UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA
ŠTUDIJ BIOTEHNOLOGIJE

Miha ŠKALIČ

INTEGRACIJA BIOLOŠKIH PODATKOV V NAPOVEDNI MODEL ZA ODKRIVANJE MOLEKULSKIH INTERAKCIJ PRI PARODONTOZI

MAGISTRSKO DELO

Magistrski študij - 2. stopnja

Ljubljana, 2016

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA
ŠTUDIJ BIOTEHNOLOGIJE

Miha ŠKALIČ

INTEGRACIJA BIOLOŠKIH PODATKOV V NAPOVEDNI MODEL ZA ODKRIVANJE MOLEKULSKIH INTERAKCIJ PRI PARODONTOZI

MAGISTRSKO DELO
Magistrski študij - 2. stopnja

BUILDING A PREDICTIVE MODEL BY INTEGRATING BIOLOGICAL DATA TO IDENTIFY MOLECULAR INTERACTIONS PRESENT IN PERIODONTITIS

M. SC. THESIS
Master Study Programmes

Ljubljana, 2016

Magistrsko delo je zaključek Magistrskega študijskega programa 2. stopnje Biotehnologije. Delo je bilo opravljeno na Biotehniški fakulteti, Oddeleku za biologijo, Katedri za biokemijo.

Študijska komisija je za mentorja magistrskega dela imenovala doc. dr. Mateja Butalo, za somentorja doc. dr. Tomaža Curka in za recenzenta prof. dr. Uroša Petroviča.

Komisija za oceno in zagovor:

Predsednica: prof. dr. Branka JAVORNIK
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za agronomijo

Član: doc. dr. Matej BUTALA
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za biologijo

Član: doc. dr. Tomaž CURK
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

Član: prof. dr. Uroš PETROVIČ
Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za biologijo

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.

Miha Škalič

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

ŠD Du2

DK UDK 616.314:577.2:575.112:004(043.2)

KG parodontoza/*Aggregatibacter actinomycetemcomitans*/biološki podatki/napovedni modeli/strojno učenje/zlivanje podatkov/molekularne interakcije/RNA-proteini

AV ŠKALIČ, Miha

SA BUTALA, Matej (mentor)/CURK, Tomaž (somentor)

KZ SI-1000 Ljubljana, Jamnikarjeva 101

ZA Univerza v Ljubljani, Biotehniška fakulteta, Študij biotehnologije

LI 2016

IN INTEGRACIJA BIOLOŠKIH PODATKOV V NAPOVEDNI MODEL ZA ODKRIVANJE MOLEKULSKIH INTERAKCIJ PRI PARODONTOZI

TD Magistrsko delo (Magistrski študij - 2. stopnja)

OP XII, 60, [10] str., 14 pregl., 32 sl., 3 pril., 61 vir.

IJ sl

JI sl/en

AI Seve bakterije *Aggregatibacter actinomycetemcomitans* povezujejo z nastankom parodontoze. Mehanizmi virulence tega oportunističnega ustnega patogena niso povsem raziskani. Za to bakterijo je značilno, da izloča številne proteine, ki bi lahko imeli toksični učinek na gostiteljske celice. V tej nalogi smo z zlivanjem podatkov s simultano matrično faktorizacijo iskali kandidatne proteine, ki bi lahko vezali molekule mRNA v človeških celicah. Analizirali smo proteom seva *A. actinomycetemcomitans* D7S in integrirali 11 matričnih relacij z namenom prioritizacije RNA vezavnih proteinov, ki se izločajo. Integrirali smo genomske, transkriptomske in proteomske podatke, anotacije ter napovedi programov. S prečnim preverjanjem smo pokazali, da je z informacijami v integriranih podatkih mogoče napovedati prisotnost RNA vezavnih domen (mediana AUC = 0,750) in sekrecijo (mediana AUC = 0,627). Napovedali smo več proteinov, ki jih bakterija izloča in hkrati kažejo sposobnost vezave RNA – prisotne so ustrezne domene v proteinih oziroma homologni proteini kandidatov so anotirani kot RNA vezavni. Obogatitvena analiza genskih skupin je pokazala, da so med skupinami genske ontologije najbolj obogatene skupine povezane s translacijo. Pokazali smo tudi, da je uporabljena metoda primerljiva oziroma v določenih pogledih celo boljša od klasičnih metod strojnega učenja. Za eksperimentalno preverbo napovedi smo interakcijo med molekulami človeške mRNA in izbranimi bakterijskimi proteini preučili z metodo na osnovi površinske plazmonske resonance. Metodo smo validirali s komponentama genotoksina CDT bakterije *A. actinomycetemcomitans*.

KEY WORDS DOCUMENTATION

ND Du2

DC UDC 616.314:577.2:575.112:004(043.2)

CX periodontitis/*Aggregatibacter actinomycetemcomitans*/biological data/predictive modelling/machine learning/data fusion/molecular interactions/RNA-proteins

AU ŠKALIČ, Miha

AA BUTALA, Matej (supervisor)/CURK, Tomaž (co-supervisor)

PP SI-1000 Ljubljana, Jamnikarjeva 101

PB University of Ljubljana, Biotechnical Faculty, Academic Study Programme in Biotechnology

PY 2016

TY BUILDING A PREDICTIVE MODEL BY INTEGRATING BIOLOGICAL DATA TO IDENTIFY MOLECULAR INTERACTIONS PRESENT IN PERIODONTITIS

DT M. Sc. Thesis (Master Study Programmes)

NO X, 60, [10] p., 14 tab., 32 fig., 3 ann., 61 ref.

LA sl

Al sl/en

AB Strains of bacteria *Aggregatibacter actinomycetemcomitans* are often found in association with periodontitis. Mechanisms of virulence of this opportunistic oral pathogen are not yet fully known. However, it is known that the bacterium has an abundant secretome, of which several proteins have a putative toxic effect on eukaryotic cells. In this thesis, we applied data fusion using simultaneous matrix factorization in order to identify candidate proteins that could bind to mRNA of human cells. We analyzed the proteome of D7S strain and by integrating 11 relations encoded in matrices we constructed secretion constrained prioritization list of RNA-binding proteins. We integrated genomic, transcriptomic and proteomic data in addition to annotations and program predictions. Using cross-validation we showed that it is possible to infer secretion of proteins (median AUC = 0.627) and presence of RNA biding domains (median AUC = 0.750). Final list reveals that there are several proteins that show mRNA binding capability and at the same time are secreted. Gene set enrichment shows that from our suggested list the candidates belong mainly to the set of genes related to translation. In addition, we have shown that our method is competitive with other machine learning techniques. Finally, we experientially tested our interactions predictions with develop human mRNA–bacterial proteins interaction probing based on surface plasmon resonance. Validation was carried out using components of *A. actinomycetemcomitans* genotoxin CDT.

KAZALO VSEBINE

KAZALO PREGLEDNIC

KAZALO SLIK

KAZALO PRILOG

OKRAJŠAVE IN SIMBOLI

AUC	Površina pod krivuljo ROC (angl. *Area under curve; ROC - Receiver Operating Characheristics*)
BLAST	Orodje za lokalno poravnavanje zaporedij *(*angl. *Basic Local Alignment Search Tool)*
CDT	Citoletalni toksin (angl. *Cytolethal distending toxins*)
COG	Gruče ortolognih skupin (angl. *Clusters of Orthologous Groups*)
DBP	DNA vezavni protein (angl. *DNA binding protein*)
FDR	Stopnja lažnih odkritij (angl. *False Discovery Rate*)
GEO	Podatkovna zbirka genskih ekspresij (angl. *Gene expression omnibus*)
GO	Genska ontologija
LC-MS/MS	Tekočinske kromatografije sklopljene s tandemsko masno spektrometrijo
RBP	RNA vezavni protein (angl. *RNA binding protein*)
rRF	Relativni rang faktorizacije
RU	Enota signala refraktometra; odzivna enota (angl. *Response unit*)
RNP	Ribonukleoprotein; kompleks med ribonukleinsko kislino in proteinom
SPR	Površinska plazmonska resonanca (angl. *Surface Plasmon Resonance*)

1 UVOD

Periodentalne bolezni sodijo med največje probleme dentalne medicine. Te bolezni so večinoma posledica infekcije in vnetja dlesni ter kosti v bližini zob. V začetku, stanju imenovanem gingivitis, dlesni postanejo otekle in lahko krvavijo. Pri napredujočem stanju, periodontitisu, pride do odstopanja dlesni, majavosti zob in izgube kostnega tkiva (CDC.gov, 2015).

Juvenilni periodontitis, imenovan tudi parodontoza, je pogosto povezan s prisotnostjo bakterije *Aggegatibacter actinomycetemcomitans* v ustni votlini. Bakterija se pojavlja v 90 % lokalizirane agresivne in 30 do 50 % primerih kronične oblike te bolezni (Raja in sod., 2014).

1.1 OPREDELITEV PROBLEMA

Poznanih je več virulenčnih dejavnikov bakterije *A. actinomycetemcomitans*, a kljub temu so mehanizmi virulence le delno raziskani. Na voljo je vse več eksperimentalnih podatkov, napovednih orodij in celovitih anotacij ter opisov lastnosti molekul bakterije *A. actinomycetemcomitans*. Integracija teh podatkov in raziskovanje struktur v podatkih je lahko pomembno za odkrivanje do sedaj neraziskanih mehanizmov virulence.

Za mehanizme virulence so potencialno zanimive molekulske interakcije med proteini patogenih bakterij in molekulami RNA gostitelja. Medvrstne interakcije RNA-protein so v precejšnji meri neraziskane. Znano je, da imajo po Gramu negativni patogeni razvite mehanizme, s katerimi spremenijo oziroma zaobidejo mehanizme imunskega sistema gostitelja. Bakterije v ta namen izločajo efektorje, najpogosteje s sistemi izločanja, predvsem tipoma III in IV, ter z izločanjem z vezikli. Sistemi izločanja omogočijo injiciranje efektorskih proteinov v celice gostitelja, kjer vplivajo na mehanizme kot so proizvodnja citokinov in zorenje lizosomov (Baxt in sod., 2013). Domnevamo, da lahko bakterije z interakcijami RNA-protein uravnavajo metabolizem RNA v gostiteljskih celicah.

1.2 CILJI NALOGE

V magistrski nalogi želimo združiti heterogene biološke podatke z namenom, da zgradimo prioritetni seznam proteinov, ki bi lahko vplivali na razvoj periodontitisa. Zanimajo nas interakcije bakterijskih proteinov patogene bakterije *A. actinomycetemcomitans* s človeško mRNA.

Nadalje želimo analizirati pridobljeni seznam z *in vitro* testom vezave proteinov z molekulami mRNA. V ta namen bomo postavili test za oceno zanesljivosti metode za napoved interakcij.

1.3 HIPOTEZE

Postavili smo naslednje delovne hipoteze:

- Metodo simultane matrične faktorizacije lahko uspešno uporabimo za integracijo heterogenih bioloških podatkov v model za napovedovanje interakcij med RNA in proteini. Med drugim tudi interakcije proteinov, za katere »preprostejše« metode, na primer metoda iskanja klasičnih RNA vezavnih domen, ne bi napovedale vezave RNA.

- Omenjeno metodo lahko uporabimo za sočasno napovedovanje vezave RNA in sekrecije molekul in tako vzpostavimo seznam proteinov, ki imajo potencialni vpliv na evkariontsko celico.

- Z biokemijskim oziroma biofizikalnim pristopom lahko okarakteriziramo interakcijo med mRNA ter proteini in s tem ocenimo uspešnost napovedi.

2 PREGLED OBJAV

2.1 LOKALIZIRANA AGRESIVNA PARODONTOZA

Agresivna oblika parodontoze povzroča hitro in huda poškodbo obzobnega tkiva. Etiologija periodontitisa je zapletena in vključuje bakterije v dentalnem biofilmu, katere lahko povzročijo vnetni odziv imunskega sistema. Ta interakcija vodi v poškodbe obzobnega tkiva. Patogene bakterije v dentalnem biofilmu so glavni vzrok patogeneze. Ustni biofilm je kompleksen ekosistem, gradi ga več kot 700 vrst bakterij. Vendar je bilo le nekaj teh bakterij identificiranih kot povzročiteljice bolezni. To so običajno po Gramu negativne, anaerobne bakterije, ki vzpostavijo svojo ekološko nišo v biofilmu. Kot najbolj pomembna in najbolj pogosta povzročiteljica agresivne parodontoze je bila identificirana bakterija *A. actinomycetemcomitans*. Druge bakterije, ki so bile povezane z napredovanjem bolezni, so *Porphyromonas gingivalis, Tannerella forsythia, Treponema denticola, Fusobacterium nucleatum, Prevotella intermedia, Prevotella nigrescens, Campylobacter rectus, Eikenella corrodens* in *Parvimonas micra* (Feng in Weinberg, 2006). Vse te bakterijske vrste sintetizirajo virulentne dejavnike, ki jim omogočajo kolonizacijo subgingvinalnega prostora, povzročitev poškodbe tkiva in odpornost proti obrambnim mehanizmom gostitelja (Chahboun in sod., 2015).

2.2 BAKTERIJA *AGGREGATIBACTER ACTINOMYCETEMCOMITANS*

Bakterija *A. actinomycetemcomitans* je fakultativno anaerobna in negibljiva baterija, ki ne tvori spor. Ta paličasta bakterija je velika 0,4-0,5 µm krat 1-1,5 µm. Mlajše kulture izolatov izgledajo kot kokobacili, medtem ko celice starejših kultur oziroma kultur gojenih na tekočih glukoznih gojiščih izgledajo bolj podolgovate. Bakterija je zahtevna za gojenje v laboratorijskih pogojih (Henderson in sod., 2010).

Visokomolekularni O-polisaharid, ki je del lipopolisaharida, je dominanten antigen. Trenutno poznamo šest serotipov *A. actinomycetemcomitans*, označenih kot serotipi a do f (Kaplan in sod., 2001). Večina oseb, katerih ustno votlino poseljuje bakterija *A. actinomycetemcomitans*, imajo v ustni flori prisoten samo en serotip *A. actinomycetemcomitans,* ki se konsistentno ohranja skozi čas pri posamezniku. Opisani pa so tudi primeri posameznikov, kjer sta bila odkrita dva oziroma celo trije serotipi hkrati. Med prisotnimi serotipi v ustni flori in geografskimi regijami ter etnično pripadnostjo gostiteljastjo je bila dokazana korelacija. Za razliko od populacij Japonske, Kitajske, Koreje in Turčije, kjer je v populaciji najpogostejši serotip c, je v populaciji ZDA pri lokalizirani juvenilni parodontozi pogostejši serotip b. Pri nordijskih narodih Finske, Švedske in Danske je porazdelitev najpogosteje zastopanih serotipov a, b in c enakomerna (Rylev in Kilian, 2008).

2.2.2 Genom bakterije *A. actinomycetemcomitans*

Eden najbolj raziskanih sevov *A. actinomycetemcomitans* je sev D7S-1, ki je uvrščen v serotip a. Sev, ki je bil izoliran iz subgingivalnega zobnega plaka, je bil prisoten pri pacientki afroameriškega porekla, diagnosticirani z agresivno parodontozo. Ima naravno sposobnost kompetence, torej prevzema tuje DNA. Chen in sod. (2010) so organizmu določili nukleotidno zaporedje in deponirali genom v podatkovno bazo GenBank (akcesijska številka: ADCF00000000).

Pangenom bakterije *A. actinomycetemcomitans* je zelo raznolik. Rezultati primerjalne genomike (Kittichotirat in sod., 2011) kažejo, da glede na podobnost obstajata dve večji skupini znotraj vrste. V eno skupino spadajo izolati serotipov a, d, e in f, v drugo pa b in c. Obstajajo tudi sevi, ki močneje odstopajo – na primer sev SC1083, ki pripada serotipu e. Glede na študijo 14 genomov *A. actinomycetemcomitans* je bilo odkritih 3.301 genov, med katerimi je 2034 ohranjenih med vrstami. Ostali geni variirajo med sevi. Sevi imajo tako med 16,7 % in 29,4 % variabilnega genoma.

2.2.3 Transkriptom bakterije *A. actinomycetemcomitans*

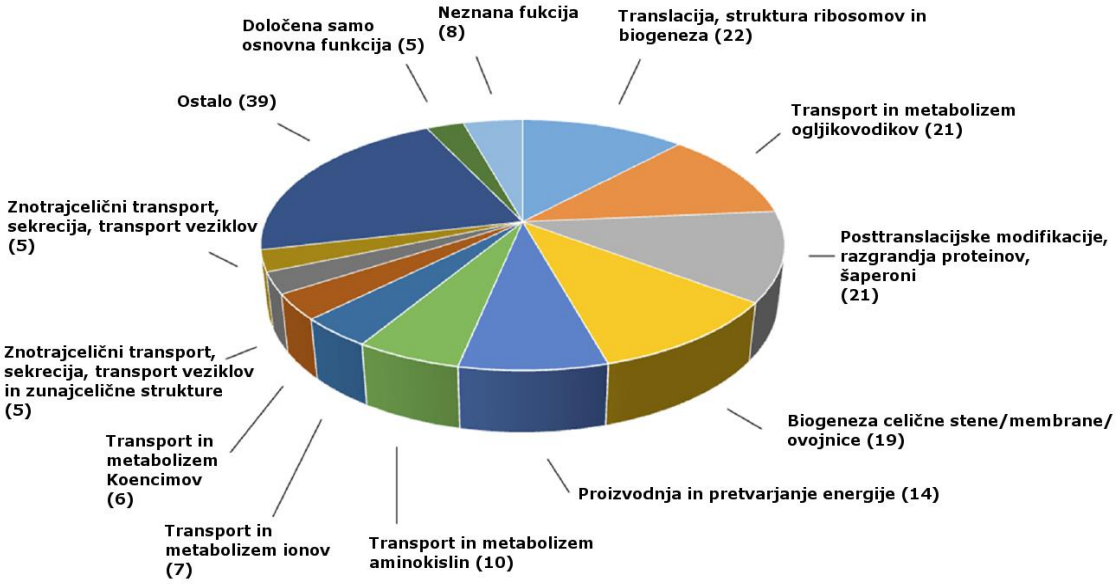
Kolonizacija tkiva gostitelja in rast patogena v gostitelju je nujna za razvoj infekcije. V ta namen patogen prilagodi svoj metabolizem in sintetizira virulentne dejavnike. Jorth in sod. (2013) so z metodo sekvenciranja RNA rekonstruirali transkriptom *A. actinomycetemcomitans* in raziskali diferencialno izražanje genov med rastjo na biofilmu *in vitro* in v mišjem abscesu *in vivo*. Od 691 kodirajočih transkripcijskih enot in 210 nekodirajočih RNA so odkrili, da se glede na testirane pogoje, diferencialno izraža ~14 % genov.

2.2.4 Sekretom bakterije *A. actinomycetemcomitans*

Za molekulske interakcije med bakterijo in gostiteljem je ključnega pomena, da se proteini, ki vstopajo v te interakcije, izločajo iz celice. Zijnge in sod. (2012) so z metodo tekočinske kromatografije sklopljene s tandemsko masno spektrometrijo (LC-MS/MS) zaznali 179 izločenih proteinov iz bakterije *A. actinomycetemcomitans* med gojenjem v biofilmu. Ugotovili so, da je delež ekstracelularnih proteinov povezanih z virulenco dosti večji, kot je prej bilo dokazano. Poleg tega so z računskimi metodami pokazali, da sev D7S uporablja sekrecije tipa I, II in V za direktno translokacijo proteinov ali pa translokacijo proteinov v dveh korakih v ekstracelularni prostor.

Preprost mehanizem sekrecije bakterije *A. actinomycetemcomitans*, kot tudi drugih prokariontov in evkariontov, je tvorjenje membranskih veziklov. Vezikli po Gramu negativnih kot tudi po Gramu pozitivnih bakterij lahko prenašajo virulenčne dejavnike, lipopolisaharide in fragmente peptidoglikana, ki stimulirajo imunski odgovor gostitelja. Raziskave veziklov z LC-MS/MS, kjer je bil analiziran klinični izolat *A. actinomycetemcomitans* (serotip e), je prav tako pokazala na številne proteine, skupno 151 proteinov v vsaj treh od štirih ponovitvah, ki se izločajo v veziklih. Večina proteinov,

identificiranih, da se izločajo z vezikli, je domnevno virulentnih (Kieselbach in sod., 2015), velik delež proteinov pa je tudi povezan s translacijo, strukturo ribosomov in biogenezo (Slika 1). Transport z vezikli je že bil podrobneje preučen za toksine CDT in levkotoksin bakterije *A. actinomycetemcomitans* (Rompikuntal in sod., 2012; Demuth in sod., 2003).



Slika 1: Deleži funkcijski skupin gruč ortolognih skupin (COG) identificiranih v veziklih bakterije *A. actinomycetemcomitans* (Kieselbach in sod., 2015).

2.2.5 Mehanizmi virulence

Do sedaj sta bila odkrita in karakterizirana dva toksina bakterije *A. actinomycetemcomitans*: RTX leukotoksin in citotoksični toksin CDT (angl. *Cytolethal Distending Toxin*). Tretji, le domnevni toksin, je kodiran v genu *cagE* – toksin CagE.

Glede na to, da tudi druge patogene bakterije, ki kolonizirajo druga tkiva v našem telesu, kot je na primer *Escherichia coli*, lahko proizvajajo ortologe naštetih toksinov, ni povsem jasno, kako so ti virulentni dejavniki povezani s patologijo dlesni. Domnevno molekule izločene iz *A. actinomycetemcomitans* vstopajo v interakcije s tkivom v stiku z bakterijo in povzročajo patološka stanja, kot sta na primer indukcija proliferacije osteoklastov ali inhibicija aktivnosti osteoblastov (Henderson in sod., 2010).

Dobro raziskan in specifičen protein *A. actinomycetemcomitans* je 113 kDa velik leukotoksin (LtxA), kodiran na operonu *ltx*: *ltxA* – *ltxD*. Produkt gena *ltxC* je odgovoren za aktivacijo toksina, medtem ko sta produkta genov *ltxB* in *ltxD* potrebna za sekrecijo toksina. Kloni JP2 bakterije *A. actinomycetemcomitans*, ki so bili povezani s hudimi oblikami parodontoze, imajo značilno delecijo v promotorski regiji operona *ltx*. Posledično

je povečana sinteza toksina. LtxA povzroči pri človeku in človeku podobnih opicah specifičen propad levkocitov, natančneje granulocitov in makrofagov (Kachlany, 2010). Toksin prepoznava antigen LFA-1 (beta2 integrin na membrani) in delno prizadene tudi eritrocite (Munksgaard in sod., 2012). Novejše študije so pokazale tudi delovanje toksinov na podganah (Schreiner in sod., 2013). Specifičnost delovanja proti celicam imunskega sistema omogoča bakteriji oslabitev imunskega odziva gostitelja. Pri velikih koncentracijah toksina se tvorijo pore v membrani gostiteljskih celic, kar sproži nekrozo. Pri manjših in fiziološko pomembnih koncentracijah je smrt celic posledica apoptoze občutljivih celic. Znano je, da se ob kontaktu s toksinom spremeni celično signaliziranje, vendar natančni mehanizmi propada celice še niso raziskani (Kachlany, 2010).

Podobno kot druge po Gramu negativne bakterije (na primer *Campylobacter jejuni*, *E. coli*, *Salmonella enterica* in *Shigella dysenteriae*), bakterija *A. actinomycetemcomitans* sintetizira tudi holotoksin CDT. Citotoksin je sestavljen iz treh proteinov. Pripadajoči geni *cdtA*, *cdtB* in *cdtC* so zapisani v operonu *cdt*. Odkrito je bilo, da Cdt onemogoči prehod iz faze G_2 v mitozo in tako delitev celic gostitelja. Pred prehodom v mitozo gostiteljska celica zazna poškodbo DNA in namesto mitoze se sproži apoptoza. Za poškodbe so dovzetne predvsem celice epitela (DiRienzo, 2014a). Aktivna komponenta holotoksina je produkt gena *cdtB*, ki naj bi deloval kot genotoksin, saj vstop CdtB v jedro celice sproži signalno kaskado, ki se sproži tudi ob poškodbah DNA. Direktni dokaz, da CdtB deluje kot Dnaza, še ni bil objavljen. Za transport do jedra gostiteljske celice sta pomembni podenoti CdtA in CdtC. Podenote se sestavijo v periplazmatskem prostoru in se izločijo iz celice. Holotoksin prepozna receptorje tarčne celice in z endocitozo preide v notranjost, kjer so izrabljeni gostiteljevi celični mehanizmi za dostavo komponente B do celičnega jedra (DiRienzo, 2014b).

Tretji zanimiv in manj raziskani kandidatni toksin je protein CagE. Pri bakteriji *H. pylori* se ta protein injicira v celice gostitelja s sekrecijskim sistemom tipa IV. Protein spremeni delovanje celic s tem, da sproži podvojevanje celic, apoptozo in morfološke spremembe. Vendar protein CagE v primeru *A. actinomycetemcomitans* okužbe zaenkrat ostaja neraziskan (Teng in Zhang, 2005).

2.2.6 Odziv gostitelja na okužbo

Transkriptomska analiza kulture epitelnih celic z uporabo mikromrež je pokazala, da ob okužbi z bakterijo *A. actinomycetemcomitans* pride do diferencialnega izražanja genov vključenih v p53 apoptotične poti. Tu prihaja do razlik v primerjavi z bakterijo *Porphyromonas gingivalis*, še eno bakterijo povezano s parodontozo. Pri tej okužbi p53 metabolna pot namreč ni bila aktivirana. Ugotovljeno je še bilo, da apoptotična pot ni bila aktivirana s strani Fas ali TNFα. Na splošno je bil odgovor človeških celic precej drugačen glede na vrsto patogenega organizma (Handfield in sod., 2005).

2.3 MOLEKULSKE INTERAKCIJE RNA-PROTEIN

V celicah evkariontov organizmov potekajo obsežne post-translacijske modifikacije mRNA, kar prispeva k dodatnemu nivoju genske regulacije. Pri procesiranju RNA sodelujejo tako trans-delujoče RNA kot tudi RNA vezavni proteini (*angl. RNA binding proteins* - RBP), ki ob vezavi tvorijo ribonukleoproteine (RNP). Čeprav RBPji vežejo RNA, se med sabo razlikujejo v specifičnosti in afiniteti glede na zaporedje in strukturo RNA. Ključnega pomena so domene proteinov, ki vstopajo v interakcijo z RNA in drugimi proteini. Pogosto so ti proteini post translacijsko modificirani. Rezultat je raznolikost RNPjev. RBPji sodelujejo v celotni verigi procesiranja RNA od transkripcije, izrezovanja pre-mRNA in poliadenilacije do RNA modifikacij, transporta, lokalizacije, translacije in končnega razkroja (Glisovic in sod., 2008).

Med klasične RNA vezavne domene spadajo RNA prepoznavni motiv, dvoverižna RNA vezavna domena, homologija K, zaporedje RGG in domene PUM. Obstaja pa še veliko domen, ki niso anotirane. Teh domen ne moremo zaznati z iskanji homologije (Livi in sod., 2015). Primer, kjer so eksperimentalno odkrili velik del takih proteinov, je študija Kwona in sod. (2013). Preučevali so interaktom RNA-protein na mišjih matičnih celicah. Ob prečnem povezanju parov molekul z UV so z masno spektrometrijo odkrili 555 RNA-vezavnih proteinov. Od teh je bilo kar 283 identificiranih proteinov brez do sedaj poznane RNA vezavne domene. Podobno sliko je pokazal tudi interaktom celične linije HeLa, kjer pri 402 od 860 proteinov ni bilo odkritih RNA vezavnih domen (Castello in sod., 2012).

2.3.1 Interakcije ob okužbah

Za razliko od DNA vezavnih proteinov, ki so že precej raziskani in pri katerih poznamo tako inhibitorne učinke, kot je učinek toksina CdtB pri *A. actinomycetemcomitans,* kot tudi stimulatorne učinke, na primer mimika transkripcijskih dejavnikov TAL efektorjev pri rastlinskih patogenih (Deslandes in Rivas, 2012), pa o RNA vezavnih proteinih ne vemo veliko. Znano je, da se ob okužbah nivo translacije upočasni. Ta efekt je bil preučen pri več patogenih organizmih: *Pseudomonas aeruginosa*, *Pseudomonas entomophila*, *Salmonella spp.*, *Shigella flexneri* in *Legionella pneumophila*. Pri virusnih okužbah je ta prilagoditev del obrambnega sistema, saj virusi uporabljajo gostiteljev sistem translacije za proizvodnjo lastnih peptidov. Težje razložimo, zakaj prihaja do inhibicij v primeru bakterijske okužbe, če pa imajo bakterijske celice lasten sistem za translacijo. Prevladujeta dve hipotezi. Po prvi je taka inhibicija pri bakterijskih okužbah posledica prilagoditve celic na stres, po druga naj bi patogeni tako zavirali delovanje imunskega sistema (Lemaitre in Girardin, 2013). Pri bakteriji *Legionella pneumophila* je za pet proteinov, od tega tri glukoziltransferaze, že bilo pokazano, da proteini v gostitelju inhibirajo translacijo. Glukoziltransferaze in še en protein z neznano molekulsko funkcijo vplivajo na elongacijski dejavnik eEF1A in inhibirajo translacijo. Mutanti z onesposobljenim vsemi kodirajočimi geni teh proteinov še vedno kažejo delno inhibicijo translacije (Fontana in sod., 2011). Prisotnost drugih bakterijskih inhibitornih dejavnikov je v skladu z domnevo, da bi lahko z vezavo mRNA bakterijski RBPji inhibirali translacijo.

Bolj kot pri bakterijah so raziskane interkacije z molekulami RNA pri virusih. Vsi virusi z negativno verigo RNA namreč kodirajo protein, ki nespecifično in z veliko afiniteto veže enoverižno molekulo RNA. Prvotna naloga teh proteinov je, da obdajo virusni genom za namene RNA traskripcije, podvojevanja in pakiranja. Proteini pa tudi vplivajo na procese v gostiteljskih celicah (Portela in Digard, 2002).

2.3.2 Napovedovanje interakcij

Za *in silico* napovedovanje RNA vezave moramo vedeti, ali določeni protein veže RNA, kateri aminokislinski ostanki so v neposrednem stiku z RNA, kateri nukleotidi reagirajo s proteinom in kakšna je struktura kompleksa RNA-protein. Znana terciarna struktura proteina bistveno olajša odkrivanje vezavnih mest. Vezavna mesta so aminokislinski ostanki v stiku z okolico, ki so običajno, vendar ne nujno, v neprekinjenem zaporedju. Mesta so praviloma pozitivno nabita in privlačijo negativno nabito RNA.

Metode za napovedovanje interakcij lahko razdelimo v dve skupini: metode, ki temeljijo na strukturah in metode, ki temeljijo na zaporedjih. V strukturnih metodah je porazdelitev naboja lahko opisano neposredno, medtem ko je pri metodah, ki uporabljajo zaporedja, ta informacija posredna.

Metoda catRAPID (Bellucci in sod., 2011) omogoča napoved interakcij med parom proteina in molekule RNA. Napovedni algoritem je bil naučen na podatkih Protein Data Bank tako, da diskriminira med molekulami, ki vstopajo in tistimi, ki ne vstopajo v interakcije. Upoštevane so lastnosti sekundarne strukture, vodikove vezi in van der Waalsove vezi (Cirillo in sod., 2012).

Različica programa s spletnim vmesnikom, imenovana catRAPID omics, omogoča primerjavo enega zaporedja proteina oziroma RNA s proteomom oziroma transkriptomom želenega organizma. S tem programom lahko parom RNA-protein napovedujemo dovzetnost za interakcije, zanesljivost napovedi, jakost interakcije in prisotnost proteinskih domen oziroma RNA motivov, ki vplivajo na vezavo (Agostini in sod., 2013).

2.4 ZLIVANJE PODATKOV Z MATRIČNO FAKTORIZACIJO

Tehnike integracije podatkovnih virov lahko razdelimo v skupine s tremi strategijami (Žitnik, 2015a):

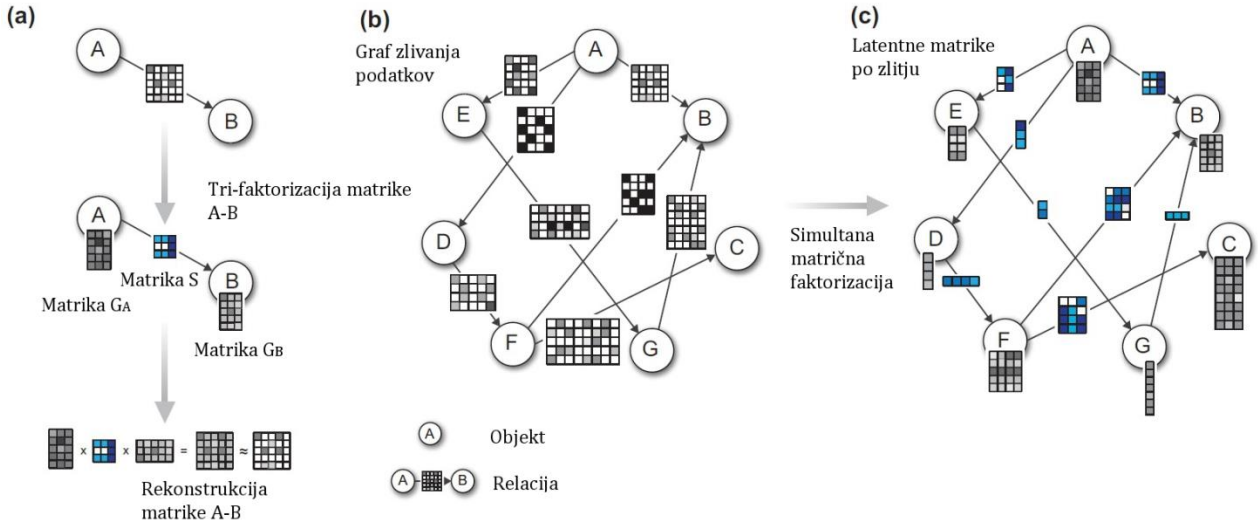
- zgodnja integracija združuje podatke v eno učno množico pred začetkom učnega procesa.
- vmesna integracija zajema izračune relacij med podatkovnimi viri in proizvede kombinirane poglede, ki so nato dani učnemu algoritmu.
- pozna integracija uporabi učni algoritem ločeno v vsaki predstavitvi in nato združi rezultate.

V analizi bioloških podatkov se velikokrat srečujemo s heterogenimi podatki, katerih struktura lahko ima dobro napovedno moč za napovedovanje novih lastnosti. Za odkrivanje teh struktur se je kot učinkovita izkazala metoda simultane matrične faktorizacije, ki spada med metode z vmesno integracijo (Žitnik, 2015a). Simultano matrično faktorizacijo lahko uporabimo za zlivanje podatkovnih naborov, ki jih lahko predstavimo v matrični obliki. Podatki v matriki povezujejo dva objekta. Na primer, objekta bakterijski protein in človeška RNA sta lahko povezana z matriko, ki opisuje verjetnost, da prihaja do interakcije med tema dvema molekulama. Vsak objekt je lahko povezan z več objekti in tako lahko tvorimo sistem povezanih matrik (Slika 2b).

Sistem podatkovnih naborov lahko nadalje modeliramo z zlivanjem podatkov - simultano matrično faktorizacijo. V tem postopku zgostimo podatke v manjše latentne matrike, ki jih lahko rekonstruiramo v matrike, podobne prvotnim. Podatke vsake relacije zgostimo v tri latentne matrike, dve matriki objektov (matriki G) in matriko relacije (matrika S). Rekonstrukcijo matrike (Slika 2a) izvedemo po formuli (1).

$$R_{i,j} \approx G_i S_{i,j} G_j^T \quad \ldots (1)$$

Relacije, ki vključujejo isti objekt, si med sabo delijo latentno matriko G. S tem zagotovimo, da se z zlivanjem podatkov ohranjajo relacije med podatkovnimi tipi. Nadaljnja prednost algoritma je veriženje latentnih matrik. Z zgoščevanjem je mogoče ohranjati strukturo dveh objektov, čeprav ta nista neposredno povezana z relacijsko matriko (Žitnik in sod., 2015).



Slika 2: Shematski prikaz zlivanja podatkov z matrično faktorizacijo. (a) Tri-faktorizacija in rekonstrukcija posamične matrike. (b) Relacijske matrike pred zlivanjem podatkov. (c) Latentne matrike po simultani matrični faktorizaciji (Žitnik in sod., 2015).

Vsakemu objektu matrične faktorizacije se določi rang faktorizacije (k). Od ranga k je odvisna oblika latentnih matrik in posledično podobnost rekonstruiranih matrik prvotnim. Če objektu A dodelimo rang k_i in objektu B rang k_j, dobimo latentne matrike oblike (2).

$$G_A \in \mathbb{R}^{|A| \times k_i};\ S_{AB} \in \mathbb{R}^{k_i \times k_j};\ G_A \in \mathbb{R}^{|B| \times k_j} \quad \ldots (2)$$

Algoritem matrične faktorizacije poteka iterativno, pri čemer se išče vsota najkrajših razdalj med relacijami in njihovimi rekonstrukcijami (enačba 3). Pri tem pa lahko še vključimo omejitvene matrike, ki povezujejo člene istega objekta (enačba 4).

$$\min_{G_i \geq 0, S_{i,j}} \sum_{R_{i,j} \in R} \left\| R_{i,j} - G_i S_{i,j} G_j^T \right\|^2 \quad \ldots (3)$$

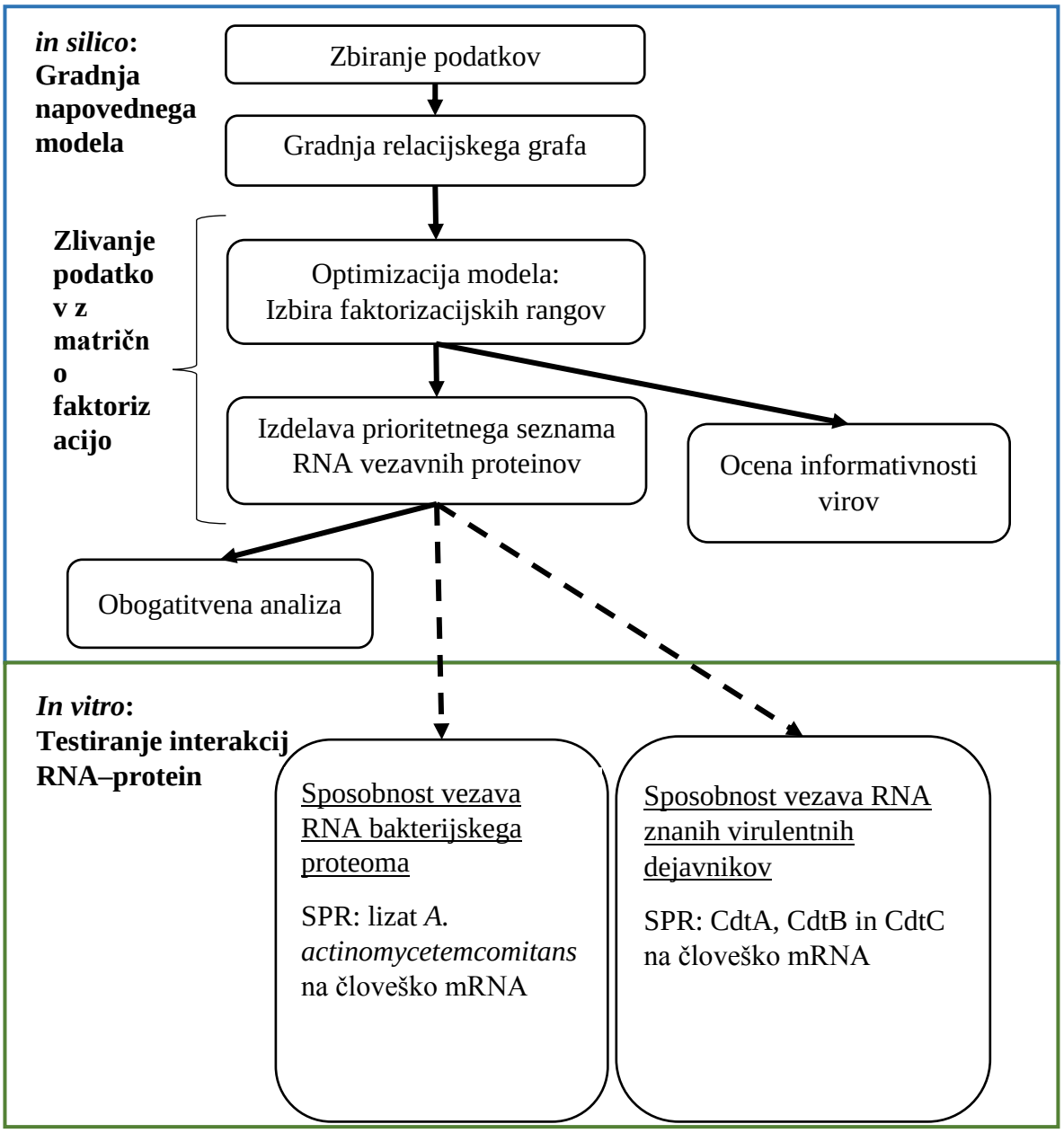
$$\min_{G_i \geq 0, S_{i,j}} \sum_{R_{i,j} \in R} \left\| R_{i,j} - G_i S_{i,j} G_j^T \right\|^2 + \sum_{t=1}^{max_i t_i} tr(G^T \theta^{(t)} G^T) \quad \ldots (4)$$

Pri enačbah predstavlja prvi člen seštevanja Frobeniusovo razdaljo in drugi člen vsota elementov na diagonali matrike – sled matrike (Žitnik in Zupan, 2015a).

3 MATERIAL IN METODE

3.1 POTEK DELA

Nalogo smo razdelili v dva dela (Slika 3). V prvem, *in silico*, delu smo zgradili relacijski graf, testirali obnašanje napovednega modela ob zlivanju podatkov, zgradili prioritetni seznam kandidatnih proteinov in preverili obogatenost genskih skupin. Drugi del je bil namenjen testiranju RNA vezave znanih viruletnih dejavnikov in celotnega proteoma bakterije *A. actinomycetemcomitans* z biokemijskimi tehnikami.



Slika 3: Potek dela.

3.2 MATERIALI

Bakterije in celične linije:

- Izolat *A. actinomycetemcomitans*, serotip C izoliran iz slovenskega pacienta s parodontozo
- Humana celična linija A549, epitelijske celice pljuč
- Humana celična linija MG-63, celice kostnega tkiva

Kemikalije, kompleti kemikalij, pripomočki in drugi (potrošni) material:

PolyATtract mRNA Isolation Systems 1000 (Promega), Sensor Chip SA (GE Healthcare Life Sciences), NativePAGE Novex 4-16 % Bis-Tris Protein Gels - 1.0 mm x 10-well (Life Technologies), GeneRuler 1 kb Plus DNA Ladder, PageRuler Plus Prestained Protein Ladder, cOmplete ULTRA Tablets (Roche), MOPS pufer (40 mM MOPS, 10 mM natrijev acetat, 1 mM EDTA), 4x nanašalni SDS pufer (200 mM Tris-HCl, 8 % SDS, 40 % glicerol, 4 % β-mercaptoethanol, 50 mM EDTA, 0,08 % bromophenol modro), 6X DNA Gel Loading Dye (TermoFisher), raztopina za barvanje proteinskih gelov (50 % dH_2O, 40 % MeOH, 10 % ocetna kislina, 2,5 g/L barvila Coomassie Brilliant Blue), raztopina za razbarvanje proteinskih gelov (15 % MeOH, 10 % ocetna kislina, 75 % dH_2O), tekoči dušik, 15 in 50 mL centrifugirke brez RNAz (Corning), mikrocentrifugirke in druge kemikalije navedene v besedilu naloge.

Laboratorijska oprema:

Refraktometer Biacore X (GE Healthcare Life Sciences), Spektrofotometer Nanodrop ND-1000 (Thermo Scientific), pH-meter Seven Multi (Metlar-Toledo), centrifuga Rotanta 460R (Hettich), centrifuga 3-30KS (Sigma), centrifuga 5418 (Eppendorf), sonifikator VCX 750 (Sonics), laminarij 1V2 (Iskra), Bunsenov gorilnik FireBoy (IBS Intergra Biosciences), Sistem za slikanje gelov G:BOX (Syngene), tehntnica (Sartorius), električni usmernik EPS (Amersham Pharmacia Biotech), elektroforezna kadička HE 33 Mini Submarine Unit (Hoefer), elektroforezni sistem Novex Mini Cell (Invitrogen), električni usmernik E143 (Consort), magnetna mešala, stresalnik, komora za anaerobno gojenje, hladilnik (T = 4 °C), zmrzovalnik (T = -20 in -80 °C) polavtomatske pipete (Gilson, Eppendorf in Biohit).

Strojna oprema:

- Računalnik RHEL 4.1.2, 24 jeder 64-bit, 48 GB rama
- Računalnik Windows 7, 4 jedra 64-bit, 8 GB rama

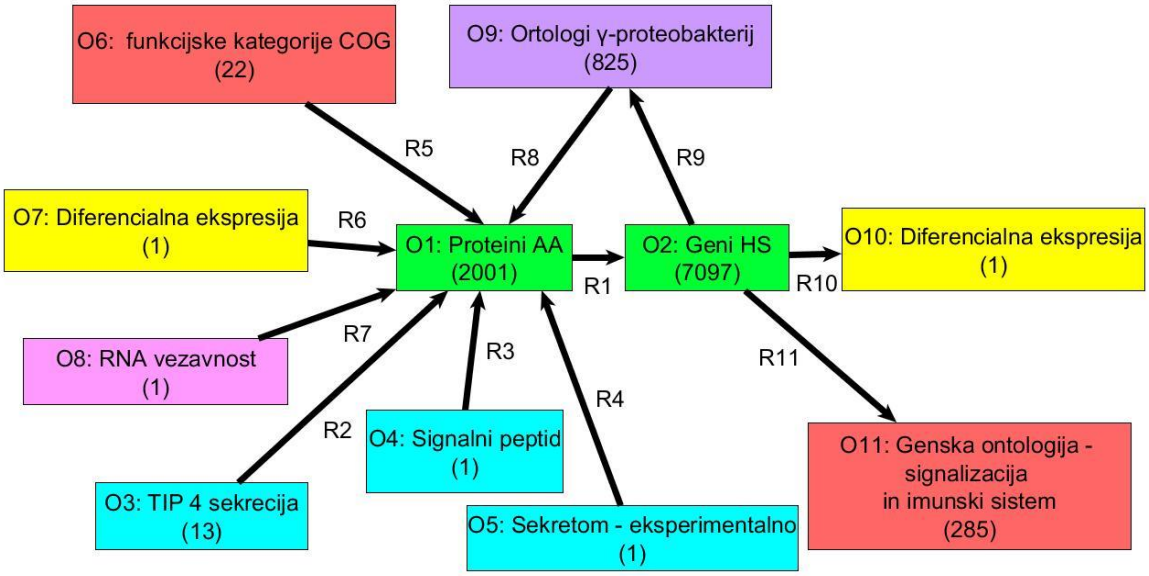
Programska oprema:

- Programski jeziki: Python (3.4.1), R (3.2.0)
- Programi: Hmmer (3.0), SignalP (4.1), S4TE (1.2), BLAST (2.2.28+), InterProScan (5.15-54.0), Blast2GO (3.1.3), Fasta (36.3.8c)

- Python knjižnice in razširitve: IPython (3.2.1), Numpy (1.9.2), Scipy (0.16.0), Matplotlib (1.4.3), Scikit-fusion (0.2.1), Scikit-learn (0.16.1), Biopython (1.65), Pandas (0.16.2), Goatools (0.5.9), Biomart (0.4.0), Statsmodels (0.7.0), Orange3 (3.2)

3.3 ZBIRANJE PODATKOV

Graf zlivanja podatkov (Slika 4) obsega 11 objektov in 11 relacij. Vključeni so tako eksperimentalni podatki kot tudi *in silico* izračuni in napovedi. Pripravo podatkov in tvorbo matrik smo izvedli v programskem jeziku Python (različica 3.4.1). Delo z zaporedji in razčlenjevanje rezultatov je bilo večinoma izvedeno s pomočjo paketa Biopython (Cock in sod., 2009).



Slika 4: Graf zlivanja podatkov. V oklepajih je zapisano število elementov objekta.

3.3.1 Genom in proteom bakterije

Analizirali smo genom bakterije *A. actinomycetemcomitans*, sev D7S-1 z NCBI referenčno številko NC_017846.1. Od proteoma, ki obsega 2255 proteinov, smo v analizo vključili 2001 proteinov, velikosti od vključno 50 do vključno 750 aminokislin. Izbor proteinov smo izvedli zaradi omejitev dolžine zaporedij pri programu catRAPID.

Za določitev podobnosti analiziranih proteinov proteomu nepatogene bakterije *E. coli* K-12 (NCBI akcesijska številka: NC_000913.3) smo uporabili globalno poravnavo programa Fasta (ggsearch36, privzete nastavitve). Podobnost smo definirali kot delež enakih aminokislin v poravnavi.

3.3.2 Interakcije protein–RNA

Za napovedovanje interakcij med molekulskimi pari protein-RNA smo uporabili program catRAPID omics v spletnem vmesniku. Analizirali smo interakcije vseh proteinov s kodirajočim transkriptom človeka (*Homo sapiens*). S pomočjo skriptnega programa smo pošiljali zahtevke za izračun interakcij z analiziranimi proteini in po tem tudi brali rezultate. Uporabljena je bila aktualna različica programa (avgust 2015), ki vsebuje RNA zaporedja iz genomske anotacije Ensembl 68.

Molekule RNA smo povezali z njihovimi geni na podlagi anotacije Ensembl 68. Tako smo vzpostavili relacijske pare *A. actinomycetemcomitans* protein – človeški gen. V relacijski matriki je kodirana mediana jakosti interakcij (Z vrednost) transkriptov, ki pripadajo določenemu genu, s proteini *A. actinomycetemcomitans*. Nabor analiziranih genov smo omejili na tiste, za katere smo imeli eksperimentalne podatke o diferencialnem izražanju (relacija R10).

3.3.3 Vezava RNA in DNA

Prisotnosti RNA in DNA vezavnih domen smo uvozili iz rezultatov catRAPID. Program ob analizi izvede tudi iskanje znanih RNA in DNA vezavnih domen v podatkovni bazi Pfam. Prisotnost RNA vezavne domene smo predstavili z vrednostjo 1, medtem ko smo DNA vezavne domene predstavili z vrednostjo 0,5. Zaradi omejitev zlivanja podatkov s paketom scikit-fusion smo iz prvotno 1 vrstice pomnožili vrstice v 5 vrstic.

3.3.4 Vključitev sekretomskih podatkov in napovedi

Prisotnost signalnih zaporedij in možnost izločevanja proteinov smo napovedovali s programoma SignalP (različica 4.1; Petersen in sod., 2011). Za izločanje proteinov s sekrecijskim sistemom tipa IV smo uporabili program S4TE (različica 1.2; Meyer in sod., 2013). Vključili smo še podatke o eksperimentalno določenih proteinih v sekretomu.

Iz skupine modulov programa S4TE smo vključili:

1. *De novo* iskanje regulatornih motivov RT-TY
2. Homologija proteinov z znanimi efektorji tipa IV
3. Prisotnost evkariontskih domen
4. Prisotnost prokariontskih domen
5. Prisotnost jedrnih lokalizacijskih signalov
6. Prisotnost prenilacijske domene
7. Struktura obvite vijačnice
8. Bazičnost karboksi-terminalnega konca
9. Naboj C-terminalnega konca
10. Hidrofilnost C-terminalnega konca
11. Hidrofilnost celotnega proteina
12. Vsebnost E bloka (niz aminokislinskih ostankov bogat z glutamatom) v zaporedju

Trinajsta vrstica v relaciji je vsota prispevkov prejšnjih vrstic normalizirana z največjo vrednostjo, ki jo dosežejo proteini. Vsakemu modulu, razen modulu homologija, smo dodelili prispevek 1. Homologiji z znanimi efektorji tipa IV smo dodelili prispevek 3.

Napovedi SignalP smo izračunali s privzetimi nastavitvami za gram negativne bakterije. V relacijo smo vstavili napovedane D vrednosti. Število vrstic smo razširili v 5.

Eksperimentalni sekretom (Zijnge in sod., 2012) smo integrirali na podlagi homologije med proteini, saj se podatki nanašajo na drugi bakterijski sev. Zbrali smo zaporedja identificiranih proteinov in jih z BLAST poravnavo primerjali z našim referenčnim proteomom. Najbolj podoben protein smo določili na podlagi najboljše poravnave po formuli (5). Na podoben način smo izvedli tudi preslikavo proteoma v veziklih (Kieselbach in sod., 2015)

$$Podobnost = \frac{(\%\ identičnost\ v\ poravnavi)\ (dolžina\ poravnave)}{(dolžina\ izločenega\ proteina\ )} \quad \ldots (5)$$

V relaciji smo homologa kodirali z vrednostjo 1, ostale vrednosti pa z 0. Število vrstic smo tudi v tem primeru razširili v 5.

3.3.5 Diferencialno izražanje genov med *in vitro* ter *in vivo* rastjo

Podatke o diferencialnem izražanju (Jorth in sod., 2013) smo integrirali na podlagi seznama diferencialno izraženih genov. Diferencialno izraženim lokusom smo poiskali pripadajoče gene v genomu (akcesijska številka: ADCF01000001). Za vsak protein smo nato poiskali identičen protein v našem referenčnem proteomu. Diferencialno izraženim genom smo pripisali dvojiški logaritem vrednosti relativne spremembe izražanja. Vrednosti smo omejili navzgor s 5 in navzdol s 5. Ostalim proteinom smo pripisali vrednost 0.

3.3.6 Ortologne skupine in funkcije ortolognih skupin

Proteinom pripadajoče ortologne skupine proteobakterij taksonomske skupine gama smo določili s pomočjo podatkovne zbirke EggNOG4.1 (Powell in sod., 2014). Našim zaporedjem smo iskali ortologe gama-proteobakterij. To smo izvedli s programom Hmmer – hmmscan (Eddy, 2009). Poiskali smo zaporedja z E-vrednostjo manjšo od 0,01. V primeru človeških genov, ki kodirajo več proteinov, smo primerjali prvi protein v zbirki Ensembl. Objekt ortolognih skupin vsebuje vse skupine, ki imajo pet ali več zadetkov v obeh primerjavah. Vrednosti so kodirane binarno z 0 oziroma 1 v primeru homologije.

Proteinom *A. actinomycetemcomitans* smo dodatno določili funkcije ortolognih skupin. V tem primer smo proteine primerjali s programom mmmscan na bazo vseh bakterijskih ortolognih skupin. Proteine smo povezali s funkcijami, ki so del anotacije zadetkov (E-vrednost > 0,01). Tudi v tem primeru smo vrednosti relacije zapisali binarno.

3.3.7 Genska ontologija

Poiskali smo skupine iz genske ontologije (http://amigo.geneontology.org/; dne 19.10.2015), ki vsebujejo termin signalizacija ali imunski (angl. *signaling* ali *immune*). Izbranim terminom smo poiskali pripadajoče gene na podlagi Ensembl 69 identifikacijskih številk (s septembrom 2015 Ensembl 68 več ni dosegljiv). V končno relacijo smo vključili skupine GO h katerim pripada 25 ali več genov. Vrednosti smo zapisali binarno z 1, v primeru pripadnosti genu, in 0, če gen ne pripada skupini.

3.3.8 Diferencialno izražanje človeških genov

Za izgradnjo ekspresijske relacije smo uporabili javno dostopne podatke iz GEO podatkovne baze, natančneje podatkovni set z oznako GSE9723. S spletnim orodjem GEO2R smo združili podatke mikromrež v kontrolno skupino in skupino, kjer so bile celice okužene z *A. actinomycetemcomitans*. Razliko v ekspresiji sonde i (DE_i) smo določili po formuli (6).

$$DE_i = log_2(mediana(okuženi)) - log_2(mediana(neokuženi)) \quad \ldots (6)$$

V premeru, da je bilo prisotnih več sond za posamezni gen, smo v relacijo zapisali vrednost DE_i, ki najbolj odstopa od vrednosti 0. Število vrstic smo tudi v tem primeru razširili v 5.

3.4 MATRIČNA FAKTORIZACIJA

Simultano matrično faktorizacijo relacijskega grafa smo izvedli s Python modulom scikit-fusion 0.2.1 (Žitnik, 2015b). Range faktorizacije (*RF*) smo vsem objektom izbirali po enakem postopku, in sicer glede na relativno velikost objektov. Za objekt *A* smo tako izračunali rang faktorizacije po enačbi (7). V enačbi (2) fukcija *ceil* zaokroži argument navzgor in rRF je relativni rang fakotrizacije, ki je enak vsem objektom v grafu.

$$RF_A = Max(\, ceil(|A|) * \, rRF, \; 2) \quad \ldots (7)$$

3.4.1 Izbira optimalnega ranga faktorizacije

Optimalni relativni rang faktorizacije smo izbrali z 10-kratnim prečnim preverjanjem. Izbirali smo med rRF vrednostmi: 1 %, 2,5 %, 5 %, 10 % 12,5 %, 15 % in 20 %. V vsakem koraku smo tako prikrili desetino podatkov in napovedovali uspešnost rekonstruiranih matrik po zlivanju. To smo izvedli z metodo metodo DFMC (angl. *Data Fusion by Matrix Completion*) iz paketa scikit-fusion. Zanimala nas je uspešnost napovedovanja vrednosti, predstavljenih v preglednici 1.

Preglednica 1: Napovedovanje vrednosti za namen izbire optimalnega ranga faktorizacije.

Napovedovanje	Oznaka relacije	Merilo
Interakcije RNA-protein	R1	Vsota abolutnih razlik
RNA vezavne domene	R7	AUC
DNA ali RNA vezavne domene	R7	AUC
Signalno zaporedje (SignalP)	R3	AUC
Prisotnost proteina v sekretomu	R4	AUC

3.4.2 Določanje informativnosti virov

Podobno kot pri izbiri optimalnega ranga faktorizacije smo tudi pri določevanju informativnosti virov uporabili 10-kratno prečno preverjanje. Pri predhodno določenem relativnem rangu faktorizacije smo posamično iz grafa izvzemali relacije in po zlivanju preverjali uspešnost napovedi. Iz grafa smo izvzemali vse relacije razen tiste, ki je povezana z napovedovano vrednostjo. Tudi v tem primeru so nas zanimale napovedi, prikazane v preglednici 1 (razen interakcije RNA-protein).

3.4.3 Primerjava metode zlivanja podatkov z ostalimi metodami strojnega učenja

Preverili smo učinkovitost naše metode v primerjavi z metodama naključnih gozdov in logistične regresije. Napovedovali smo vrednosti v preglednici 1 (razen interakcije RNA-protein) in tudi v tem primeru z 10-kratnim prečnim preverjanjem. Kot značilke smo uporabili vrednosti v sedmih relacijah, neposredno povezanih s proteini *A. actinomycetemcomitans*. Osma relacija pa vključuje informacije o napovedovanih vrednostih. Za obe metodi smo uporabili implementacijo algoritmov v knjižnici Scikit-learn. V primeru naključnih gozdov smo uporabili 100 napovednih dreves, za ostale vrednosti smo uporabili privzete nastavitve. Uspešnost klasifikatorjev smo preverjali z Nemenyi testom pri statistični značilnosti 0,05 in izrisali graf kritičnih razdalj (Demšar, 2006).

3.5 NAPOVEDOVANJE VEZAVE PROTEINOV *A. ACTINOMYCETEMCOMITANS* S ČLOVEŠKO RNA

Sposobnost vezave RNA in rang na kandidatni listi smo določali na podlagi rekonstrukcije dveh vrednosti:

1. Napovedi prisotnosti RNA vezavne domene (R7) in
2. povprečne Z vrednosti interakcij s človeškim RNA (R1)

Rang vezave proteina *i* smo določili na podlagi vsote zbirnih funkcij verjetnosti – formula (8).

$$Vezavnost_i = P\left(X_{RBP} \leq x_{i\_RBP}\right) + P\left(X_{Z\ vrednosti} \leq x_{i\_z\_vrednost}\right) \quad \ldots (8)$$

Pri tem sta $X_{Z\ vrednosti}$ in X_{RBP} β porazdelitvi, ki se najbolje prilagajata podatkom.
Iz tega seznama smo nato izbirali proteine, ki jim je bila dokazana prisotnost v sekretomu (Zijnge in sod., 2012) oziroma v veziklih (Kieselbach in sod., 2015) ali pa napoved programa signalP kaže na prisotnost signalnega zaporedja. Tak seznam smo predlagali kot kandidatni seznam RNA vezavnih proteinov s potencialnim vplivom na človeške celica.

3.5.1 Analiza obogatenosti genskih skupin in podobnost z znanimi sesalskimi RBP

Analizirana proteinska zaporedja smo okarakterizirali s skupinami GO. Povezavo smo pridobili z Blast2GO kartiranjem (privzete nastavitve). Predhodno smo poiskali zadetke BLAST (e-vrednost < 0.1, privzete nastavitve ostalih parametrov) naših proteinov v bazi RefSeq protein (različica 15. december 2015) in jih klasificirali z orodjem InterProScan (privzetne nastavitve). Verjetnost naključne obogatitve smo izračunali s pomočjo hipergeometrične porazdelitve (enačba (9) pri znanih velikostih skupin; k predstavlja velikost preseka med množico izbranih proteinov in množico proteinov prisotnih v GO skupini) in izračunali q-vrednosti s popravkom za nadzor stopnje lažnih odkritij (FDR) po postopku Benjamini–Hochberg.

$$Verjetnost\ naključne\ obogatitve = P_{hiperg.}(X \geq k) \quad \ldots (9)$$

Prav tako smo uporabili hipergeometrično porazdelitev za obogatitveno analizo humanih tarčnih mRNA. Za 20 predhodno izbranih proteinov smo poiskali molekule mRNA, s katerimi je napovedana vezava najmočnejša. Za vsak protein smo tako izbrali 10 genov z največjo napovedano vrednostjo vezave ob odštetju povprečne vezave dotičnega gena. Za izbrano množico človeških genov smo preverili, katere GO skupine so obogatene. Za povezavo med geni in GO skupinami smo uporabili povezave v bazi Ensembl, različica 67.

Z BLAST primerjavo smo preverjali, ali je vrh našega prioritetnega seznama podoben odkritim RBPjem v sesalskih organizmih. Zbrali smo zaporedja iz študij Castello in sod. (2012), Kwon in sod. (2013) ter Baltz in sod. (2012). Skupno 4670 pridobljenih zaporedij smo uporabili kot bazo za primerjavo. Protein z našega seznama smo označili kot podoben, če je imel vsaj en zadetek z e-vrednostjo nižjo od 0,01.

3.6 TESTIRANJE RNA VEZAVE ZNANIH VIRULETNIH DEJAVNIKOV

3.6.1 Gojenje bakterij *A. actinomycetemcomitans*

Predhodno zamrznjeni sev bakterije *A. actinomycetemcomitans*, ki pripada serotipu C in je bil izoliran iz pacienta s kroničnim parodontitisom (Obradović in sod., 2014), smo nacepili na trdno gojišče (Preglednica 2). Kulturo smo gojili tri dni v komori s sestavo zraka: 10 % CO_2, 5 % H_2 in ostalo N_2 pri 37 °C. Sledilo je precepljene v 5 mL tekočega gojišča 1 (Preglednica 3). Kulturo smo gojili en dan pri enakih pogojih in nato prenesli 1 mL gojišča v 75 mL g tekočega gojišča 2 (Preglednica 4). Po dveh dneh gojenja smo požeti biomaso z

10 minutnim centrifugiranjem pri 8000 g in shranili pridobljeno biomaso v zmrzovalniku pri -80 °C.

Preglednica 2: Trdno gojišče.

Komponenta	Koncentracija
Triptozni sojin bujon (TSB)	30 g/L
Kvasni ekstrakt	0,6 %
Glukoza	0,8 %
Agar	14 g/L

Preglednica 3: Tekoče gojišče 1.

Komponenta	Koncentracija
Triptozni sojin bujon (TSB)	30 g/L

Preglednica 4: Tekoče gojišče 2.

Komponenta	Koncentracija
Triptozni sojin bujon (TSB)	30 g/L
Kvasni ekstrakt	0,6 %
Glukoza	0,8 %

3.6.2 Izolacija proteinov in poliakrilamidna gelska elektroforeza (PAGE)

Zamrznene celice *A. actinomycetemcomitans* smo resuspendirali v 20 mL vezavega pufra (Preglednica 5) z dodatkom 0,5 mg/mL lizocima, 10 µg/ml DNaze in tabletko proteazih inhibitorjev cOmplete ULTRA Tablets, Mini, EASYpack (proizvajalca Roche). Suspenzijo smo inkubirali 45 minut pri temperaturi 4 °C in konstantnem mešanju z magetnim mešalom. Sledila je homogenizacija celic z ultrazvokom s sonifikatrojem Sonics VCX 750 (3-krat 10 sekund, 40 % moči). Pelet smo odstranili z dvakratnim centrifugiranjem (16000 g, 30 minut, 4 °C). Lizat smo do uporabe shranili na -80 °C.

Prisotnost proteinov in uspešnost pridobitve proteinskega lizata smo ugotavljali s PAGE elektroforezo na komercialnih gelih NativePAGE Novex Bis-Tris gelih. K 5, 20 in 30 µL proteinskega lizata smo dodali po 4, 7 in 10 µL 4-kratnega natrijevega dodecil sulfat (NaDS) nanašalnega pufra in za 5 minut inkubirali v vreli vodi. Proteinske vzorce smo pripravili z dodatkom 4 µL 4-kratnega NaDS nanašalnega pufra k 2 µg proteina. Te vzorce in 3,4 µL velikostnega standarda smo nanesli na gel. Elektroforeza je potekala v pufru MOPS in pri napetosti 180V. Ločevanje smo ustavili, ko se je 10 kDa marker približeval koncu gela. Sledila je 1,5 h barvanja v vodni ratopini ocetne kisline in metanola s barvila Coomassie Brilliant Blue in prekonočno razbarvanje v vodni raztopini ocetne kisline in metanola.

3.6.2 Izolacija mRNA in agarozna gelska elektroforeza

Kot izhodiščni material smo uporabili ~7 x 10^6 humanih epitelnih celic A549 (gojišče DMEM, 10 % FBS, 4mM L-glutamin; subkultivirane na 2-3 dni) in ~3 x 10^6 humanih kostnih celic MG-63 (gojišče DMEM, 10 % FBS, 4mM L-glutamin; subkultivirane na 3-4 dni). Celice so bile gojene v atmosferi s 5 % CO_2 in visoki (95 %) vlažnosti pri 37 °C. Iz celic smo izolirali mRNA s komercialnim kitom PolyATtract System 1000 z modificiranim protokolom. Celice smo zbrali v 50 mL centrifugirki in centrifugirali 5 minut pri sili 300 g. Odpipetirali smo gojišče in dodali 25 mL ledeno hladnega pufra PBS. Celice smo ponovno centrifugirali pri enakih pogojih. Nato smo sedimentu celic dodali 4 mL ekstrakcijskega pufra (ob predhodnem dodatku 164 µL β-merkaptoetanola). Celice smo razbili s 30 sekundnim vorteksiranjem. Nato smo dodali 8 mL redčitvenega pufra (ogretega na 70 °C in dodatku 164 µL β-merkaptoetanola) in 10 µL biotiliniranih Oligo(dt) sond. Raztopino smo inkubirali 5 minut na 70 °C. Sledilo je 15-minutno centrifugiranje pri radialnem pospešku 7500 g. Med tem časom smo 7 mL magnetnih kroglic, obdanih s streptavidinom, dvakrat sprali v 2 mL 0,5X pufra SSC. Magnetne kroglice smo z magneti zadržali v centrifugirki. H kroglicam smo dodali supernatant, raztopino pomešali in po dveh minutah polovili magnetke. Tekočino smo odstranili in ponovno dvakrat spirali z 2 mL 0,5X pufra SSC. mRNA smo eluirali v vodi (1 mL), odstranili magnetke z magnetnim privlakom in z dodatnim centrifugiranjem. Koncentracijo RNA smo določili z merjenjem absorbance na spektrofotometru Nanodrop 1000.

Za agarozno gelsko elektroforezo smo uporabljali 1,25 % gele: 0.5 g agaroze smo umešali v 40 mL predhodno pripravljenega 0,5-kratnega pufera TBE. Agar smo raztopili z gretjem v mikrovalovki in po delni ohladitvi dodali 0,5 µL etidijevega bromida. V kadički smo gel zalili z 0,5-kratni pufrom TBE. Nanašali smo 20 µL vzorcev (ob predhodnem dodatku nanašalnega pufra v razmerju 1:5) in 6 µL dolžinskega standarda. Elektroforezo smo izvajali 15 minut pri napetosti 110 V in toku 125 mA. Gele smo slikali z UV-transiluminatorjem in vizualizirali s programom GeneSnap.

3.6.3 Odkrivanje interakcij s površinsko plazmonsko resonanco

Površinsko plazmonsko resonanco smo merili z refraktometrom Biacore X v Infrastrukturnem centru za raziskave molekulskih interakcij na Biotehniški fakulteti Univerze v Ljubljani. Napravo smo uporabljali v skladu z navodili za delo z aparaturo (molekulske-interakcije.si, 2015). Za študije interakcij smo uporabljali vezavni pufer (Preglednica 5) pri pH 7,4. Poskuse smo izvajali pri temperaturi aparature 25 °C. Na čip SA, ki ima na površini nanešene molekule streptavidina, smo nanesli biotilinirane Oligo(dT) sonde (v skupni količini 5,5 µM). Sledil je nanos molekul mRNA in nato nanos analiziranega proteina. V preglednici 6 so prikazani ti koraki s parametri. Analizirali smo proteine CdtA, CdtB in kot pozitivno kontrolo protein imu3 (Črnigoj in sod., 2014). Redčitve smo naredili v vezavnem pufru. Izolirano RNA smo 4-kratno redčili (25 µL mRNA, 5 µL 2M NaCl in 75 µL vezavnega pufra). Asociacijo proteinov z molekulami mRNA smo spremljali 120 s pri pretoku 20 µl/min. Površino čipa smo regenerirali s 50 mM NaOH.

Preglednica 5: Vezavni pufer.

Komponenta	Koncentracija / Količina
HEPES	10mM
NaCl	140 mM
EDTA	3 mM
Mg	5 mM
Surfaktant P-20	0,005 %
miliQ voda	do 1 L

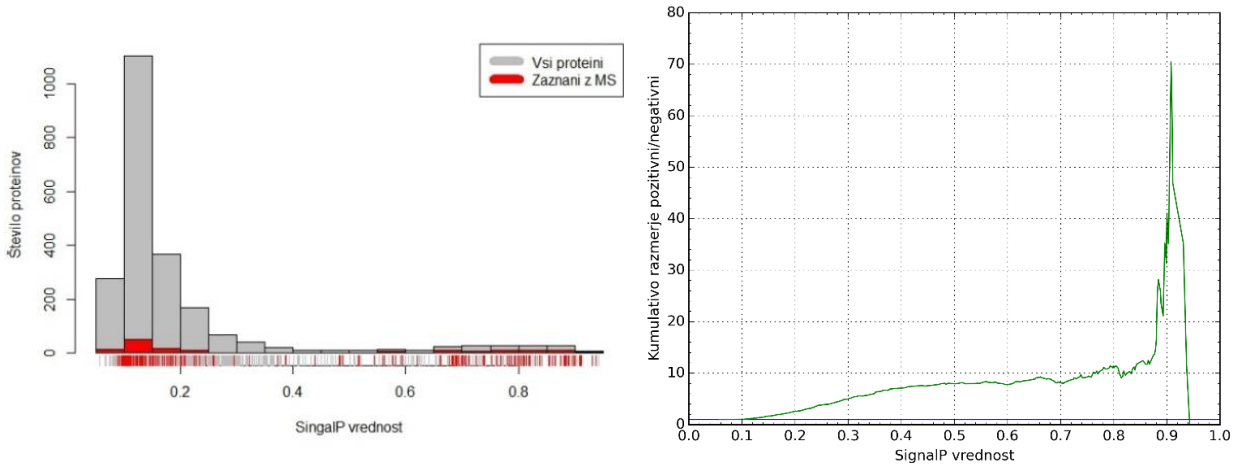
Preglednica 6: Koraki nanosa za testiranje interakcij z RNA.

Korak	Koncentracija	Pretok (Celica)	Čas pred začetkom spiranja (s)
1. Nanos biotiliniranih Oligo(dT) sond	0,5 µM in 5 µM	2 µL/min (1 in 2)	
2. Nanos RNA	~ 1,8 mM	2 µL/min (2)	
3. Nanos proteina	1 µM	20 µL/min (1 in 2)	120

4 REZULTATI

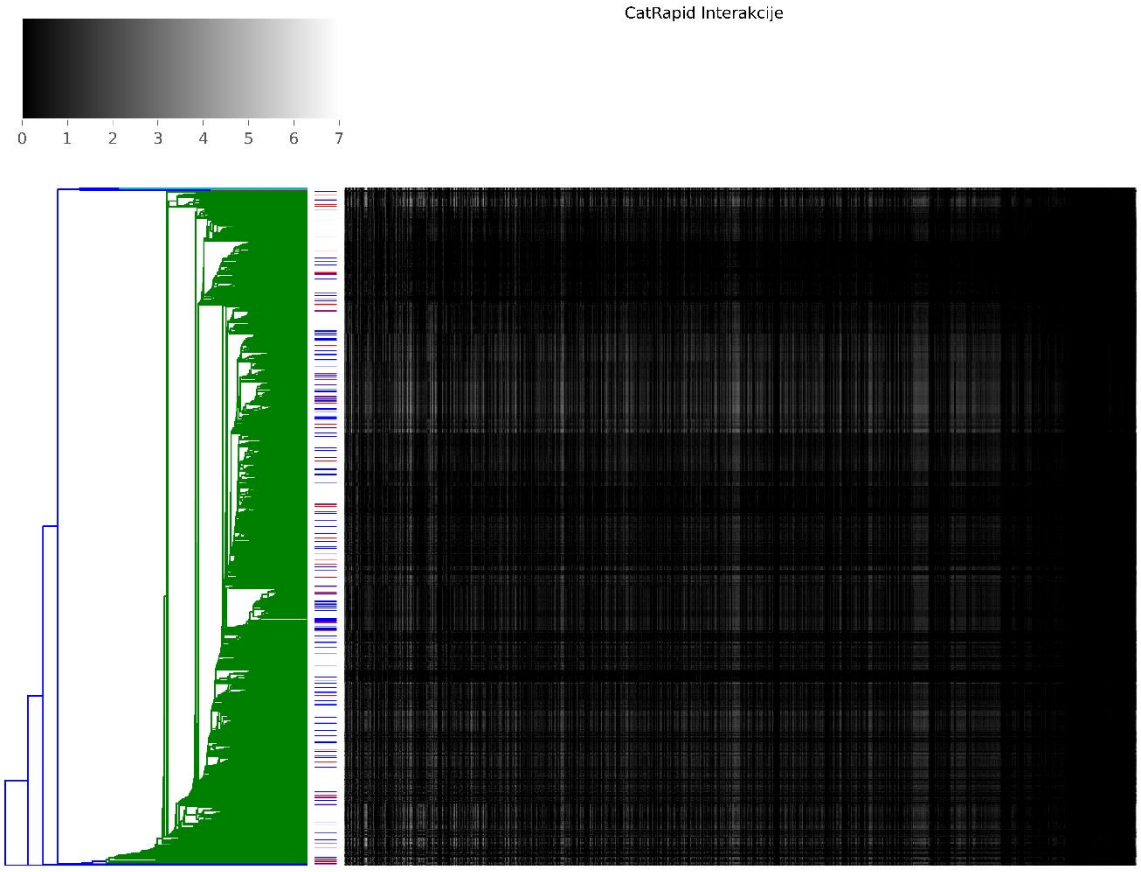
4.1 PRELIMINARNA ANALIZA PODATKOV IN ZLIVANJE PODATKOV

Pred zlivanjem podatkov nas je zanimalo, če lahko napoved prisotnosti signalnega zaporedja (SignalP) pomaga pri napovedovanju proteinov, ki se izločajo iz bakterije. V ta namen smo izrisali grafikon (Slika 5). Ugotovili smo, da lahko napoved SingalP pomaga pri določevanju sekretoma.

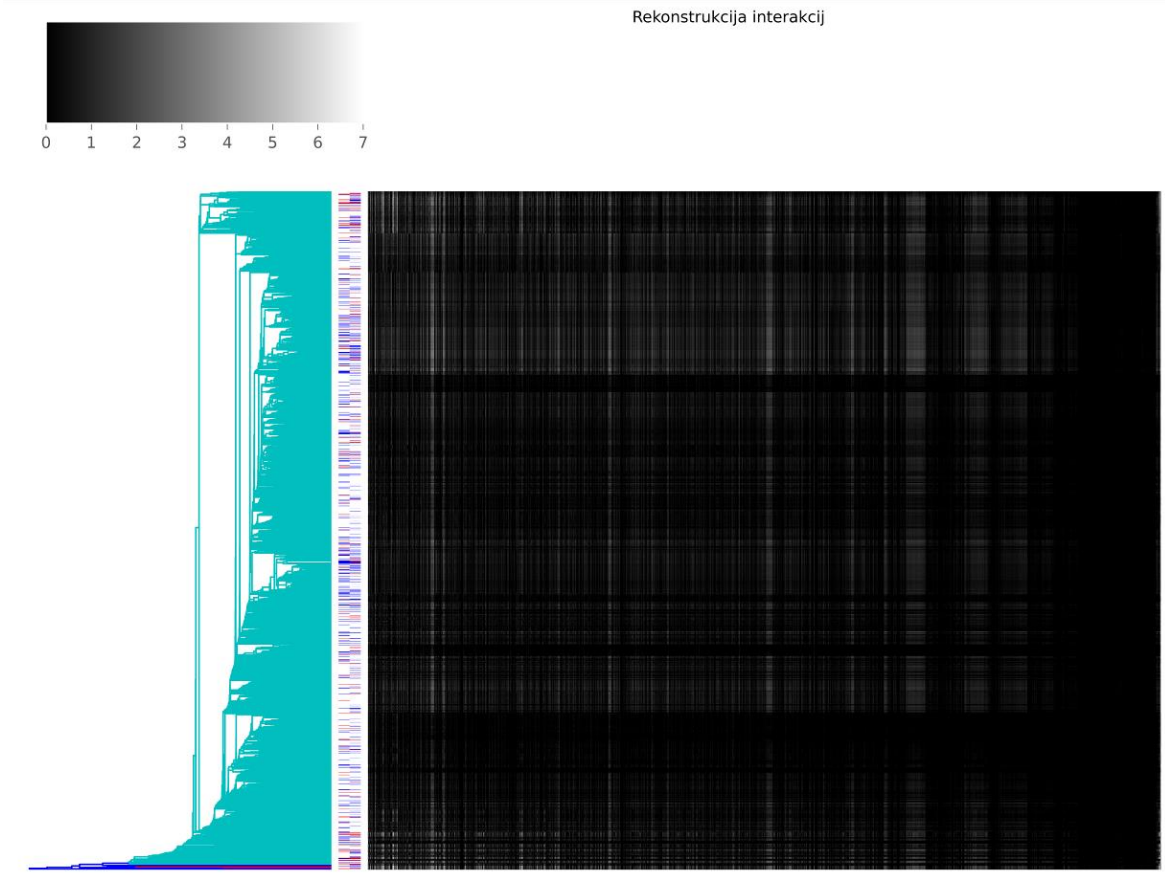


Slika 5: Porazdelitev napovedi prisotnosti signalnega peptida (SignalP; levo) in razmerje deležev proteinov v sekretomu in proteini, ki niso v sekretomu (desno). Pri določeni SignalP vrednosti smo vključili v razmerje vse proteine, ki imajo vrednost nad to mejo.

Rekonstrukcijo podatkov, vključenih v matrično faktorizacijo, smo spremljali pri 1 %, 2,5 %, 5 %, 10 %, 12,5 % in 15 % rRF. Povečanje podobnosti prvotnim podatkom z večanjem rRF je predstavljeno v prilogi A. Posebej nas je zanimalo, kako vpliva faktorizacija na centralno matriko interakcij. V ta namen smo izrisali toplotni graf (angl. *heatmap*) napovedi interakcij pred zlivanjem (Slika 6) in po zlivanju pri 5 % rRF (Slika 7). Ob primerjavi opazimo, da se bloki podobnosti v grobem ohranjajo.



Slika 6: Toplotni graf napovedih interakcij Protein A. actinomycetemcomitans (os Y) in človeška mRNA (os X) s programom catRAPID. Prag na osi Y ponazarja prisotnost RNA (rdeče) ali DNA (modro) vezavne domene.

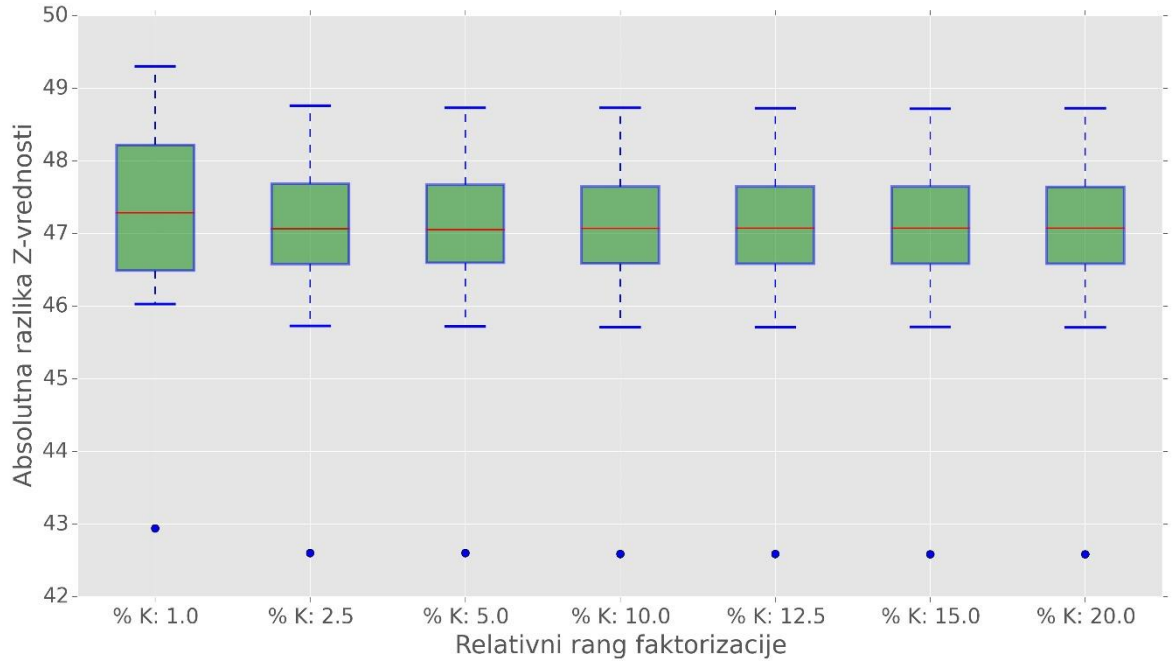


Slika 7: Toplotni graf napovedih interakcij Protein A. actinomycetemcomitans (os Y) in človeška mRNA (os X) po rekonstrukciji. Prag na osi Y ponazarja prisotnost RNA (rdeče) ali DNA (modro) vezavne domene (levo) in rekonstruirana vrednosti prisotne domene (desno).

4.2 DOLOČITEV OPTIMALNEGA RANGA FAKTORIZACIJE

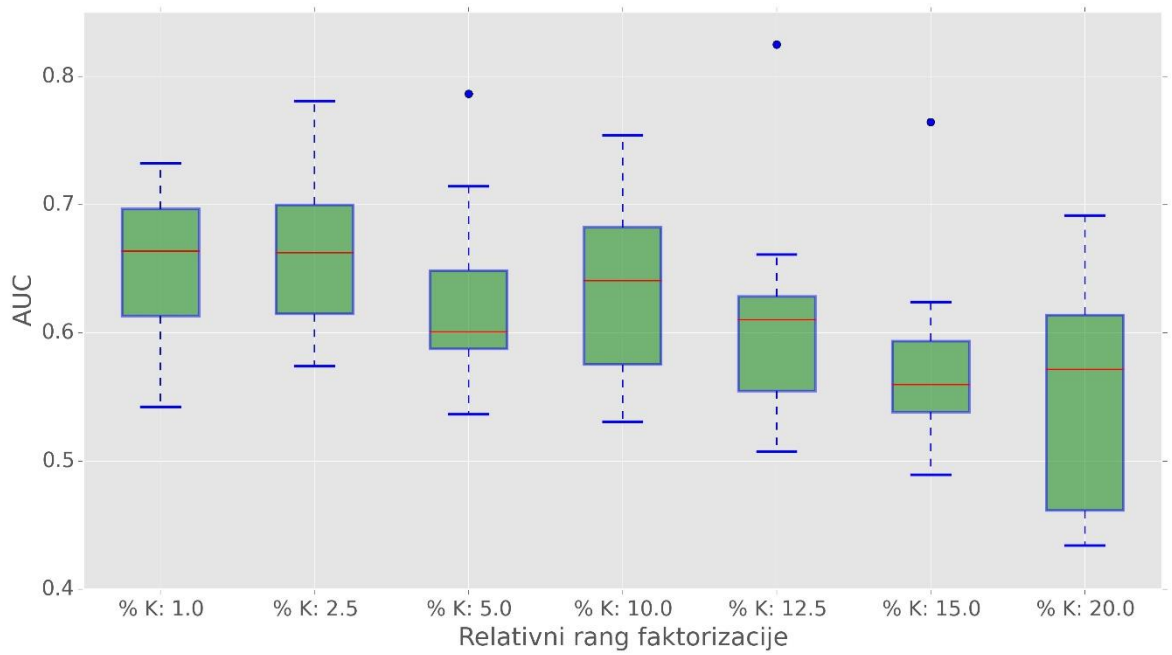
Preden smo predlagali kandidatni seznam, nas je zanimalo, kateri je optimalni rRF. Kot je prikazano v prilogi A, se z večanjem rRF podatki po rekonstrukciji vse bolj prilegajo prvotnim podatkom. Želeli smo najti optimalno vrednost kompresije, pri kateri ohranimo relevante strukture, hkrati pa odstranimo čim več »šuma« v podatkih. To vrednost smo izbrali na podlagi 10 kratnega prečnega preverjana. Pri relativnih rangih faktorizacije 1 %, 2,5 %, 5 %, 10 %, 12,5 %, 15 % in 20 % smo preverjali, kako uspešno se naš model obnaša pri napovedovanju prikritih vrednosti.

Če povprečimo absolutno razliko med prvotno Z-vrednostjo in njegovo rekonstrukcijo programa catRAPID za vsak protein (Slika 8), je pri 1 % rRF napaka v primerjavi z ostalimi rangi faktorizacije še opazna, medtem ko se med ostalimi rangi napaka bistveno ne spremeni.

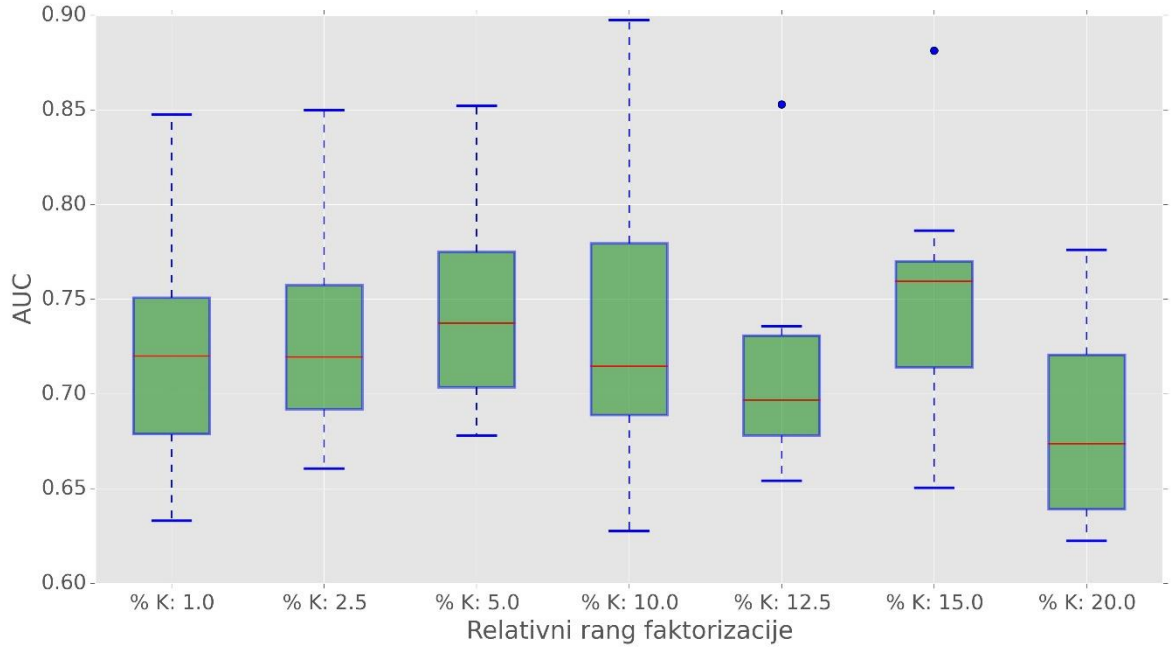


Slika 8: Absolutna razlika med povprečnimi Z-vrednostmi proteinov pred in po zlivanju.

Večje razlike med rRF opazimo pri napovedovanju sekretoma(Slika 9) in potencialnega sekretoma – proteinov s prisotnim signalnim zaporedjem (Slika 10).

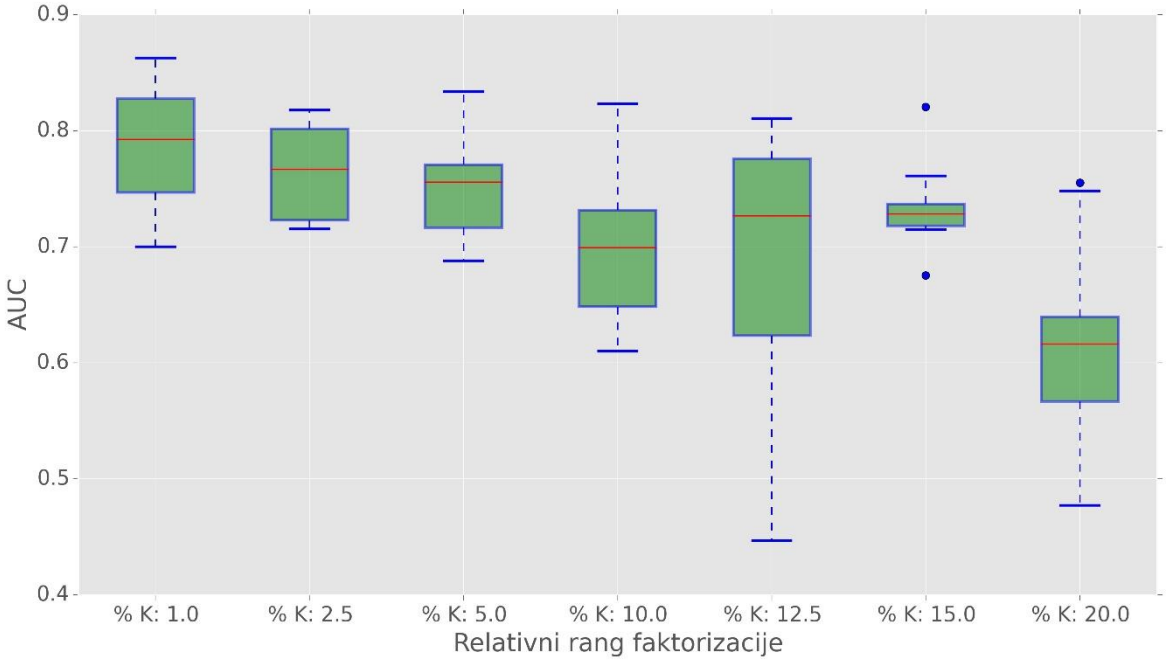


Slika 9: Napovedovanje sekrecijskih proteinov (Eksperimentalni podatki).

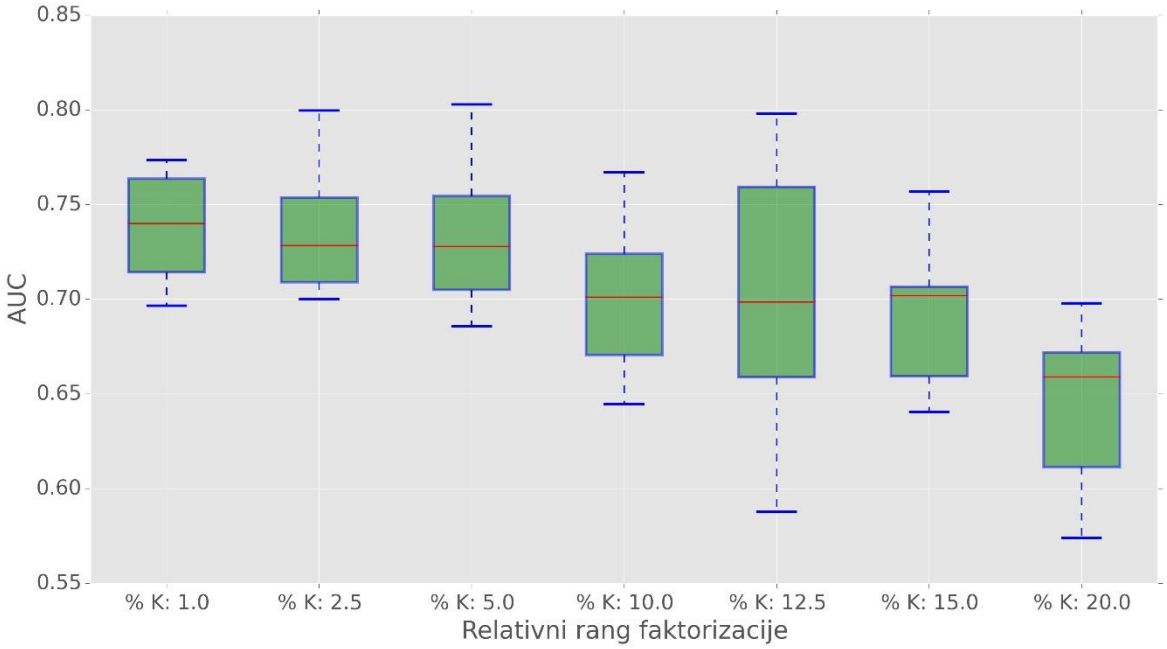


Slika 10: Napovedovanje signalnega zaporedja (SignalP).

Napovedovanje prisotnosti RNA vezavnih domen (Slika 11) in hkratno napovedovanja RNA ali DNA vezavnih domen (Slika 12) je bolj učinkovito pri nižjih (1, 2,5 in 5 %) rRF.



Slika 11: Napovedovanje prisotnosti RNA vezavnih domen.



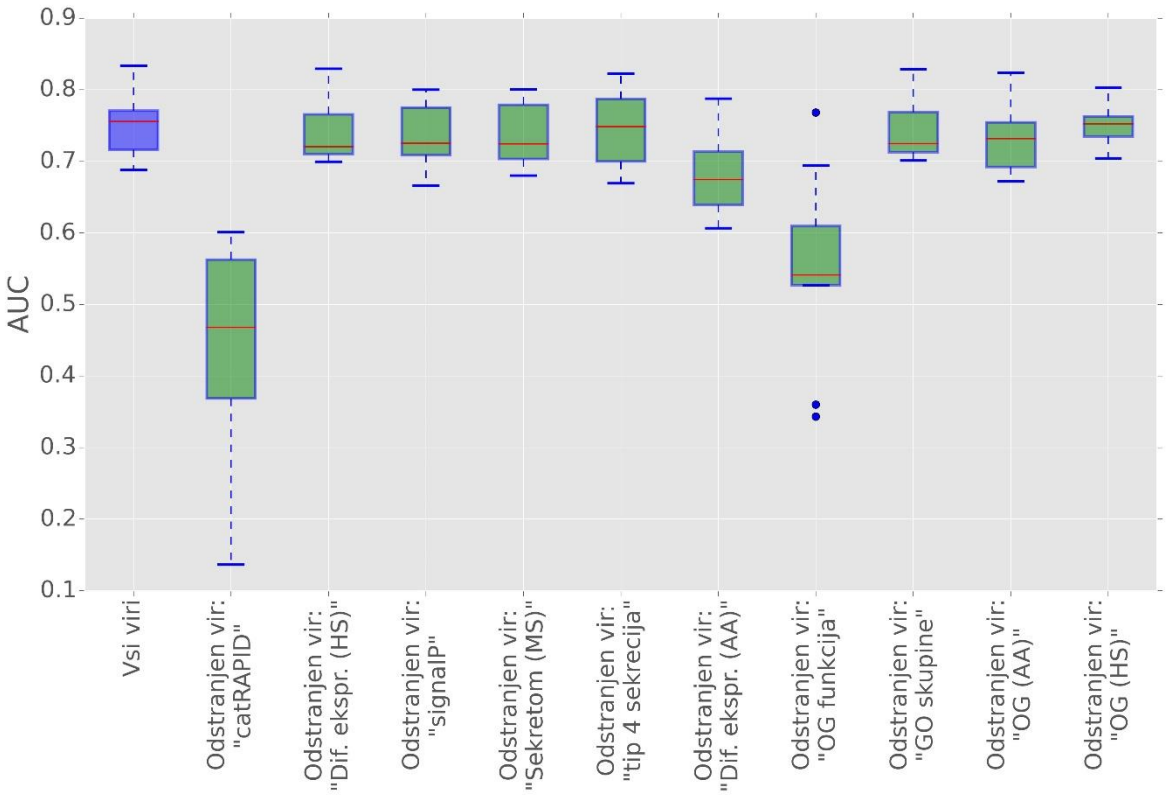
Slika 12: Napovedovanje prisotnosti RNA ali DNA vezavnih domen.

Na podlagi prikazanih podatkov smo se odločili, da bomo uporabili 5 % rRF kot osnovo za oceno informativnosti virov in tudi za sestavo kandidatnega seznama RBPjev s potencialnim vpliv na evkariontsko celico.

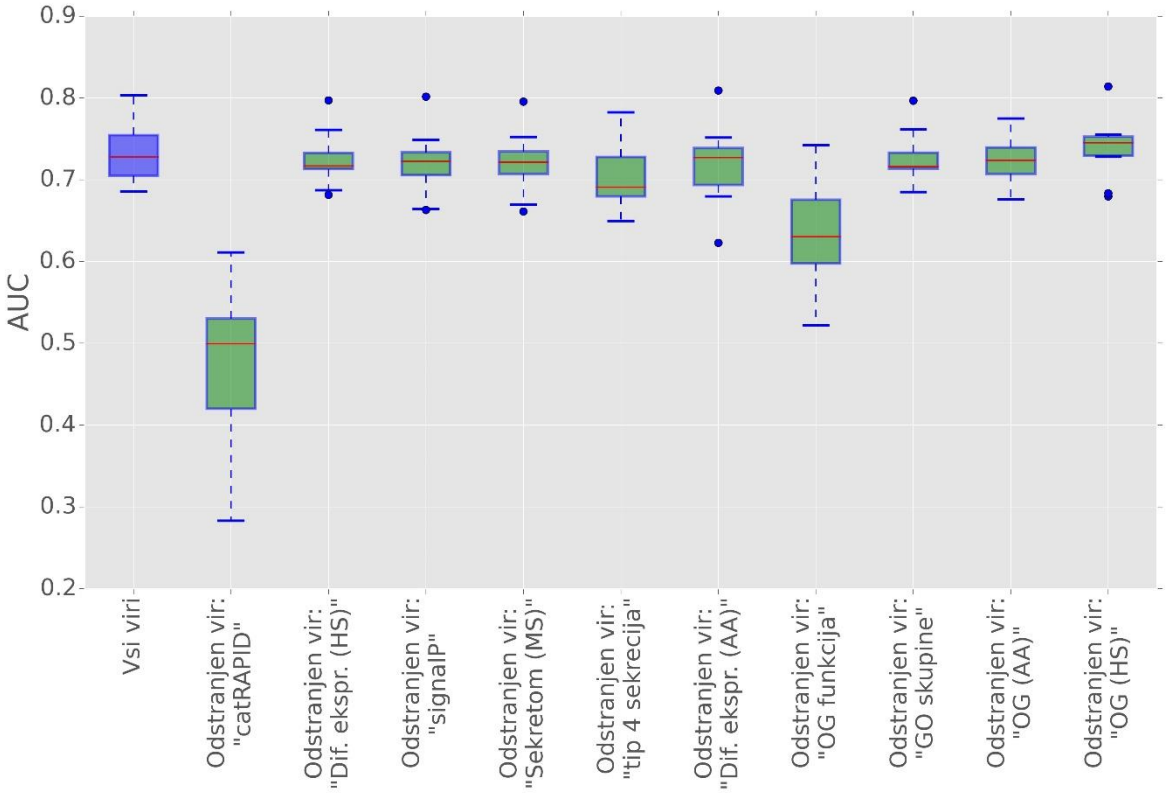
4.3 INFORMATIVNOST VIROV

Pri izbranem rRF, to je 5 %, nas je zanimalo, kako odstranitev virov vpliva na uspešnost napovedovanja izbrane relacije. Opažamo, da je ključnega pomena v vseh napovedovanjih relacija, ki povezuje proteine *A. actinomycetemcomitans* in človeške gene (relacija R1 - catRAPID). Pri odstranitvi tega vira se napovedna moč zmanjša. Domnevamo, da rezultati niso samo posledica informacij, ki jih nosi matrika, ampak tudi značilnosti matrike. Matrika ni redka, je v središču našega grafa in je po dimenzijah med večjimi. Torej, če odstranimo to relacijo, se optimizacijski postopek in konvergiranje vrednosti v latentnih matrikah bistveno spremenita.

Za napovedovanje prisotnosti RNA vezavnih domen in hkratne RNA ali DNA vezavne domene (Slika 13) je dodatno pomembna relacija funkcije ortolognih skupin (Relacija R5; »OG funkcija«), saj po odstranitvi tega vira vrednost AUC pade. Vpliv tega vira smo pričakovali, saj so med funkcijami skupin translacija, transkripcija in procesiranje RNA. To so skupine, v katerih je veliko RBP. Nekoliko nižja je tudi napovedna moč, če odstranimo vir diferencialnega izražanja *A. actinomycetemcomitans*.



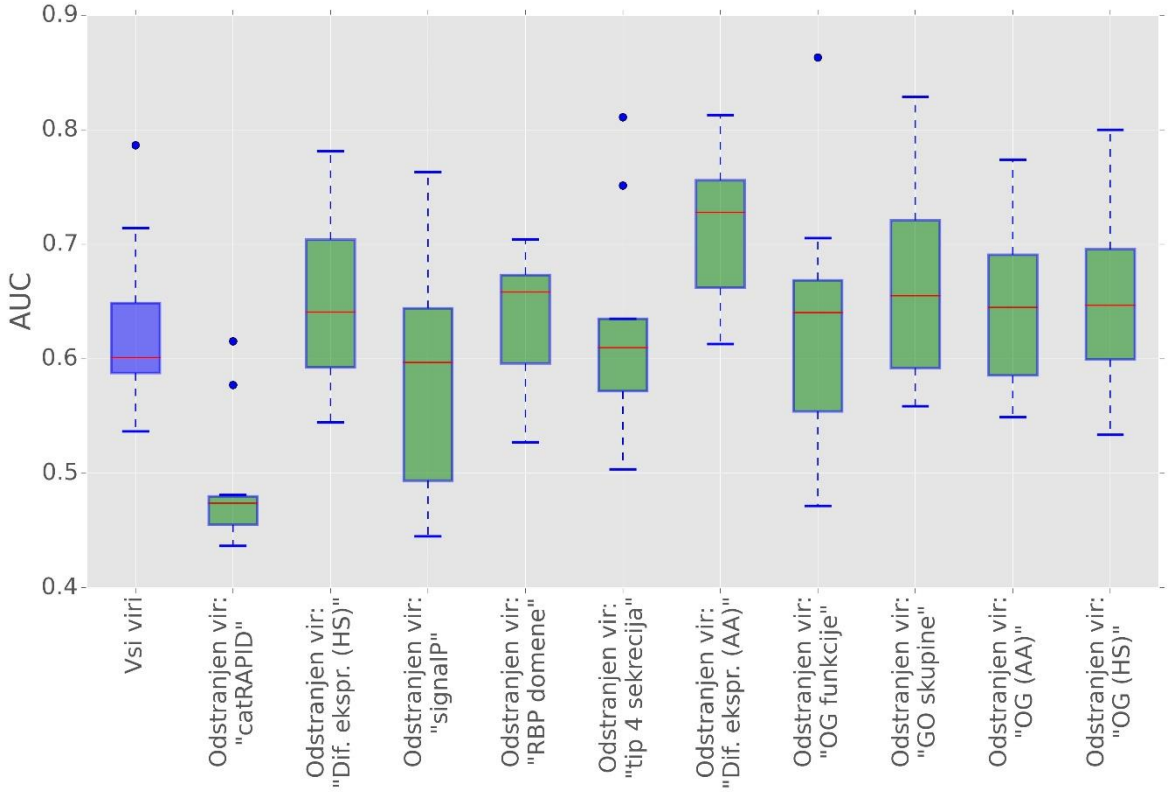
Slika 13: Uspešnost napovedovanja RNA vezavnih proteinov ob odstranitvi posameznih virov.



Slika 14: Uspešnost napovedovanja RNA in DNA vezavnih proteinov ob odstranitvi posameznih virov.

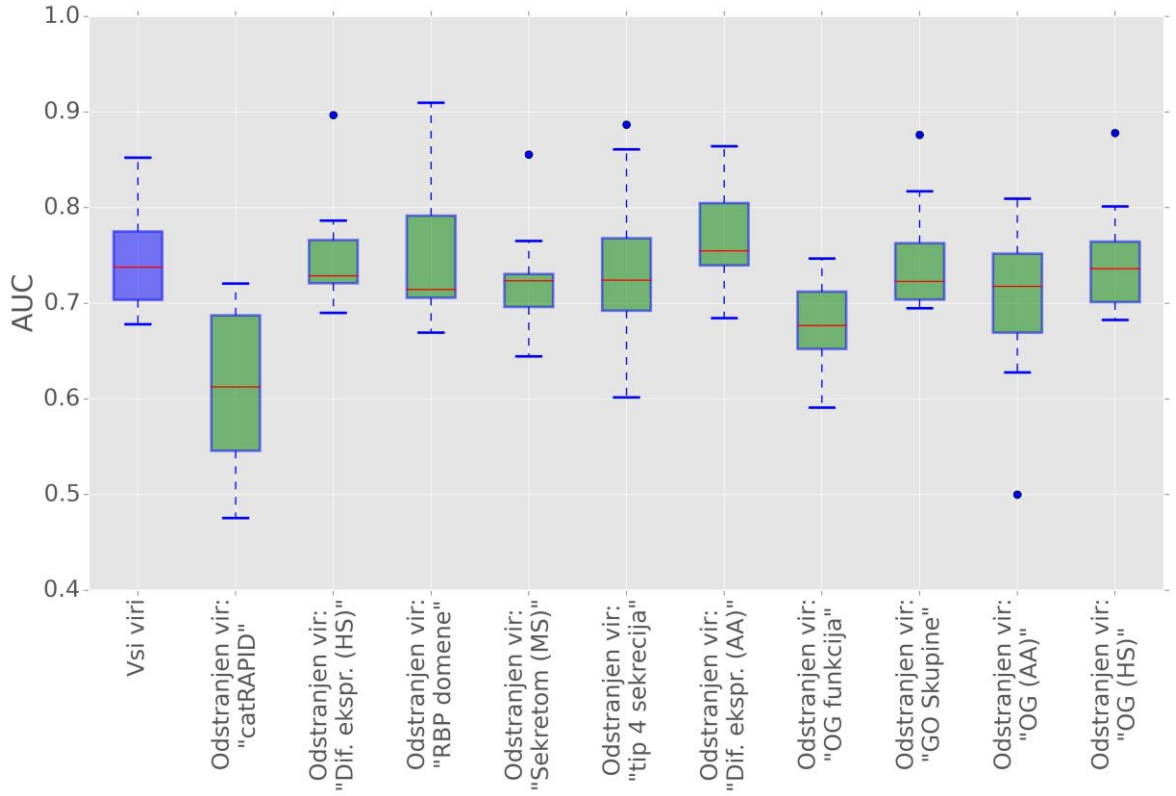
Napovedovanje proteinov v sekretomu se je izkazalo za zahtevnejši problem, saj so AUC vrednosti nižje (Slika 15) kot v primeru napovedovanja domen.

Razen odstranitve relacije »catRAPID«, odstranitev drugih relacij bistveno ne vpliva na rezultate. Do odstopanja pa lahko pride tudi zaradi hevristične narave metode. Dobljeni rezultati niso nujno optimalni minimum, ampak na končni rezultat vplivajo tudi začetne vrednosti v latentnih matrikah in nastavitve optimizacijskih parametrov.



Slika 15: Uspešnost napovedovanja Sekretoma ob odstranitvi posameznih virov.

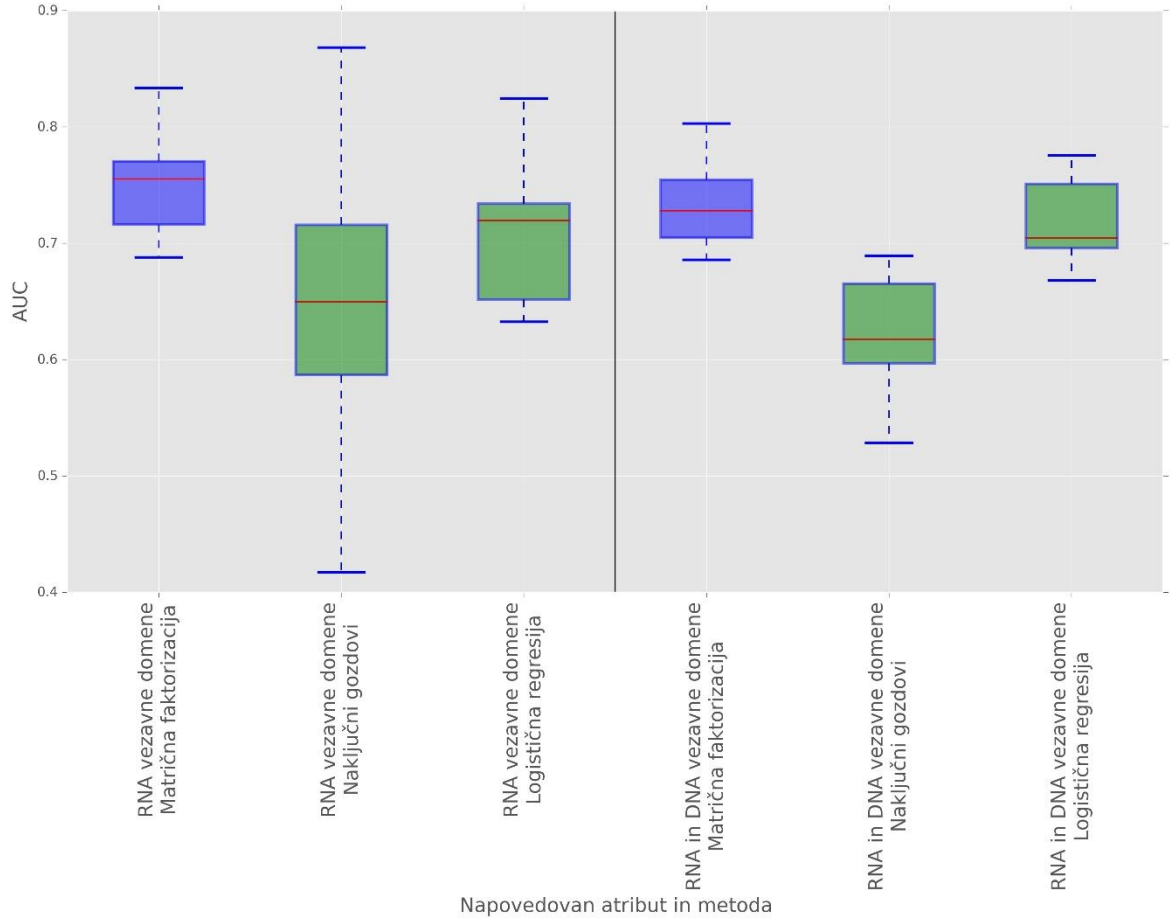
Slika 16 prikazuje, kako uspešno smo rekonstruirali napoved prisotnosti signalnega zaporedja. Tudi v tem primeru odstranitev posameznega vira, razen vira »catRAPID«, bistveno ne vpliva na pridobljene rezultate.



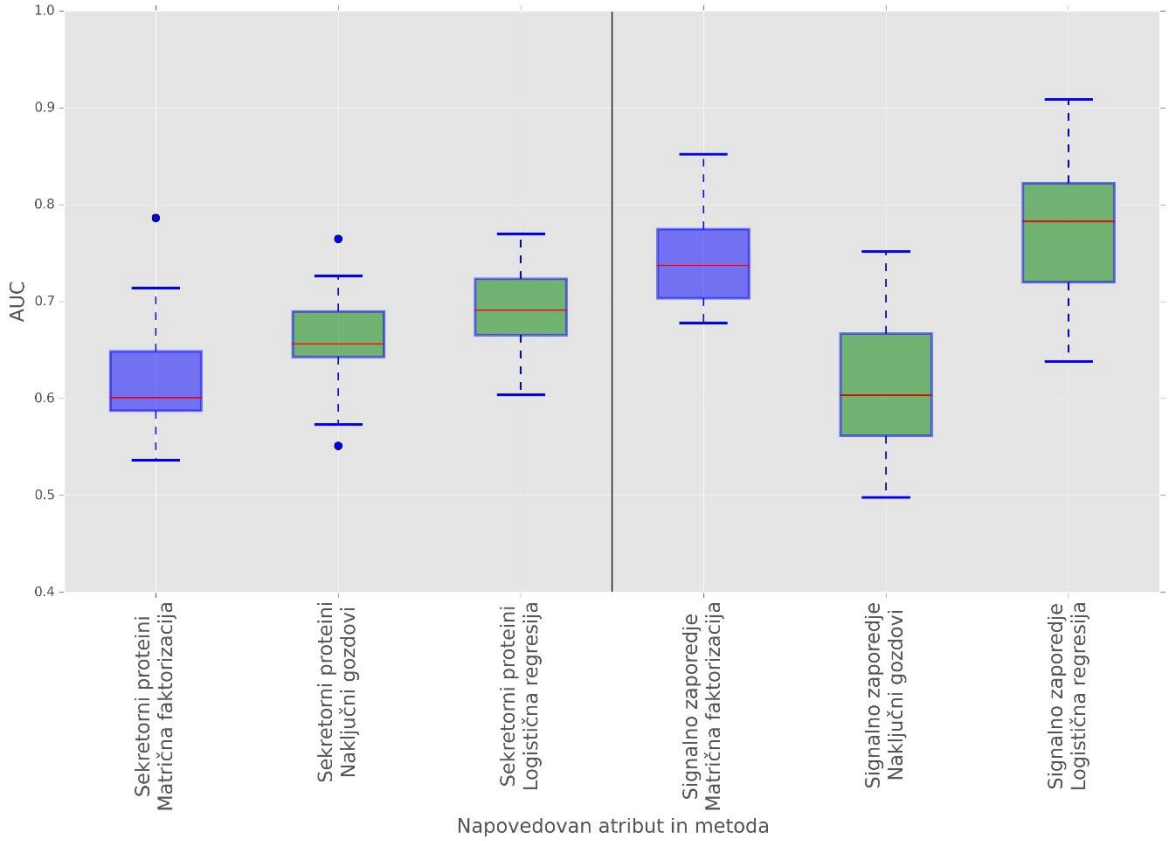
Slika 16: Uspešnost napovedovanja (rekonstrukcije) prisotnosti signalnega zaporedja.

4.4 PRIMERJAVA ZLIVANJA PODATKOV Z DRUGIMI METODAMI STROJNEGA UČENJA

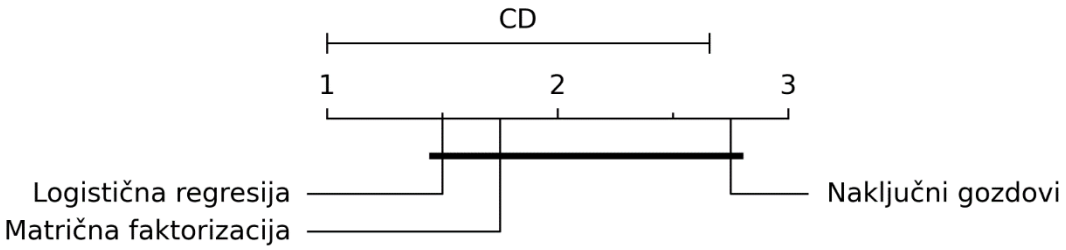
Napovedovanje prisotnosti RNA vezavnih domen in hkratno napovedovanje RNA in DNA vezavnih domen je z metodo matrične faktorizacije glede na mediano AUC boljše od metod naključnih gozdov in logistične regresije (Slika 17). Pri napovedovanju proteinov v sekretomu in pri rekonstrukciji prisotnosti signalnega zaporedja (SignalP) se je metoda logistične regresije, gledano na mediano AUC vrednosti, izkazala za nekoliko boljšo (Slika 18). Nemenyi test (Slika 19) ni pokazal, da bi katera od metod bila statistično značilno boljša od ostalih dveh pri napovedovanju več atributov.



Slika 17: Napovedovanje prisotnost (ribo)nukleinskih vezavnih domen.



Slika 18: Napovedovanje sekrecije in signalnega zaporedja.



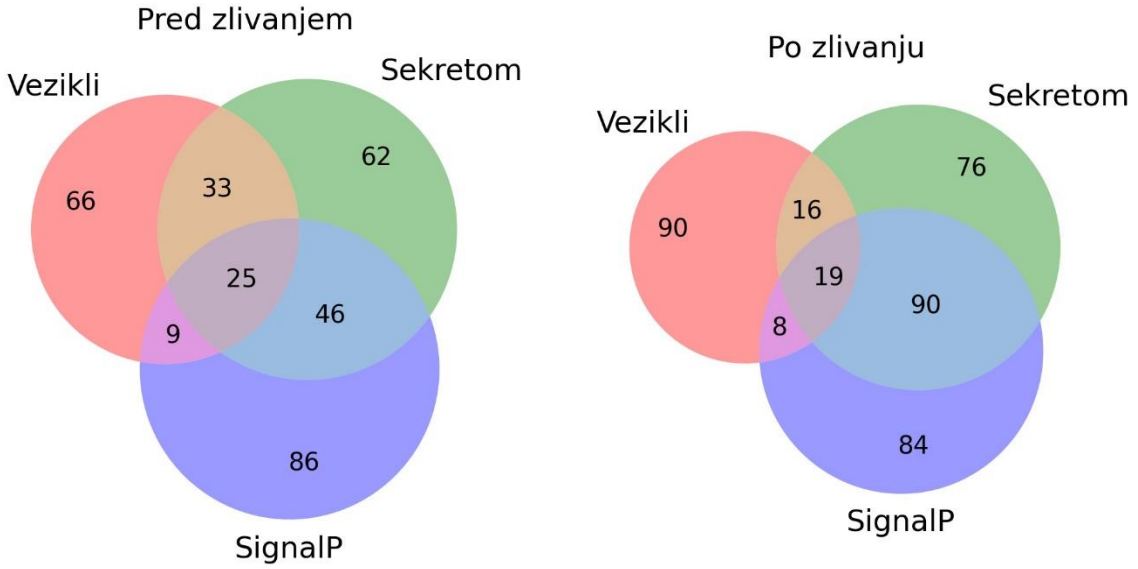
Slika 19: Primerjava klasifikatorjev. Merilo CD predstavlja kritično razdaljo za signifikantno razliko (Nemenyi test, α=0,05).

4.5 PRIORITETNI SEZNAM KANDIDATNIH RBP

4.5.1 Izbor glede na izločanje proteinov

Preseki med množicami napovedanih sekretornih proteinov (SignalP), eksperimentalno določenih proteinov (Sekretom) in množico proteinov v veziklih (Vezikli), ki je nismo vključili v zlivanje, so relativno majhni (Slika 20). Le malo več kot tretjina (35 %) proteinov iz študije sekretoma se prekriva z rezultati iz lizosomov. Če primerjamo unijo množic Sekretom in SignalP z množico Vezikli, je prekrivanje še manjše (26 %). Kljub temu pa porazdelitev ne izgleda naključna, saj je v primeru hipergeometrične porazdelitve verjetnost $P(X \geq 67)$ pri skupno 2001 proteinih zelo majhna ($1{,}045 * 10^{-27}$).

Če primerjamo prekrivanje množic po fuziji in rekonstrukciji (Slika 20 desno) z množico Vezikli, ugotovimo, da je presek še manjši. Pri tem je 127 proteinov enakih kot v prvotnih množicah, 166 pa jih je na novo napovedanih kot del sekretoma oziroma kot potencialni kandidat za sekrecijo. Naša metoda je bila manj uspešna pri napovedovanju novih proteinov, ki bi lahko bili prisotni v veziklih. Pri postavljenih pogojih smo napovedali 8 proteinov, ki prvotno niso bili v množici Sekretom oziroma SignalP, vendar so v množici Vezikli. Hkrati pa smo vključili 158 takih, za katere ni dokazov o sekreciji. V napovedovanju dodatnih proteinov, ki so v veziklih, bi tako naša metoda bila boljša od naključnega izbiranja v 71 % primerih.

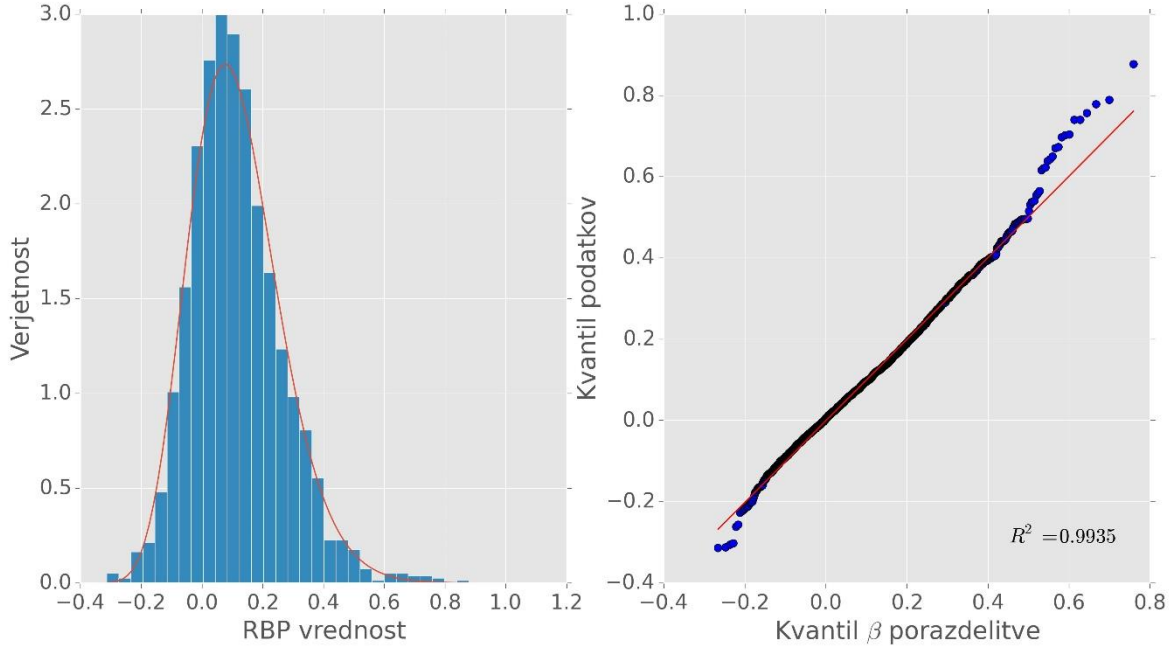


Slika 20: Vennov diagram za izločene proteine in napovedi.

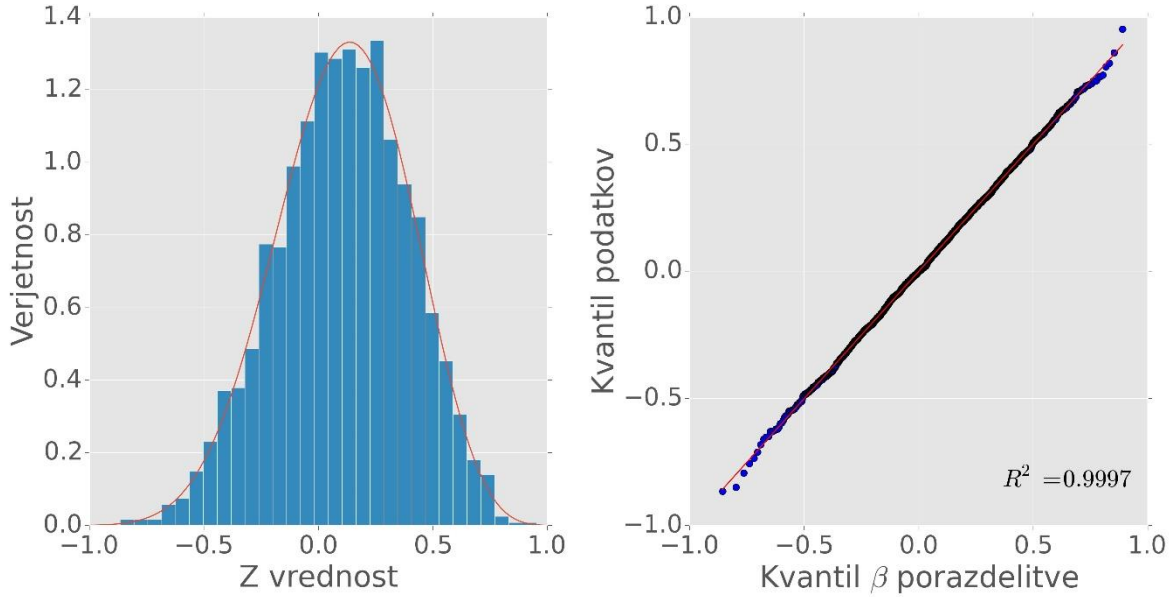
Zaradi slabših rezultatov prečnega preverjanja in manjšega preseka množic SignalP in Sekretom po zlivanju z množico Vezikli smo se odločili, da bomo v končni seznam kandidatnih RBPjev predlagali proteine iz treh prvotnih množic pred zlivanjem (SignalP, Vezikli in Sekretom).

4.5.2 Rangiranje glede na sposobnost vezave nukleinskih kislin

Normalno in beta porazdelitev smo prilegali podatkom, napovedanih RNA vezavnih domen in povprečni jakosti vezave proteinov po zlivanju. Tako v primeru RNA domen (Slika 21), kot tudi v primeru jakosti vezave (Slika 22), se je porazdelitev beta zaradi repov izkazala za ustreznejšo (Preglednica 7).



Slika 21: Porazdelitev rekonstrukcije RBP domen (levo) in kvantil-kvantil diagram (desno) v primeru beta porazdelitve.



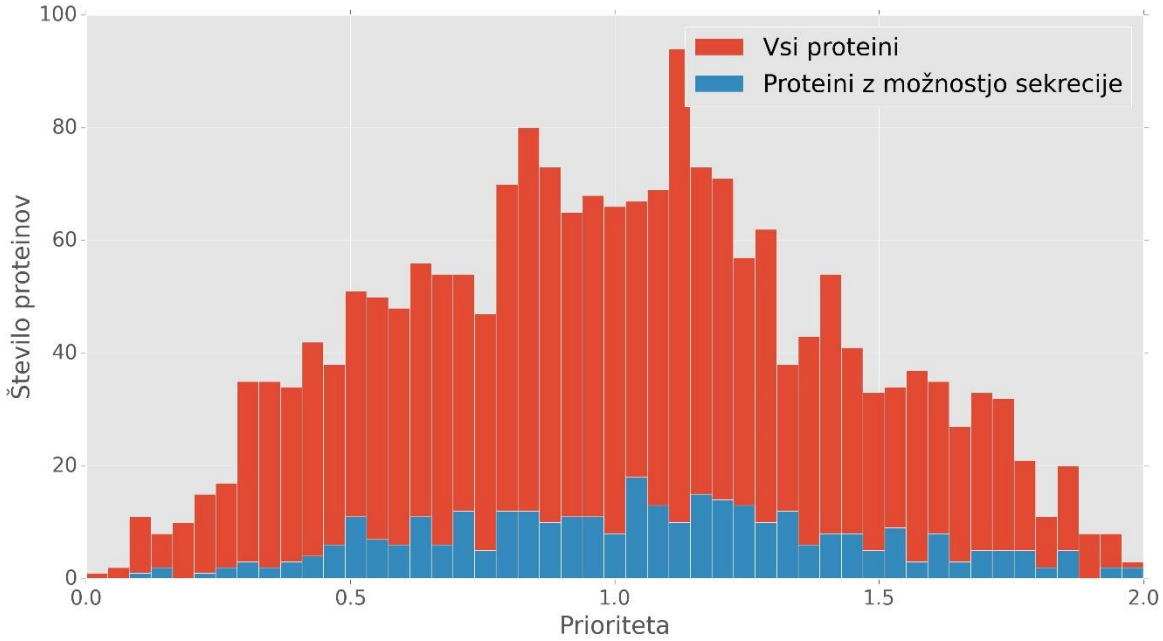
Slika 22: Porazdelitev rekonstrukcije povprečne Z vrednosti (levo) in kvantil-kvantil diagram (desno) v primeru beta porazdelitve.

Preglednica 7: Prileganje porazdelitve podatkom (p-vrednost Kolmogorov–Smirnov testa).

	Normalna porazdelitev	**Beta porazdelitev**
RNA vezavne domene	0,0002	0,8120
Povprečna Z-vrednost	0,2222	0,8841

4.5.3 Prioritetni seznam RBP z možnostjo zunajceličnega delovanja

Vsoti zbirnih funkcij verjetnosti iz prejšnje točke smo uporabili za prioritizacijo RNA vezavnosti. Nadaljnje smo izbrali le proteine, ki so se bili zaznani v sekretomu ali veziklih oziroma imajo napovedano signalno zaporedje (Slika 23). Najvišje rangirani proteini so predstavljeni v preglednici 8.



Slika 23: Porazdelitev vseh proteinov in tistih z možnostjo sekrecije glede na prioriteto po predlagani meritveni funkciji.

Preglednica 8: Napovedanih 20 najboljših RNA vezavnih proteinov, ki se (domnevno) izločajo. Pri izločanju oznaka S označuje določeno signalno zaporedje, M označuje sekretom in V označuje vezikle.

Protein (oznaka gena)	**Rang**	**RNA/DNA vezavna domena**	**Vrednost napovedane RNA domene**	**Vezavnost (Povprečna Z-vrednost)**	**Izločanje**	**Napoved sekrecije po zlivanju**	**Podobnost z *E. coli K-12* (% AA)**	**Podobnost z *E. coli K-12* (Percentil)**
DNA vezavni protein HU (D7S_00989)	1	DNA	0,994150	0,982545	V	DA	80,00	0,9620
30S ribosomalni protein S20 (D7S_01849)	3	DNA	0,973502	0,989272	M	DA	74,71	0,9160
Ribosomalni protein S1 (D7S_00172)	4	RNA	0,961256	0,974684	V	NE	77,92	0,9460
IHF podenota alfa (D7S_00047)	11	RNA	0,972960	0,947795	M	NE	67,68	0,8350
Protein TadG pilusa Flp (D7S_01444)	20	DNA	0,915399	0,961323	MV	NE	19,58	0,0185
Podenota A eksonukleaze ABC (D7S_00635)	29	0	0,940722	0,916472	SV	DA	21,83	0,1205
Domnevna L-asparaginaza (D7S_01893)	33	0	0,913081	0,935549	SM	DA	68,48	0,8460
Šaperon DnaK (D7S_02078)	34	DNA	0,908355	0,938434	MV	DA	84,17	0,9765
50S ribosomalni protein L7/L12 (D7S_01707)	36	0	0,881725	0,963947	M	NE	78,05	0,9475
Monosaharid-transportirajoča ATPaza (D7S_00956)	40	DNA	0,963968	0,868893	SM	DA	34,89	0,4265
Septacijski protein A (D7S_01522)	49	0	0,846691	0,951398	P	DA	29,32	0,3725
Košaperon GrpE (D7S_01532)	52	DNA	0,900358	0,893182	M	NE	42,79	0,5055
Protein lipoproteina LppC (D7S_00457)	54	0	0,932086	0,855552	MV	DA	31,93	0,4025
F0F1 ATP sintetaza, podenota alfa (D7S_01240)	57	0	0,801588	0,977240	MV	NE	86,55	0,9840
Protein SmpA (D7S_01009)	67	0	0,892984	0,867946	S	NE	35,51	0,4340
DNA polimeraza I (D7S_01384)	70	RNA	0,956023	0,799438	M	NE	63,42	0,7855
30S ribosomalni protein S16 (D7S_00835)	75	0	0,940064	0,809586	M	DA	76,83	0,9350
Dejavnik obnove ribosoma (D7S_00932)	85	DNA	0,856221	0,878775	M	NE	67,57	0,8330
Tiol:disulfid izmenjevalni protein DsbD (D7S_00679)	86	0	0,834887	0,900019	S	NE	46,99	0,5600
TonB-odvisni siderofor receptor (D7S_02228)	94	0	0,769576	0,957878	SV	DA	27,66	0,3500

4.5.4 Pregled literature za sposobnost vezave RNA najvišje uvrščenih proteinov

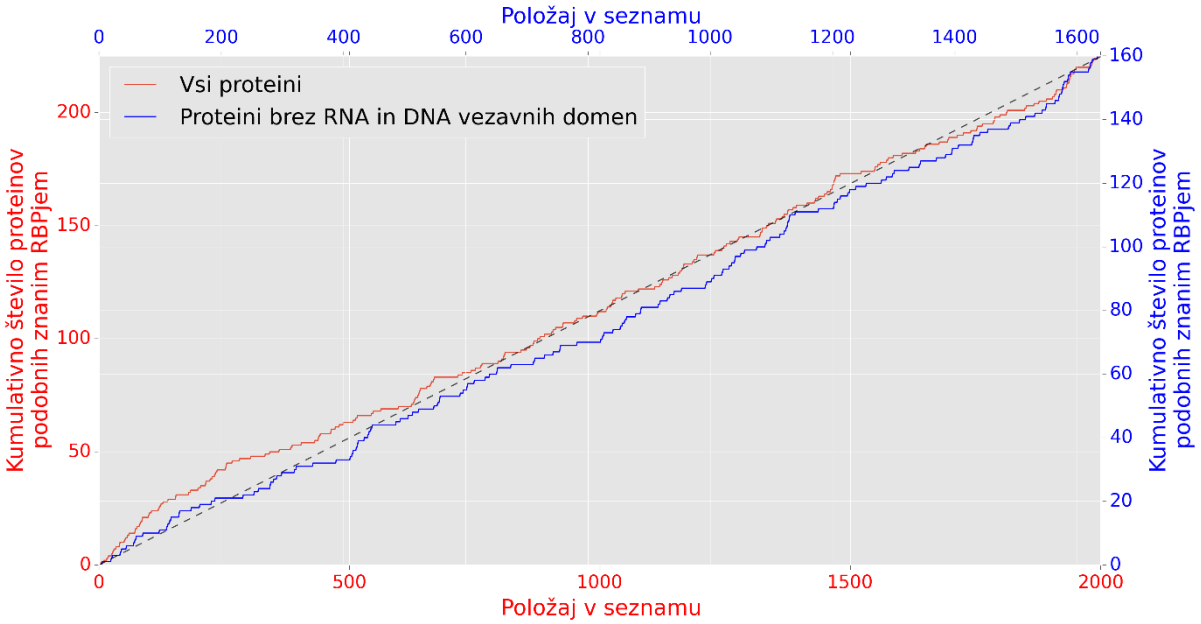
S pregledom literature smo ugotovili, da lahko za vsaj prvih 6 proteinov z najvišjo vsoto zbirnih funkcij verjetnosti bolj ali manj posredno povežemo z vezavo RNA. Pri tem imata le dva proteina klasične RNA vezavne domene (Preglednica 9).

Preglednica 9: Pregled literature za sposobnost vezave RNA 6 najvišje uvrščenih proteinov.

Protein (oznaka gena)	Vezavna domena	Kazatelj vezave RNA	Vir
DNA vezavni protein HU (D7S_00989)	DNA	Protein iz *E. Coli* veže dvoverižno RNA.	Balandina in sod., 2002
N6-adenin-specifična DNA metiltransferaza (D7S_01500)	RNA	Človeški homologni protein TRDMT1 metilira tRNAAsp.	Goll in sod., 2006
30S ribosomalni protein S20 (D7S_01849)	DNA	Ribosom brez podenote ni sposoben vezave molekul mRNA.	Tobin in sod., 2010
ribosomal protein S1 (D7S_00172)	RNA	Ribosomski protein ima RNA vezavno domeno.	/
Protein s ponovitvijo WD40 (D7S_01764)	0	Proteini z več ponovitvami WD40 imajo sposobnost vezave RNA.	Lau in sod., 2009
ATPaza vključena v popravo poškodb DNA, domnevni transmembranski protein (D7S_00521)	DNA	Protein ima domeno podobno HNH endonukleazni domeni. Za te endonukleaze je bilo dokazano, da lahko tvorijo ribonukleoproteinske komplekse.	Zimmerly in sod., 1995

4.5.5 Obogatitvena analiza prioritetnega seznama: podobnost z znanimi sesalskimi RBP

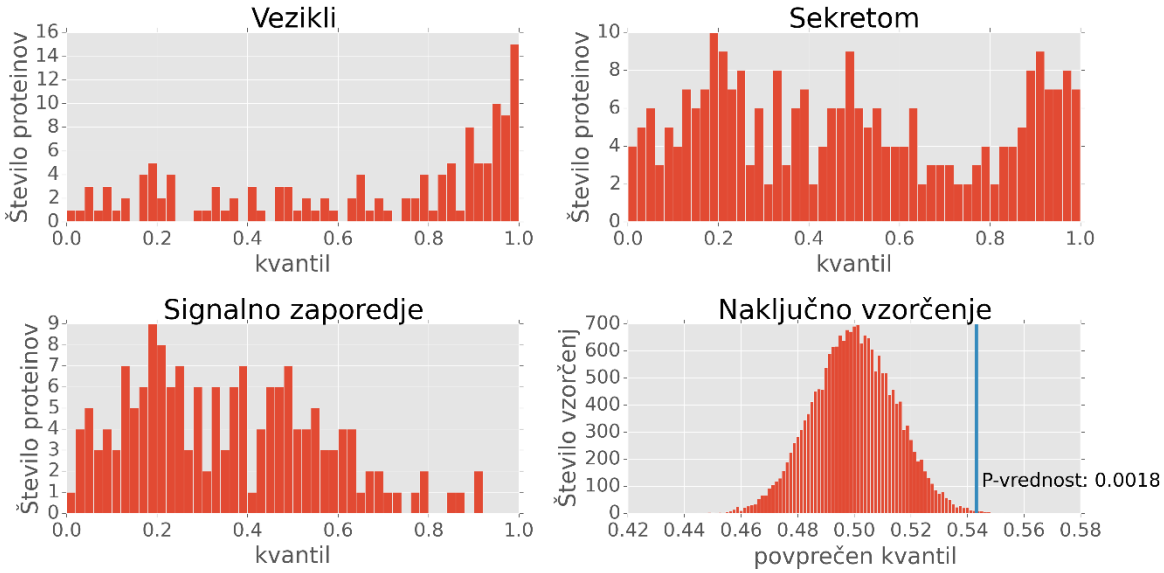
Proteini, podobni sesalskimi RBPjem, glede na prioritetni seznam niso značilno obogateni na začetku seznama (Slika 24). Če upoštevamo vse proteine, je med prvimi 10 % proteini na seznamu 34 od skupno 225 proteinov, ki so podobni RBPjem. Na sliki pa je opazno rahlo naraščanje odstopanja od naključne porazdelitve (diagonale) do ~250. mesta. Če upoštevamo samo proteine, ki nimajo RNA veznih domen, je med najvišje urejenimi 10 % le 18 od skupno 160 podobnih proteinov. Odstopanja od diagonale navzgor ne opazimo.



Slika 24: Podobnost med proteini A. actinomycetemcomitans in eksperimentalno določenim RBP.

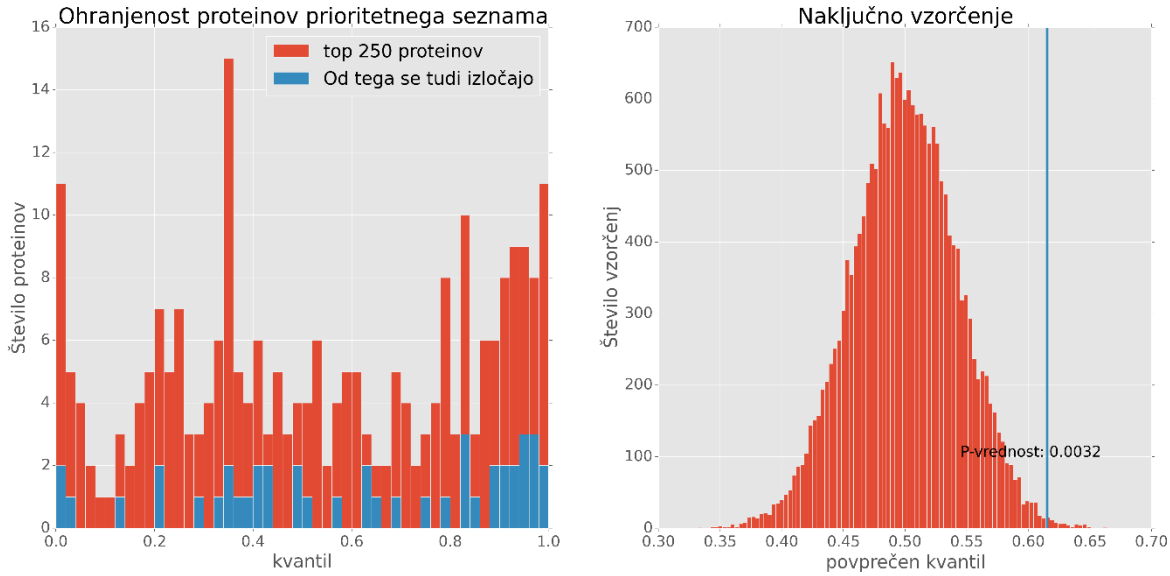
4.5.6 Podobnost proteinov *A. actinomycetemcomitans* z bakterijskimi proteini *E. coli*

Zanimalo nas je, kako podobni so proteini oziroma določene skupine proteinov *A. actinomycetemcomitans* s proteini bakterije *E. coli*. S tem smo hoteli preveriti, ali smo prioritizirali neke splošne bakterijske proteine ali pa specifične proteine bakterije *A. actinomycetemcomitans*. V ta namen smo proteine glede na podobnost razvrstili v kvantile. Pri izvenceličnih proteinih opažamo, da so tisti, ki so bili najdeni v veziklih, najbolj ohranjeni, medtem ko so proteini s signalnim zaporedjem manj ohranjeni (Slika 25). Kljub temu so v splošnem izvencelični proteini s povprečnim kvantilom 0,543 bolj ohranjeni, kot bi pričakovali ob naključnem izboru skupine 327 proteinov (Slika 25 desno spodaj).



Slika 25: Ohranjenost proteinov prisotnih v veziklih, v sekretomu in tistih, ki imajo določeno signalno zaporedje. Levo spodaj je prikazana pričakovana porazdelitev ob naključnem vzorčenju.

Tudi če omejimo na vrhnjih 250 proteinov z našega seznama in od teh izberemo zunajcelične proteine, to je 44 proteinov, vidimo, da so ti s povprečnim kvantilom 0,615 nadpovprečno ohranjeni (Slika 26).



Slika 26: Ohranjenost proteinov iz prioritetnega seznama za katere verjamemo, da so zunajcelični (levo) in pričakovana porazdelitev ohranjenosti ob naključnem vzorčenju (desno).

4.5.7 Obogatitvena analiza: GO

Preverili smo obogatenost vrhnjih 50 proteinov z našega seznama RNA vezavnih proteinov. Pri 5 % FDR je obogatenih 12 skupin (Preglednica 10), pri 25 % FDR pa jih je 27. Med temi 50 proteini smo osmim z Blast2GO anotacijo določili molekularno funkcijo vezave RNA. Če analiziramo vrhnjih 50 kandidatov s seznama, kjer so izvzeti proteini z odkritimi RNA in DNA vezavne domenami, pri 25 % FDR ne odkrijemo nobenih obogatenih skupin. V tem primeru je najboljše rangirana skupina eksonukleazna aktivnost (p-vrednost: 0,0059).

Nadalje smo analizirali vrhnjih 20 proteinov z našega prioritetnega seznama, za katere domnevamo, da se izločajo (iz preglednice 8). Pri 25 % FDR smo ugotovili, da med obogatene skupine spadajo biološki procesi translacije in njegove starševske skupine (Preglednica 11). Če iz analize izključimo gene povezane s translacijo (Preglednica 12), imajo GO skupine manjši presek z našim spremenjenim seznamom. S tremi proteini v preseku je na vrhu seznama skupina šaperonov. Preverili smo tudi, kakšna bi bila obogatitev skupin, če ne bi upoštevali RNA vezavnih domen ampak samo povprečno moč vezave, napovedano z catRAPID, proteinov (Priloga B). V tem primeru vrhnja skupina proteinov spremeni.

Preglednica 10: Obogatenost GO skupin (pri 5 % FDR) za vrhnjih 50 napovedanih RNA veznih proteinov. Okrajšave: BP - biološki proces, MF – molekulska funkcija.

Skupina	Opis	Domena	Št. genov skupine	Št. genov v preseku	p-vrednost	q-vrednost
GO:0003676	Vezava nukleinskih kislin	MF	327	24	1,22E-07	0,000341
GO:1901363	Vezava heterocikličnih skupin	MF	606	30	1,00E-05	0,005455
GO:0097159	Vezava organskih cikličnih spojin	MF	606	30	1,00E-05	0,005455
GO:0043170	Proces metabolizma makromolekul	BP	539	28	1,05E-05	0,005455
GO:0044260	Proces metabolizma celičnih makromolekul	BP	476	26	1,14E-05	0,005455
GO:0003677	Vezava DNA	MF	176	15	1,17E-05	0,005455
GO:0043565	Vezava specifičnega zaporedja DNA	MF	25	6	2,21E-05	0,008829
GO:0034645	Celični proces biosinteze makromolekul	BP	238	17	2,85E-05	0,00995
GO:0090304	Proces metabolizma nukleinskih kislin	BP	296	19	3,77E-05	0,010099
GO:0006259	Proces metabolizma DNA	BP	107	11	3,94E-05	0,010099
GO:0009059	Proces biosinteze makromolekul	BP	244	17	3,98E-05	0,010099
GO:0034641	Celični proces metabolizma dušikovih spojin	BP	513	26	4,81E-05	0,011191

Preglednica 11: Obogatenost GO skupin (pri 25 % FDR) za vrhnjih 20 proteinov glede na napovedano RNA vezavo in hkratno sekrecijo. okrajšave: BP - biološki proces.

Skupina	Opis	Domena	Št. genov v skupini	Št. genov v skupini	p-vrednost	q-vrednost
GO:0043603	Celični metabolni proces amidov	BP	120	7	9,46E-05	0,1830
GO:0044267	Celični metabolni proces proteinov	BP	176	8	0,000153	0,1830
GO:0006412	translacija	BP	97	6	0,000248	0,1830
GO:0043043	Proces biosinteze peptidov	BP	98	6	0,000263	0,1830
GO:0006518	Metabolni proces peptidov	BP	102	6	0,000328	0,1830
GO:1901564	Metabolni proces organskih dušikovih spojin	BP	324	10	0,000436	0,2031
GO:0043604	Proces biosinteze amidov	BP	111	6	0,000521	0,2079

Preglednica 12: 5 najbolj obogatenih GO skupin za vrhnjih 20 proteinov glede na napovedano RNA vezavo in hkratno sekrecijo, brez genov povezanih s translacijo. okrajšave: BP - biološki proces, MF – molekulska funkcija, CC – celični predel.

Skupina	Opis	Domena	Št. genov v skupini	Št. genov v skupini	p-vrednost	q-vrednost
GO:0006457	Zvijanje proteinov	BP	19	3	0,00125	1
GO:0009279	Zunanja celična membrana	CC	23	3	0,00222	1
GO:0030234	Regulacija encimske aktivnosti	MF	7	2	0,00275	1
GO:0098772	Regulacija molekulske funkcije	MF	8	2	0,00364	1
GO:0050790	Regulacija katalitske aktivnosti	BP	8	2	0,00364	1

4.5.8 Obogatitvena analiza: GO za skupino vezavnih parov RBPjev

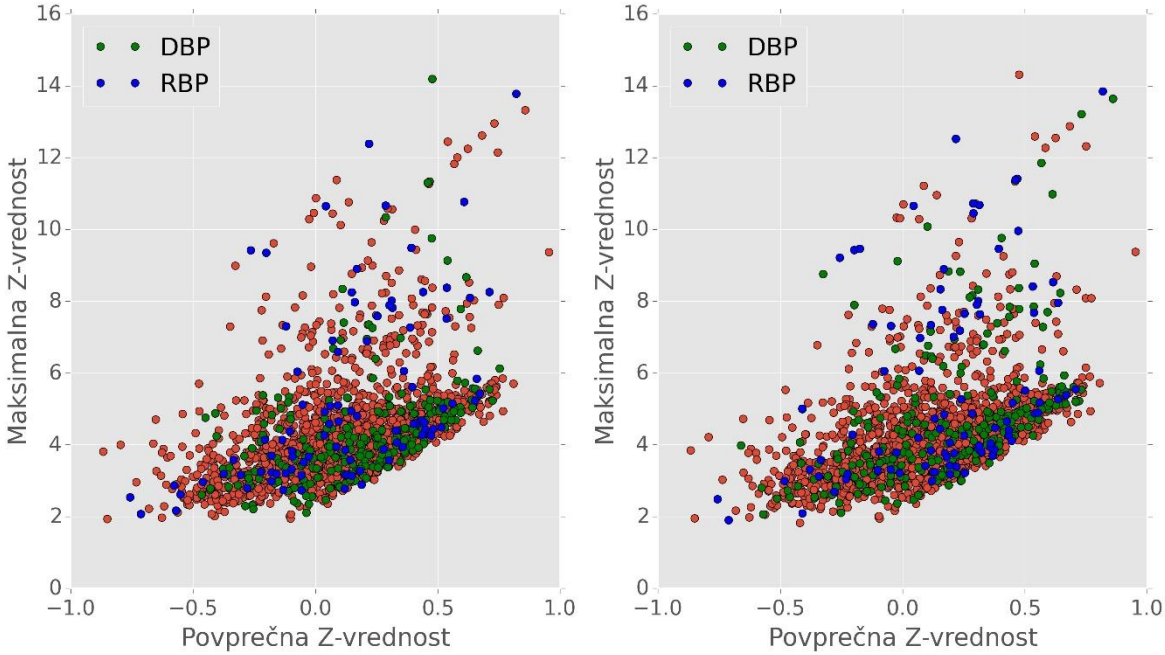
Obogatene GO skupine za najboljše mRNA pare (Priloga C) izbranih RBPjev (iz Preglednica 8) so predstavljene v Preglednici 13. Pri izboru 10 najboljših vezavnih tarč za vsakega od vrhnjih 20 *A. actinomycetemcomitans* proteinov je v uniji 105 unikatnih človeških genov.

Preglednica 13: 5 najbolj obogatenih GO skupin človeških genov, ki kažejo veliko verjetnost za vezavo s proteini iz vrha prioritetnega seznama. Okrajšave: BP - biološki proces, MF – molekulska funkcija, CC – celični predel.

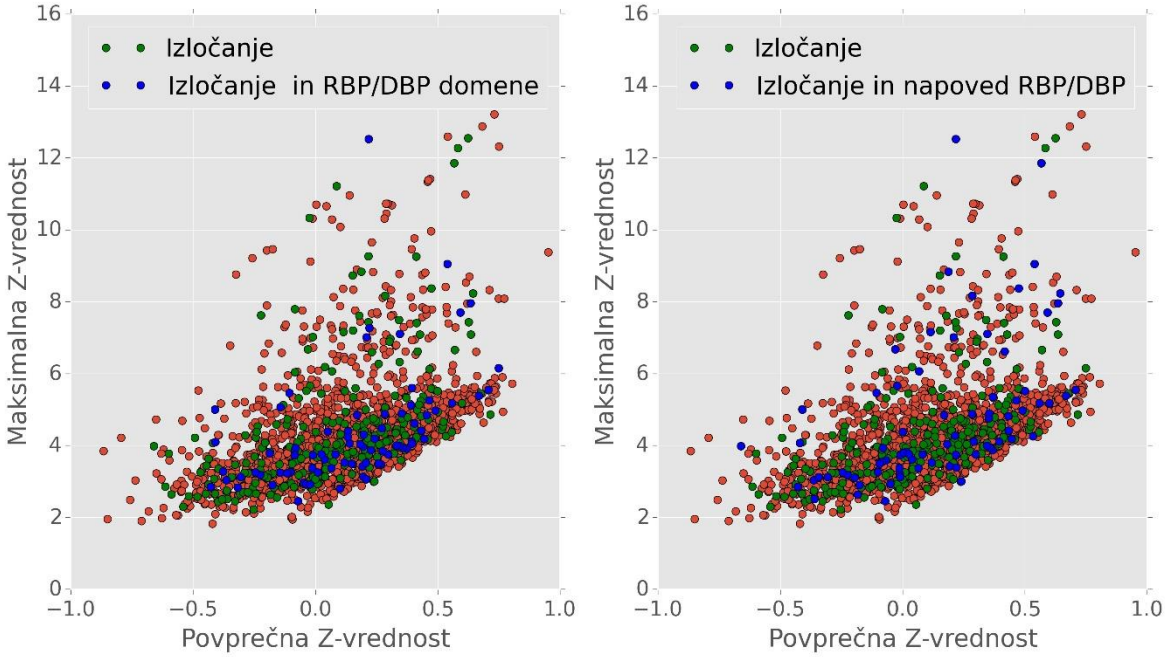
Skupina	Opis	Domena	Št. genov v skupini	Št. genov v preseku	p-vrednost	q-vrednost
GO:0042668	Določitev usode slušnih receptorskih celic	BP	2	2	0,0002	1
GO:0046658	Zasidrane spojine plazemske membrane	CC	11	3	0,0005	1
GO:0021527	Diferenciacija nevronov povezanih s hrbtenjačo	BP	5	2	0,0021	1
GO:0030901	Razvoj srednjih možganov	BP	19	3	0,0026	1
GO:0042491	Diferenciacija slušnih receptorskih celic	BP	6	2	0,0031	1

4.5.9 Specifična vezava RBP

V točki 4.5.1 smo določili kandidatne proteine, ki kažejo sposobnost sekrecije. Nato smo jih v točki 4.5.2 rangirali glede na splošnost sposobnost vezave molekul RNA. Te podatke smo uporabili za izgradnjo splošnega seznama. Pri obogatitvenih analizah smo analizirali, če kandidatni RBPji pripadajo določenim skupinam, oziroma če vplivajo na določeno skupino genov v človeškem organizmu. Zanimalo pa nas je tudi, če kakšni proteini izstopajo po specifičnosti in če napovedi kažejo, da se z določenimi molekulami mRNA posebej dobro vežejo. Iz Slike 27 je razvidno, da nekateri proteini res izstopajo in so vidni kot točke, ki na ordinatni osi presegajo vrednost 6. Te proteine predlagamo kot kandidate za nadaljnje analize specifične vezave na mRNA in s tem tarčnega vplivanja na evkarionsko celico. Dodatno smo še filtrirali kandidate za sekrecijo (Slika 28). Izbor predlaganih kandidatov je predstavljen v Preglednici 14.



Slika 27: Povprečna in maksimalna Z vrednost proteinov pred zlivanjem podatkov ob prisotnosti RNA/DNA vezavnih domen (levo) in po zlivanju podatkov (desno). Na desni so proteini obarvani kot RBP ali DBP glede na napovedene vrednosti prisotnosti RNA/DNA domen. Številčno razmerje je ohranjeno.



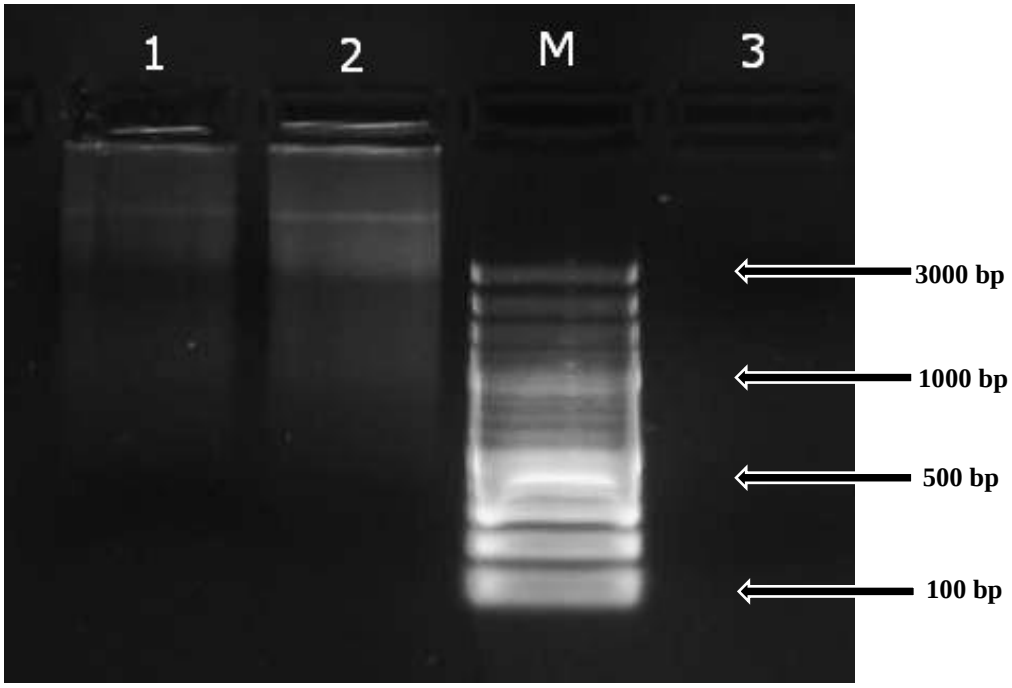
Slika 28: Povprečna in maksimalna Z vrednost proteinov po zlivanju podatkov ob določeni sekrecije in prisotnosti RNA/DNA vezavnih domen (levo) ter določena sekrecija in napovedana RNA/DNA vezavnost (desno).

Preglednica 14: Napovedani BRP, ki izstopajo s specifično vezavnostjo (maksimalna z vrednost > 7) in se (domnevno) izločajo. Pri izločanju oznaka S označuje določeno **s**ignalno zaporedje, M označuje sekretom in V označuje **v**ezikle.

Protein (oznaka gena)	**RNA/DNA vezavna domena**	**Vrednost napovedane RNA domene**	**Vezavnost (Maksimalna Z vrednost)**	**Izločanje**
polynucleotide phosphorylase/polyadenylase (D7S_01727)	RNA	0,4800	12,52476	MV
intracellular septation protein A (D7S_01525)	0	0,2663	11,86061	S
chaperone protein DnaK (D7S_02078)	DNA	0,3214	9,054249	MV
heat shock protein 90 (D7S_01489)	0	0,3523	8,836367	MV
thiol:disulfide interchange protein DsbD (D7S_00679)	0	0,2578	8,369340	S
F0F1 ATP synthase subunit alpha (D7S_01240)	0	0,2360	8,235792	MV
high-affinity Fe2+/Pb2+ permease (D7S_00378)	0	0,2522	8,168870	S
ribosomal protein S1 (D7S_00780)	RNA	0,4037	7,958123	V
Flp pilus assembly protein TadG (D7S_01444)	DNA	0,3295	7,698370	MV
threonyl-tRNA synthetase (D7S_02333)	0	0,2425	7,158458	V
cytochrome c nitrite reductase (D7S_01946)	DNA	0,2400	7,112221	SM
preprotein translocase subunit SecF (D7S_01729)	RNA	0,8342	7,017544	V

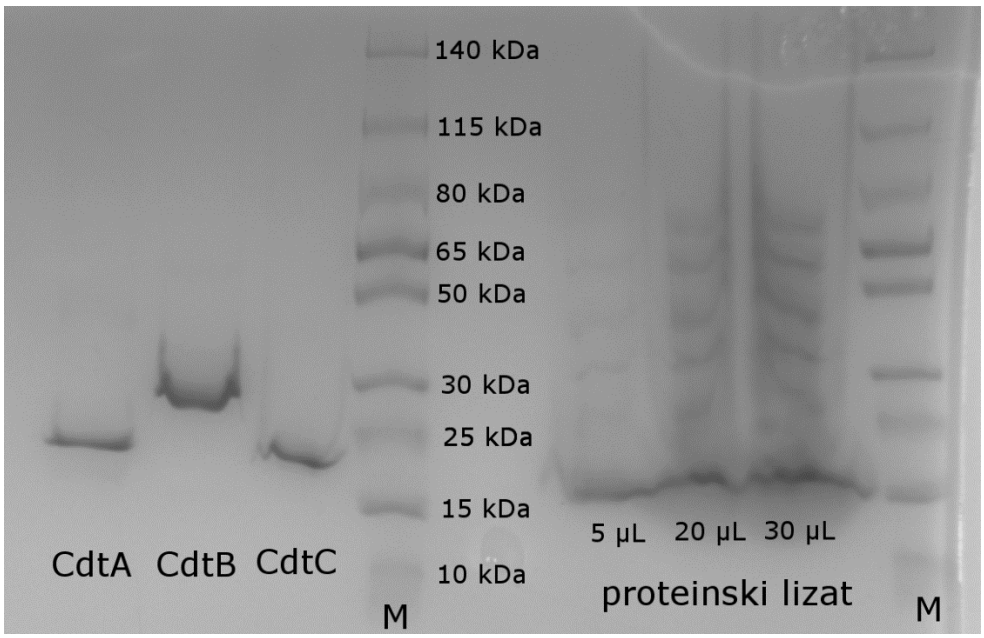
4.6 ANALIZA VEZAVE KOMPONENT CDT ALI CELIČNEGA LIZATA BAKTERIJE *A. ACTINOMYCETEMCOMITANS* NA MOLEKULE mRNA

Meritve absorbance izolirane mRNA so pokazale, da smo izolirali 7.1 ng/µL mRNA z razmerjem absorbance izmerjene pri 260 nm ter 280 nm ($A_{260/280}$) 2,45. Ta izolat smo uporabili za poskuse vezave s proteini. Pred testiranjem vezave proteinov na mRNA z metodo SPR smo analizirali kakovost in stabilnost izolirane mRNA na agarozni gelski elektroforezi. Preverili smo stabilnost mRNA v uporabljenem vezavnem pufru za 2 h pri sobni temperaturi in razgradnjo ob dodatku RNaze (Slika 29). Ugotovili smo, da smo izolirali zadostno količino mRNA in da je pripravljen pufer primeren za nadaljnje analize.



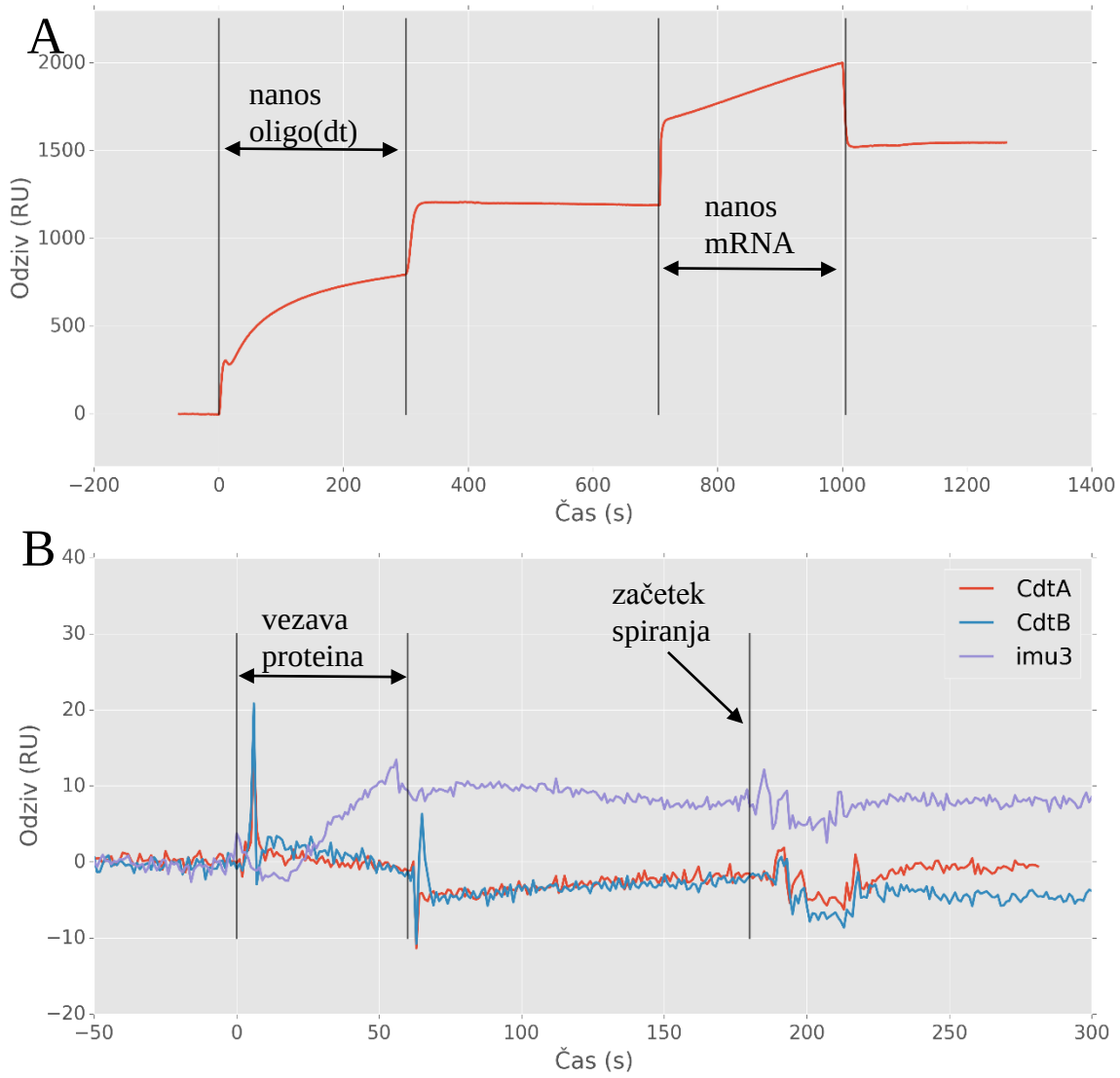
Slika 29: Agarozna gelska elektroforeza izolirane mRNA. Nanos glede na jamice: mRNA v vezavnem pufru (1), mRNA v vodi (2), dolžinski standard (M) in mRNA ob dodatku RNAze (3).

Prav tako smo z elektroforezo ugotovili, da imamo izolirane čiste CDT proteine (CdtA mase ~23,2 kDa, CdtB mase ~30 kDa, CdtC mase ~19,5 kDa) in da smo pridobili grobi celični lizat bakterije *A. actinomycetemcomitans* (Slika 30).



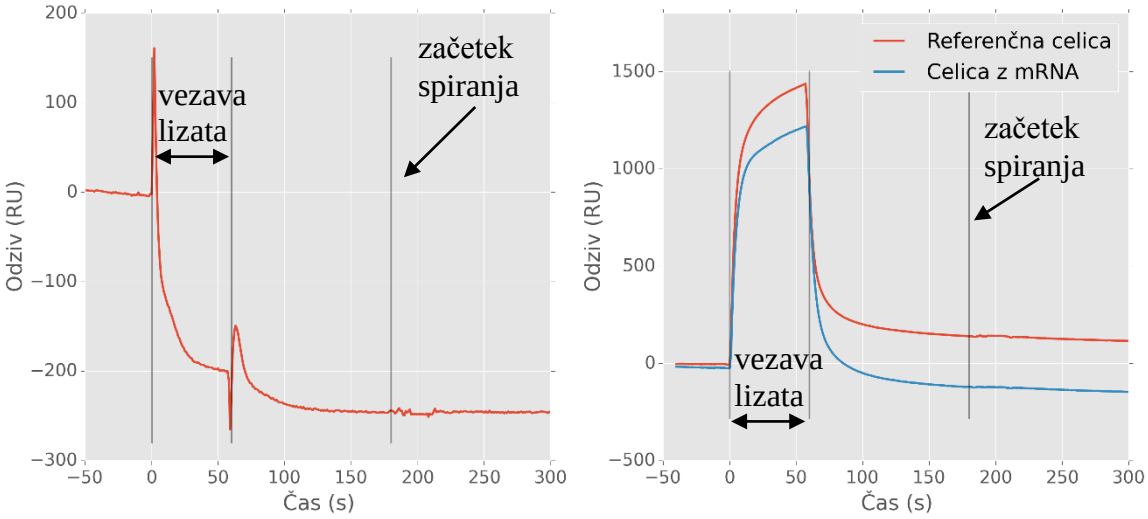
Slika 30: PAGE gel proteinov CDT in proteinskega lizata bakterije *A. actinomycetemcomitans*.

Z merjenjem površinske plazmonske resoncance (Slika 31) smo odkrili, da se CdtA ali CdtB ne vežeta na molekule mRNA v meri ki bi jo zaznali z refraktometrom, medtem ko se je kontrolni protein imu3 vezal stabilno. Zaradi težav ob nanosu vzorca CdtC preko mRNA, signal ni prikazan.



Slika 31: Senzogram imobilizacije mRNA molekul na čip SA in SPR študija interakcije CdtA, CdtB ali proteina imm3 z imobilizirano mRNA. A) Imobilizacija oligo(dT) ter mRNA na čip SA. B) Sensorgram interakcij komponent toksina CDT (1 µM) ali proteina imu3 (1 µM).

Z namenom prepoznave, ali sev bakterije *A. actinomycetemcomitans* izoliran iz pacienta sintetizira molekule, ki interagirajo s človeško mRNA, smo preko mRNA vezane na čip nanesli grobi celični lizat bakterije *A. actinomycetemcomitans*. Ob nanosu celotnega lizata (Slika 32) smo zaznali nespecifično vezave molekul lizata s površino čipa, kar je vidno kot negativna razlika pri odzivnih enotah med odzivom na celici 2 na katero so bile hibridizirane molekule mRNA ter referenčno, prazno celico 1 (Slika 32 levo). Hkrati pa je prišlo tudi do upada signala pod začetno vrednost na celici 2 po nanosu (Slika 32 desno), kar pripisujemo prisotnosti mRNA nukleazne aktivnosti v lizatu.



Slika 32: SPR senzogram vezave lizata na molekule mRNA. Levo: razlika signalov med celicamo 2, ki nosi molekule mRNA, in signalom zabeleženim na referenčni celici 1. Desno: signal na obeh celicah po vbrizgu lizata.

5 RAZPRAVA IN SKLEPI

V nalogi smo napovedovali interakcije med RNA in proteini v kontekstu okužbe in razvoja parodontoze. Napovedovanje interakcij je zahteven problem, saj je na voljo omejena količina neposrednih podatkov o teh interakcijah. Napovedni modeli so v zgodnjih fazah razvoja in večina omskih raziskav je praviloma omejena na modelne organizme. Zato smo se zanašali na posredne vire informacij: eksperimentalne podatke iz drugih raziskovalnih kontekstov, napovedne modele in opise (anotacije) elementov našega sistema.

5.1 ZLIVANJE PODATKOV IN ODKRIVANJE INTERAKCIJ

Odločili smo, da bomo uporabili metodo zlivanja podatkov s simultano matrično faktorizacijo. Metoda je že bila uspešno aplicirana za prioritizacijo genov (Žitnik in sod., 2015), za napovedovanje preživetja bolnikov obolelih z rakom (Žitnik in Zupan, 2015b), za napovedovanje asociacij med boleznimi (Žitnik in sod., 2013) in druge aplikacije. V tem delu smo metodo aplicirali še na odkrivanje potencialnih medvrstnih interakcij RNA–protein. Z uporabo metode smo se izognili ročnemu izbiranju kandidatov, ki bi lahko bilo časovno zamudno in pristransko. S takim pristopom zlivanja podatkov se namreč lahko izognemo filtriranju podatkov po korakih in lahko prioritiziramo tudi take kandidatne proteine, ki bi jih v nasprotnem primeru izpustili. Ta metoda pa omogoča avtomatizirano iskanje struktur v podatkih, ki lahko imajo prisoten šum, manjkajoče ali napačne vrednosti. Z zlivanjem lahko napovedujemo oziroma popravljamo obstoječe vrednosti na podlagi struktur. Tolerantnost do napak oziroma manjkajočih vrednosti se je izkazala za koristno, saj so v raziskavah *A. actinomycetemcomitans* uporabljeni različni sevi, ki se med seboj lahko močno razlikujejo (Kittichotirat in sod., 2011). Tako lahko, na primer, prioritiziramo proteine, ki nimajo klasičnih RNA vezavnih domen, vendar struktura v podatkih kaže, da imajo podoben profil relacij tistim proteinom, ki imajo klasične domene. Dodatno se lahko z uporabo te metode zanašamo na vire informacij, ki niso neposredno povezani s preučevanim objektom. V našem primeru smo lahko povezali proteine bakterije *A. actinomycetemcomitans* z ekspresijo človeških genov. Izkazalo se je, da je matrična faktorizacija za namen napovedovanja vrednosti primerljiva z drugimi metoda strojnega učenja (Poglavje 4.4).

Tako smo v tej nalogi na nov način preučevali možne vzroke razvoja parodontoze skozi prizmo interakcij RNA-protein. V napovedni model smo vključili relacije, ki opisujejo sposobnost vezave RNA in relacije, za katere smo domnevali, da so relevantne za nastanek bolezenskega stanja. Skupno smo integrirali 11 relacij. Verižili smo podatke od izražanja genov v *A. actinomycetemcomitans*, sposobnosti sekrecije nastalih proteinov, do interakcij s človeškimi molekulami mRNA in diferencialnega izražanja le-teh. Ugotovili smo, da običajno pri nižjih rRF dobivamo boljše rezultate (Poglavje 4.2). Zato smo za nadaljnje analize uporabili 5 % rRF. Pri preverjanju informativnost virov (Poglavje 4.3), smo ugotovili, da izvzem posameznega vira, razen centralne matrike interakcij, ne vpliva izrazito na napovedovanje drugih vrednosti. Predpostavljamo, da to ni samo posledica informacije, ki jo matrika nosi, ampak tudi njene strukture. Matrika je namreč velika in ni redka, zato za optimizacijo v latentnem prostoru porabi večji delež prostora. Pri napovedovanje RNA in DNA vezavnih domen se je kot pomemben vir izkazala še funkcija

ortolognih skupin. Domneve, da bi lahko viri, kot so diferencialno ekspresija človeških oziroma genov bakterije *A. actinomycetemcomitans* ali sestava sekretoma, bistveno pripomogla k napovedovanju prisotnosti RNA vezavnih domen, nismo potrdili. Pri napovedovanju RNA vezavnih domen odstopa še vir diferencialna ekspresija *A. actinomycetemcomitans* med *in vivo* in *in vitro* rastjo. Predpostavljamo, da je lahko vzrok diferencialno izražanje genov za RBP, ki so odgovorni za prilagajanje metabolizma na spremenjeno okolje. Raziskave namreč kažejo, da je lahko virulenca bakterij uravnavana na postranskripcijskem nivoju (Oliva in sod., 2015).

5.2 KANDIDATNI SEZNAM

Ker nismo imeli na voljo referenčnih profilov vezave, na katere bi se lahko upirali za razporeditev, smo predlagali ocenjevalno funkcijo, podobno tisti, ki jo uporabljajo avtorji metode catRAPID omics. Pri njej upoštevamo kriterij Z-vrednosti in prisotnost RNA oziroma DNA vezavnih domen. Naš primer se je razlikoval v tem, da smo rekonstruirane Z-vrednosti povprečili, s čimer smo izpostavili splošne oziroma nespecifične kandidatne RBPje. Pri končni formuli smo dodatno upoštevali še verjetnostno porazdelitev zlitih podatkov. Kot potencialne proteine z delovanjem v evkariontski celici pa smo predlagali tiste, ki so bili odkriti v sekretomu oziroma veziklih ali pa imajo napovedano signalno zaporedje. Želeli smo izpustiti čim manj potencialno zanimivih proteinov, tudi če imamo lažno pozitivne kandidate v končnem izboru. Primeri lažno pozitivnih proteinov so lahko na primer proteini s signalnim zaporedjem, ki lahko ostanejo del celične membrane.

Ugotavljamo, da uspešno prioritiziramo kandidate, saj so prioritizirani kandidate smiselni. Na primer, visoko so uvrščeni ribosomalni proteini brez klasičnih RNA vezavnih domen (S20, L23, L7/L12, S14). S pregledom literature je mogoče povezati prioritizirane proteine z vezavo RNA tudi za druge proteine. Tak je na primer protein HU, ki je prvi na seznamu in ima določeno samo DNA vezavno domeno. Za ta protein je znano, da sodeluje pri oblikovanju nukleoida ter da ima pomembno vlogo pri podvajanju DNA, rekombinaciji in popravljalnih mehanizmih. Sodeluje tudi pri transkripciji in podrobneje je preučena njegova vloga pri preživetju v stacionarni fazi in stresnih pogojih. Protein je zelo podoben histonom in drugim evkariontskim proteinom. Raziskave pa so pokazale, da ta protein ni sposoben vezave samo na dvojno vijačnico DNA, ampak tudi na RNA in DNA-RNA hibridne molekule (Balandina in sod., 2002). Nadalje lahko v literaturi najdemo tudi opise o izločanju proteina HU v povezavi z vnetji. Pri bakteriji *Helicobacter pylori*, ki povzroča vnetja želodca in poškodbe epitelnega tkiva, je bilo odkrito, da je HU eden izmed trinajstih proteinov, ki se izločajo iz celice v visoki koncentraciji s specifičnim mehanizmom in ne z nespecifično celično lizo (Kim in sod., 2002). Podobno lahko najdemo v literaturi namige, da bi tudi peti protein na seznamu in prvi brez RNA ali DNA vezavnih domen, protein »WD40 repeat containing«, lahko bil RNA vezaven. Namreč, za proteine, ki imajo WD40 domene, to so domene iz tandemskih ponovitev WD40, je bilo dokazano, da so RNA vezavni. Vezava je bila dokazana tako za snRNA (Lau in sod., 2009), kot tudi za molekule RNA s poli-A repom (Kwon in sod., 2013). Relevantnost seznama smo preverjali še avtomatizirano s primerjavo BLAST na identificirane RBPje iz sesalskih celic (Poglavje 4.5.5). Pričakovali smo, da bo na vrhu seznama več RNA vezavnih proteinov. Obogatitev

opažamo do položaja 250, vendar ta ni izrazita. Razlog ni nujno v slabi prioritizaciji, lahko je tudi v manjši podobnosti med RBPji sesalcev in bakterijo *A. actinomycetemcomitans*.

Z obogatitveno analizo smo pokazali, da prioritiziramo skupine, povezane s translacijo (Poglavje 4.5.7). To je v skladu s pričakovanji, saj translacijski mehanizem vstopa v stik z molekulami RNA. Van Assche in sod. (2015) so že povzeli mehanizme delovanja RNA vezavnih proteinov pri bakterijah in opisali delovanje najbolj preučenih RBPjev. Najbolj preprost mehanizem je inhibicija translacije ob vezavi molekule mRNA. Vendar v pregledu literature nismo zasledili, da bi kdo preučeval RBPje kot toksine oziroma efektorje v evkariontskih celicah. Domnevamo, da bi lahko bakterijski proteini v evkariontski celici delovali kot RBPji in blokirali translacijo. Če izvzamemo proteine, ki so povezani s translacijo, najdemo na vrhu obogatitvenega seznama, čeprav le s tremi elementi v preseku, šaperone, ki sodelujejo pri zvitju proteinov. Tudi ta skupina je zanimiva za nadaljnjo analizo, saj nekateri bakterijski proteini kažejo možnosti interakcij tako z RNA kot tudi s proteini (Kovacs in sod., 2009).

Pri našem napovedovanju smo se v veliki meri zanašali na napovedi programa catRAPID. Ta napovedni model je bil treniran na 592 znanih interakcijah, mi pa smo za napovedovanje interakcij uporabili 2001 proteinov, ki so si po podobnosti bolj oddaljeni. Zaradi take ekstrapolacije podatkov moramo biti previdni. Problem napačnih napovedi pri zaporedjih, kjer ni homologov in podatkov o interakcijah, navajajo tudi avtorji metode catRAPID (Bellucci in sod., 2011).

Domnevali smo, da bomo lahko na strani humanih mRNA tarč z obogatitveno analizo dobili obogatene GO skupine bioloških procesov kot so vnetje (Herbert in sod., 2015), celična apoptoza, uravnavanje razvoja osteoblasti-osteoklasti (Handfield in sod. 2005) in podobnih. V naši analizi teh skupin nismo identificirali kot obogatenih (Poglavje 4.5.8). Najvišje se uvrščajo skupine, povezane z nevralnim razvojem (Preglednica 13), ki imajo malo elementov. V seznamu tarčnih človeških genov sicer najdemo skupine povezane z apoptozo (ENSG00000116824, ENSG00000100290), vnetjem (ENSG00000172156) in diferenciacijo osteoklastov (ENSG00000164326), vendar moramo biti zaradi že prej omenjene ekstrapolacije in povezanih težav napačnega napovedovanja ter lažno pozitivnih rezultatov zaradi velikosti izbranega seznama pri sklepanju zaključkov previdni.

5.3 DOKAZOVANJE INTERAKCIJ RNA-PROTEIN

Uporabljen sistem, pri katerem smo izkoriščali mehanizem površinske plazmonske resonance za odkrivanje interakcij, se je izkazal za učinkovitega (Poglavje 4.6). Uspešno smo izvedli kaskadno vezavo več bioloških molekul, in sicer smo na streptavidinski čip integrirali molekule biotiliniranega poli-deoksitimidin-monofosfata. Na to smo na te molekule hibridizirali molekule mRNA preko poliadeliniranega repa in šele po tem izvedli vezavo analiziranega proteina. Pokazali smo, da je mogoče analizirati celotni transkriptomski profil in ne samo ene molekule, kot je primer pri klasičnih SPR analizah vezav mRNA (Katsamba in sod., 2002). Med potekom poskusov je sistem ostal stabilen, kajti nismo zaznali postopnega upadanja signala zaradi morebitnih kontaminant RNaz.

Tako smo lahko zaporedno testirali nanos proteinov brez ponovnega nanašanja molekul mRNA v primeru, da se analiziran protein ni vezal na liganda.

Laboratorijska analiza vezave komponent A in B virulentnega dejavnika CDT bakterije *A. actinomycetemcomitans* ni pokazala, da bi ta proteina interagirala z molekulami mRNA. Pozitivni signal pa smo dobili za kontrolni protein imu3, kateremu je že bila dokazana vezava bakterijske RNA (Črnigoj in sod., 2014). Od teh dveh proteinov iz kompleksa CDT, ki ga preučujejo na Katedri za biokemijo Oddelka za biologijo Biotehniške fakultete, je bila zanimiva analiza predvsem proteina CdtB, ki ima predvideno DNA vezavno domeno in opisano sposobnost restrikcije molekul DNA. Poleg tega novejše študije kažejo, da ima veliko DNA vezavnih proteinov sposobnost vezave tudi molekul RNA (Hudson in Ortlund, 2014). V našem prioritetnem seznamu je bil ta protein uvrščen na 897. mesto, je pa nazadoval z 249. mesta pred zlivanjem podatkov.

Pri uporabljeni metodi se moramo zavedati, da smo uporabili samo nabor poliadeniliranih molekul RNA iz dveh celičnih linij. Posledično moramo dopuščati možnost, da obstajajo interakcije z molekulami mRNA, ki se ne izražajo v analiziranih celicah oziroma se izražajo v tako majhnih količinah, da jih ne moremo zaznati. Kljub temu smo z izborom epitelne in kostne celične linije zajeli dva pomembna mehanizma, na katera vpliva prisotnost bakterije. To sta povzročanje nekroze epitelnih celic in inhibicija delovanja osteoklastov.

Uporabljena metoda validacije vezave proteinov celičnega lizata bakterije *A. actinomycetemcomitans* z mRNA je potrebna nadaljnje optimizacije. Potrebno je izničiti razgradnjo mRNA in minimizirati nespecifične interakcije med molekulami in površino čipa tekom poskusa. Za problem razgradnje mRNA predlagamo dodatek RNaznih inhibitorjev lizatu pred refraktometersko analizo. Poskus z dodatkom govejega serumskega albumina ni rešil problema nespecifične vezave RNA. Zato za naslednje poskuse predlagamo obogatitev RBPjev pred samo analizo na refraktometru. Na primer, z uporabo molekul mRNA lahko afinitetno selekcioniramo proteine, za katere v naslednjem koraku testiramo še vezavo na SPR. Za te poskuse bo potrebna nadaljnja optimizacija pogojev. Dodati je še potrebno, da bakterije virulentne dejavnike pogosto sintetizirajo le ob stiku s celico gostitelja (Jorth in sod., 2013), zato bi bilo pomembno pripraviti lizat iz bakterij, gojenih na tkivu ali raslih v prisotnosti izrabljenega gojišča tkivnih kultur.

Zaključujemo, da smo v tej nalogi uspešno predlagali kandidatni seznam proteinov. S tem smo pokazali, da obstajajo kandidatni proteini s potencialno vezavo molekul mRNA. Ni pa nam uspelo predlaganega seznama tudi eksperimentalno preveriti. Kljub temu delo postavlja izhodišče in validacijski test za nadaljnje študije molekulskih interakcij v ustni votlini, kar je nujno za razumevanje izvora in poteka parodontoze.

5.4 SKLEPI

Na podlagi rezultatov, ki smo jih dobili v tej raziskavi, lahko zaključimo sledeče:

- Z integracijo podatkov smo sestavili prioritetni seznam RNA vezavnih proteinov.
- Proteini z vrha seznama so bolj podobni sesalskim RBPjem kot nižje rangirani proteini.
- Uporabljena metoda matrične faktorizacije je primerljiva z metodama naključnih gozdov in logistično regresijo pri napovedovanju RNA/DNA vezavnih domen in proteinov v sekretomu.
- Prečno preverjanje je pokazalo, da lahko na podlagi podatkov bakteriji *A. actinomycetemcomitans* napoveduje pristnost RNA vezavnih domen (mediana AUC = 0,750). Nekoliko slabše lahko napovedujemo proteine, ki se izločajo (mediana AUC = 0,627).
- Med proteini, ki se izločajo iz bakterije *A. actinomycetemcomitans* in so na vrhu prioritetnega seznama RNA vezavnih proteinov, je obogatena genska skupina povezana s translacijo.
- Pri obogatitveni analizi humanih mRNA tarč ni izstopala nobena skupina proteinov.
- Proteinom bakterijskega holotoksina CDT nismo dokazali sposobnosti vezave mRNA.
- Metoda na osnovi površinske plazmonske resonance je uporabna za validacijo napovedanih RBPjev.

6 POVZETEK

Okoli 50% ljudi, starejših od 50 let, ima določeno obliko bolezni dlesni. Parodontalna bolezen je kronična vnetna bolezen dlesni, ki se začne z oteklino in rdečino dlesni, v napredovani fazi pa vodi do propada pozobničnega ligamenta, kar lahko vodi do majavosti in izgube zob. Bakterija *Aggregatibacter actinomycetemcomitans* je močno povezana z nastankom parodontoze. Do sedaj sta bila preučena dva toksina bakterije *A. actinomycetemcomitans*, ki toksično vplivata na evkariontske celice. To sta toksin CDT, ki povzroča dvoverižne prelome DNA, in levkotoksin, ki sproži apoptotične poti v celicah, oziroma v velikih koncentracijah povzroča nekrozo tarčnih celic, saj tvori pore v membrani. Virulenca bakterije je domnevno odvisna še od drugih mehanizmov in virulentnih dejavnikov. Za bakterijo je znano, da s pomočjo veziklov oziroma drugih sistemov izločanja translocira številne proteine, tudi virulentne dejavnike, iz bakterije. V tej nalogi nas je zanimalo, ali lahko z zlivanjem podatkov z metodo simultane matrične faktorizacije odkrijemo kandidatne proteine, ki bi lahko delovali kot RNA vezavni in vplivali na delovanje ekvariontske celice. Med glavnimi cilji naloge je bila izgradnja prioritetnega seznama RNA vezavnih proteinov in izbira tistih proteinov, ki se lahko potencialno prenesejo v celice gostitelja.

Ker je področje medvrstnih interakcije protein-RNA slabo raziskano, smo se zanašali na posredne podatke in iskanje strukture v le-teh. Izgradili smo relacijski graf z enajstimi povezavami, kjer so vključeni eksperimentalni podatki, anotacije in rezultati napovednih programov. Eksperimentalni podatki so zajemali diferencialno izražanje genov človeških celic ob okužbi z bakterijo *A. actinomycetemcomitans,* sekretomske podatke bakterije in diferencialno izražanje genov bakterije tekom rasti *in vivo*. Kot referenčni genom smo uporabili genom bakterijskega seva D7S-1.

S prečnim preverjanjem smo pokazali, da je mogoče uporabiti metodo zlivanja podatkov za napovedovanje RNA (in DNA) vezavnih domen ter nekoliko manj zanesljivo za napovedovanje proteinov, ki se izločajo. Napovedovanje teh vrednosti z matrično faktorizacijo, ob uporabi enakega nabora podatkov, je primerljivo z metodama naključnih gozdov in logistične regresije. Predlagali smo prioritetni seznam RNA vezavnih proteinov bakterije *A. actinomycetemcomitans*, ki bi se lahko izločali iz prokarionta in v celicah evkariontskega organizma vstopale v interakcije z molekulami mRNA. Proteini z vrha seznama (približno do mesta 250) kažejo večjo podobnost z identificiranimi RBPji sesalcev kot pa nižje uvrščeni proteini. Obogatitvena analiza vrhnjega seznama je pokazala, da smo prioritizirali gene, povezane s translacijo. Na strani človeških mRNA tarč nismo zaznali značilne obogatitve GO skupin.

Za namen eksperimentalne validacije našega seznama RBPjev smo razvili protokol vezave proteinov na molekule mRNA s pomočjo refraktometrije. Podenotama toksina CDT nismo določili sposobnosti vezave RNA. Nespecifična vezava proteinskega lizata bi lahko bila vzrok, da z napravo za refraktometrijo nismo uspeli dokazati vezave proteinov iz grobega celičnega lizata bakterije *A. actinomycetemcomitans* na človeške molekule mRNA in ločitev vezavnih proteinov za nadaljnjo analizo. Sistem pa ponuja izhodišče za *in vitro* teste vezave heterologno izraženih proteinov *A. actinomycetemcomitans* z molekulami mRNA.

7 VIRI

Agostini F., Zanzoni A., Klus P., Marchese D., Cirillo D., Tartaglia G. G. 2013. catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. Bioinformatics, 29, 22: 2928-2930

Balandina A., Kamashev D., Rouviere-Yaniv J. 2002. The bacterial histone-like protein HU specifically recognizes similar structures in all nucleic acids. DNA, RNA, and their hybrids. The Journal of Biological Chemistry, 277, 31: 27622-27628

Baltz A. G., Munschauer M., Schwanhäusser B., Vasile A., Murakawa Y., Schueler M., Youngs N., Penfold-Brown D., Drew K., Milek M., Wyler E., Bonneau R., Selbach M., Dieterich C., Landthaler M. 2012. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Molecular Cell, 46, 5: 674-690

Baxt L. A., Garza-Mayers A. C., Goldberg M. B. 2013. Bacterial subversion of host innate immune pathways. Science, 340, 6133: 697-701

Bellucci M., Agostini F., Masin M., Tartaglia G. G. 2011. Predicting protein associations with long noncoding RNAs. Nature, 8, 6: 444-445

Castello A., Fischer B., Eichelbaum K., Horos R., Beckmann B. M., Strein C., Davey N. E., Humphreys D. T., Preiss T., Steinmetz L. M., Krijgsveld J., Hentze M. W. 2012. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell, 149, 6: 1393-1406

CDC.gov. Periodontal Disease http://www.cdc.gov/oralhealth/periodontal_disease/ (november, 2015)

Chahboun H., Arnau M. M., Herrera D., Sanz M., Ennibi O. K. 2015. Bacterial profile of aggressive periodontitis in Morocco: a cross-sectional study. BMC Oral Health, 15, 25, doi: 10.1186/s12903-015-0006-x: 8 str.

Chen C., Kittichotirat W., Chen W., Downey J. S., Si Y., Bumgarner R. 2010. Genome sequence of naturally competent *Aggregatibacter actinomycetemcomitans* serotype a strain D7S-1. Journal of Bacteriology, 192, 10: 2643-2644

Cirillo, D., Agostini, F. and Tartaglia, G. G. 2012. Predictions of protein–RNA interactions. WIREs Computational Molecular Science, 3, 2: 161-175

Cock P. J., Antao T., Chang J. T., Chapman B. A., Cox C. J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., de Hoon M. J. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 11: 1422-1423

Črnigoj M., Podlesek Z., Budič M., Zgur-Bertok D. 2014. The Escherichia coli uropathogenic-specific-protein-associated immunity protein 3 (Imu3) has nucleic acid binding activity. BMC Microbiology, 14, 16, doi: 10.1186/1471-2180-14-16: 8 str.

Demšar J. 2006. Statistical Comparisons of Classifiers over multiple Data Sets. Journal of Machine Learning Research, 7: 1-30

Deslandes L., Rivas S. 2012. Catch me if you can: bacterial effectors and plant targets. Trends in Plant Science, 17, 11: 644-655

DiRienzo J. M. 2014a. Breaking the Gingival Epithelial Barrier: Role of the *Aggregatibacter actinomycetemcomitans* Cytolethal Distending Toxin in Oral Infectious Disease. Cells, 3, 2: 476-499

DiRienzo J. M. 2014b. Uptake and processing of the cytolethal distending toxin by mammalian cells. Toxins, 6, 11: 3098-3116

Demuth D. R., James D., Kowashi Y., Kato S. 2003. Interaction of *Actinobacillus actinomycetemcomitans* outer membrane vesicles with HL60 cells does not require leukotoxin. Cellular Microbiology, 5,2: 111-121

Eddy S.R. 2009. A new generation of homology search tools based on probabilistic inference. Genome informatics. International Conference on Genome Informatics, 1: 205-211

Feng Z., Weinberg A. 2006. Role of bacteria in health and disease of periodontal tissues. Periodontol 2000, 40: 50-76

Fontana M. F., Banga S., Barry K. C., Shen X., Tan Y., Luo Z. Q., Vance R. E. 2011. Secreted bacterial effectors that inhibit host protein synthesis are critical for induction of the innate immune response to virulent Legionella pneumophila. PLoS Pathogens, 7, 2: e1001289, doi: 10.1371/journal.ppat.1001289: 15 str.

Glisovic T., Bachorik J. L., Yong J., Dreyfuss G. 2008. RNA-binding proteins and post-transcriptional gene regulation. FEBS Letters, 582, 14: 1997-1986

Goll M.G., Kirpekar F., Maggert K.A., Yoder J.A., Hsieh C.L., Zhang X., Golic K.G., Jacobsen S.E., Bestor T.H. 2006 Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science, 311, 5759: 395-398

Handfield M., Mans J. J., Zheng G., Lopez M. C., Mao S., Progulske-Fox A., Narasimhan G., Baker H. V., Lamont R. J. 2005. Distinct transcriptional profiles characterize oral epithelium-microbiota interactions. Cellular Microbiology, 7, 6: 811-823

Herbert B. A., Novince C. M., Kirkwood K.L. 2015. *Aggregatibacter actinomycetemcomitans*, a potent immunoregulator of the periodontal host defense system and alveolar bone homeostasis. Molecular Oral Microbilogy 22 (pred tiskom), doi: 10.1111/omi.12119: 21 str.

Henderson B., Ward J. M., Ready D. 2010. *Aggregatibacter (Actinobacillus) actinomycetemcomitans*: a triple A* periodontopathogen? Periodontology 2000, 54, 1: 78-105

Hudson W. H., Ortlund E. A. 2014. The structure, function and evolution of proteins that bind DNA and RNA. Nature Reviews. Molecular Cell Biology, 15, 11: 749-760

Jorth P., Trivedi U., Rumbaugh K., Whiteley M. 2013. Probing Bacterial Metabolism during Infection Using High-Resolution Transcriptomics. Journal of Bacteriology, 195, 22: 4991–4998

Kachlany S. C. 2010. *Aggregatibacter actinomycetemcomitans* Leukotoxin. Journal of Dental Research, 89, 6: 561-570

Kaplan J. B., Perry M. B., MacLean L. L., Furgang D., Wilson M. E., Fine D. H. 2001. Structural and genetic analyses of O polysaccharide from *Actinobacillus actinomycetemcomitans* serotype f. Infection and Immunity, 69, 9: 5375-5384

Katsamba P. S., Park S., Laird-Offringa I. A. 2002. Kinetic studies of RNA-protein interactions using surface plasmon resonance. Methods, 26, 2: 95-104

Kieselbach T., Zijnge V., Granström E., Oscarsson J. 2015. Proteomics of *Aggregatibacter actinomycetemcomitans* Outer Membrane Vesicles. PloS One, 10, 9: e0138591, doi: 10.1371/journal.pone.0138591: 21 str.

Kim N, Weeks D. L., Shin J. M., Scott D. R., Young M. K., Sachs G. 2002. Proteins Released by *Helicobacter pylori* in vitro. Journal of Bacteriology, 184, 22: 6155-6162

Kittichotirat W., Bumgarner R. E., Asikainen S., Chen C. 2011. Identification of the pangenome and its components in 14 distinct *Aggregatibacter actinomycetemcomitans* strains by comparative genomic analysis. PLoS One, 6,7: e22420, doi: 10.1371/journal.pone.0022420: 12 str.

Kovacs D., Rakacs M., Agoston B., Lenkey K., Semrad K., Schroeder R., Tompa P. 2009. Janus chaperones: assistance of both RNA- and protein-folding by ribosomal proteins. FEBS Letters, 583, 1: 88-92

Kwon S. C., Yi H., Eichelbaum K., Föhr S., Fischer B., You K. T., Castello A., Krijgsveld J., Hentze M. W., Kim V. N. 2013. The RNA-binding protein repertoire of embryonic stem cells. Nature Structural & Molecular Biology, 20, 9: 1122-1130

Lau C. K., Bachorik J. L., Dreyfuss G. 2009. Gemin5-snRNA interaction reveals an RNA binding function for WD repeat domains. Nature Structural & Molecular Biology, 16, 5: 486-491

Lemaitre B., Girardint S. E. 2013. Translation inhibition and metabolic stress pathways in the host response to bacterial pathogens. Nature Reviews Microbiology, 11, 6: 365-369

Livi C. M., Klus P., Delli Ponti R., Tartaglia G. G. 2015. catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics (pred tiskom), doi: 10.1093/bioinformatics/btv629: 3 str.

Meyer D. F., Noroy C., Moumène A., Raffaele S., Albina E., Vachiéry N. 2013. Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. Nucleic Acids Research, 41, 20: 9218-9229

Molekulske-interakcije.si. Navodila_za_BiacoreX_2015 http://www.molekulske-interakcije.si/sl/uporabno.html (december, 2015)

Munksgaard P. S., Vorup-Jensen T., Reinholdt J., Söderström C. M., Poulsen K., Leipziger J., Praetorius H.A., Skals M. 2012. Leukotoxin from *Aggregatibacter actinomycetemcomitans* causes shrinkage and P2X receptor-dependent lysis of human erythrocytes Cellular Microbiology, 14, 12: 1904-1920

Obradović D., Gašperšič R., Seme K., Maček P., Butala M. 2014. Genotipska karakterizacija sevov oportunističnega patogena aggregatibacter actinomycetemcomitans pri slovenskih bolnikih s kroničnim parodontitisom, preliminarna raziskava. Zobozdravstveni vestnik, 69: 21-27

Oliva G., Sahr T., Buchrieser C. 2015. Small RNAs, 5' UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence. FEMS Microbiology Reviews, 39, 3, 331-349

Petersen T. N. , Brunak S., von Heijne G., Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 8, 10: 785-786

Portela A., Digard P. 2002. The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. The Journal of General Virology, 83, 4: 723-734

Powell S., Forslund K., Szklarczyk D., Trachana K., Roth A., Huerta-Cepas J., Gabaldón T., Rattei T., Creevey C., Kuhn M., Jensen L. J., von Mering C., Bork P. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Research, 42: 231-239

Raja M., Ummer F., Dhivakar C. P. 2014. *Aggregatibacter Actinomycetemcomitans* – A Tooth Killer? Journal of Clinical and Diagnostic Research, 8, 8: 13-16

Rylev M., Kilian M. 2008. Prevalence and distribution of principal periodontal pathogens worldwide. Journal of Clinical Periodontology, 35, 8: 346-361

Rompikuntal P. K., Thay B., Khan M. K., Alanko J., Penttinen A. M., Asikainen S., Wai S. N., Oscarsson J. 2012. Perinuclear localization of internalized outer membrane vesicles carrying active cytolethal distending toxin from *Aggregatibacter actinomycetemcomitans*. Infection and Immunity, 80, 1: 31-42

Schreiner H., Li Y., Cline J., Tsiagbe V. K., Fine D. H. 2013. A comparison of *Aggregatibacter actinomycetemcomitans* (Aa) virulence traits in a rat model for periodontal disease. PLoS One, 8, 7: e69382, doi: 10.1371/journal.pone.0069382: 8 str.

Teng Y.T., Zhang X. 2005. Apoptotic activity and sub-cellular localization of a T4SS-associated CagE-homologue in *Actinobacillus actinomycetemcomitans*. Microbial Pathogenisis, 38, 2: 125-132

Tobin C., Mandava C. S., Ehrenberg M., Andersson D. I., Sanyal S. 2010. Ribosomes lacking protein S20 are defective in mRNA binding and subunit association. Journal of Molecular Biology, 397, 3: 767-777

Van Assche E., Van Puyvelde S., Vanderleyden J., Steenackers H. P. 2015. RNA-binding proteins involved in post-transcriptional regulation in bacteria. Frontiers in Microbiology, 6, 141, doi: 10.3389/fmicb.2015.00141: 16 str.

Zimmerly S., Guo H., Eskes R., Yang J., Perlman P. S., Lambowitz A. M. 1995. A group II intron RNA is a catalytic component of a DNA endonuclease involved in intron mobility. Cell, 83, 4: 529-538

Zijnge V., Kieselbach T., Oscarsson J. 2012. Proteomics of protein secretion by *Aggregatibacter actinomycetemcomitans*. PLoS One, 7, 7: e41662, doi: 10.1371/journal.pone.0041662: 11 str.

Žitnik M. 2015a. Learning by Fusing Heterogeneous Data. PhD thesis. University of Ljubljana, Faculty of Computer and Information Science: 337 str.

Žitnik M. 2015b. Scikit-fusion version 0.2.1 http://github.com/marinkaz/scikit-fusion (december, 2015)

Žitnik M., Janjić V., Larminie C., Zupan B., Pržulj N. 2013. Discovering disease-disease associations by fusing systems-level molecular data. Scientific Reports, 3, 3202, doi: 10.1038/srep03202: 9 str.

Žitnik M., Nam E. A., Dinh C., Kuspa A., Shaulsky G., Zupan B. 2015. Gene Prioritization by Compressive Data Fusion and Chaining. PLoS Computational Biology, 11, 10: e1004552, doi: 10.1371/journal.pcbi.1004552: 18 str.

Žitnik M., Zupan B. 2015a. Data Fusion by Matrix Factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37, 1: 41-53

Žitnik M., Zupan B. 2015b. Survival regression by data fusion. Systems Biomedicine, 2, 3: 47-53

ZAHVALA

Na prvem mestu se zahvaljujem mentorju doc. dr. Mateju Butali in somentorju doc. dr. Tomažu Curku. Obema se zahvaljujem za vodenje, nasvete in strokovno pomoč. Dr. Curku hvala, da je z mano podelil svojo ekspertizo s področja računalništva in strojnega učenja. Dr. Butali hvala za predlagano edinstveno tematiko in vso pomoč v laboratoriju.

Zahvaljujem se izr. prof. dr. Urošu Petroviču za hitro in strokovno recenzijo naloge.

Rad bi se zahvalil tudi izr. prof. dr. Jerneju Jakšetu, da mi je omogočil uporabo računalnika Katedre za genetiko, biotehnologijo, statistiko in žlahtnjenje rastlin za reševanje računsko zahtevnejših operacij.

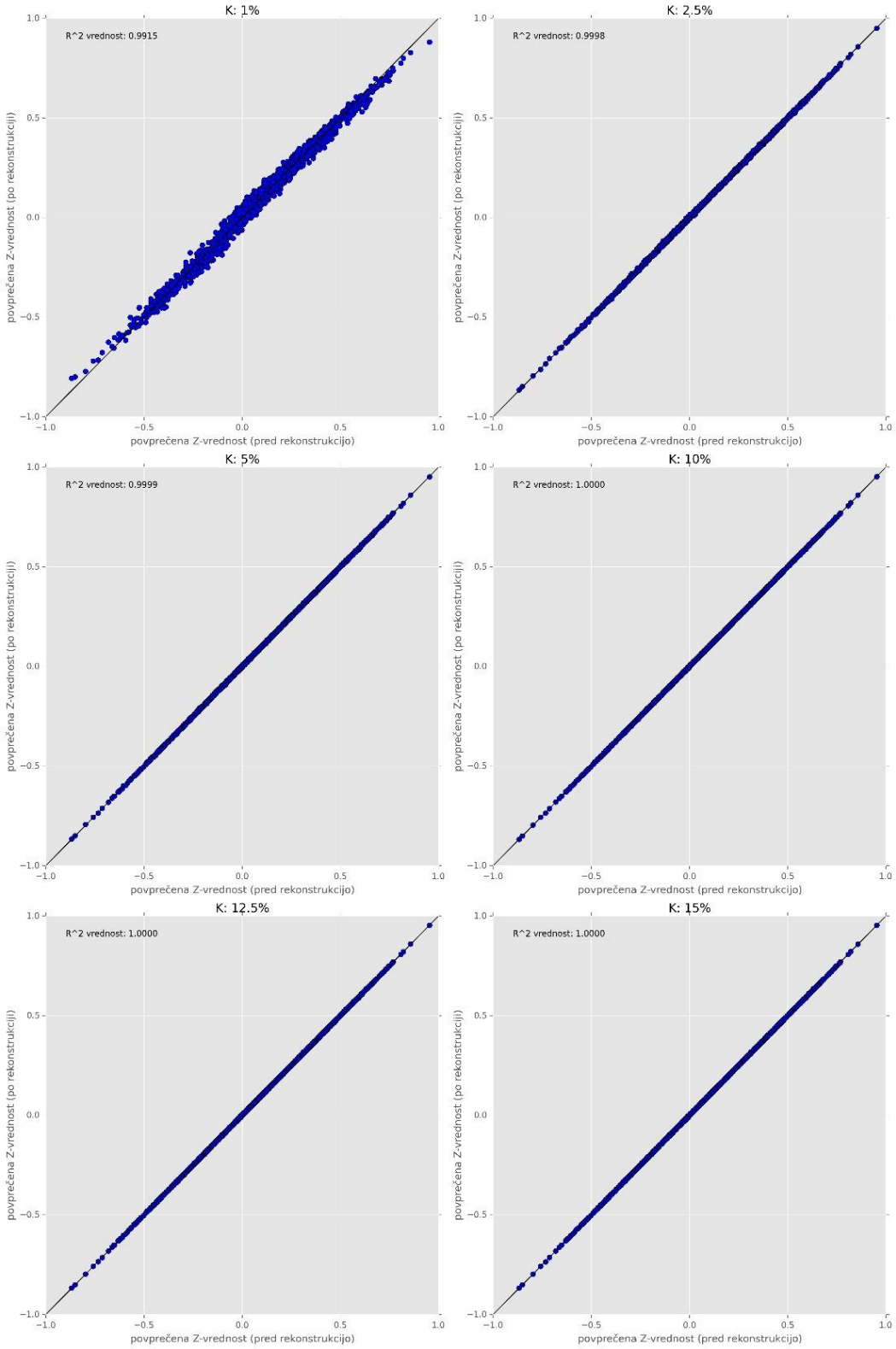
Mag. Francu Kuzmiču se zahvaljujem za lekturo naloge.

Hvala vsem na katedri za biokemijo, ki so mi pomagali pri izvedbi laboratorijskih poskusov.

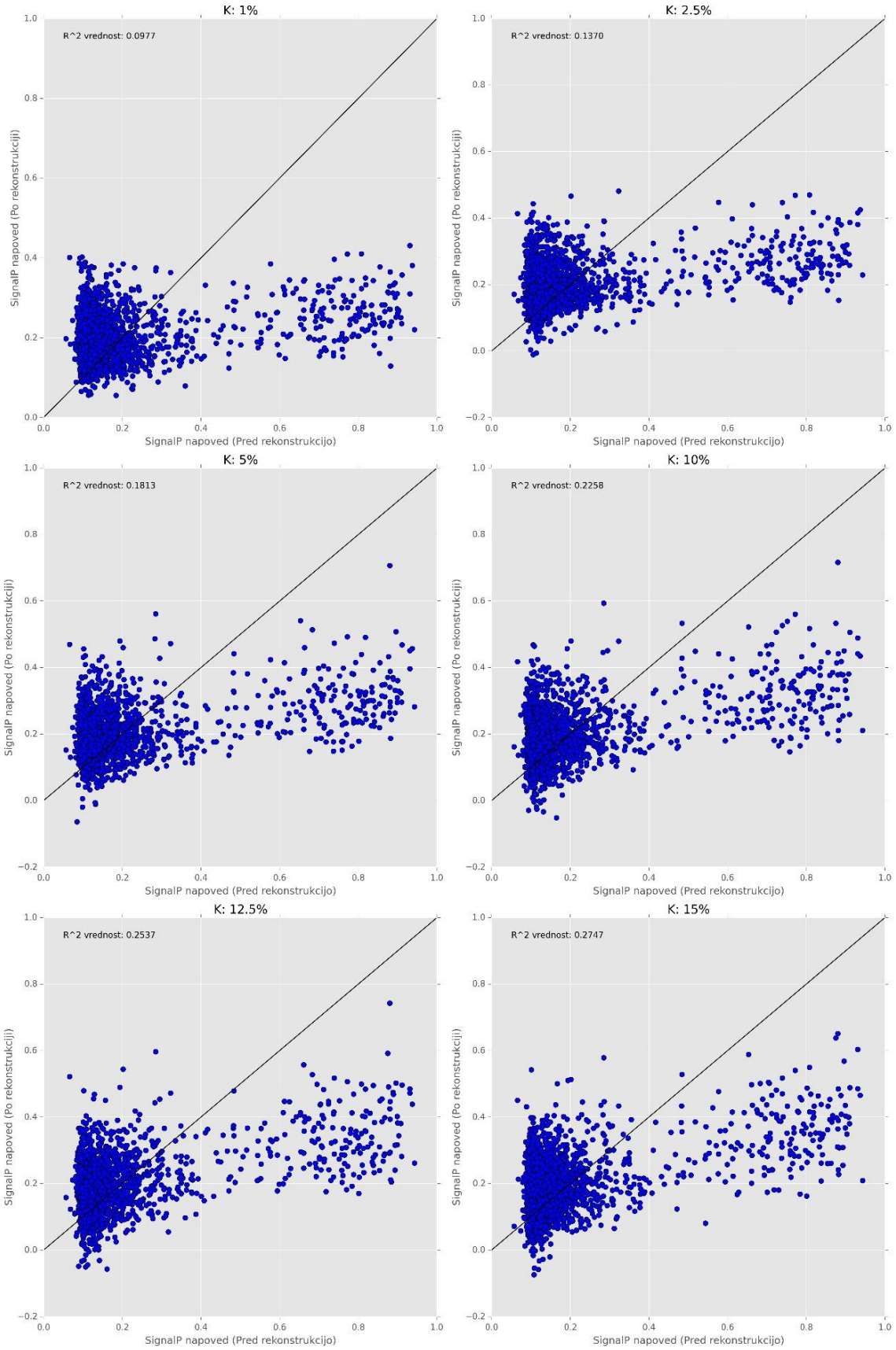
Hvala tudi moji družini in prijateljem za vso podporo in spodbude v času študija in priprave tega dela.

PRILOGA A

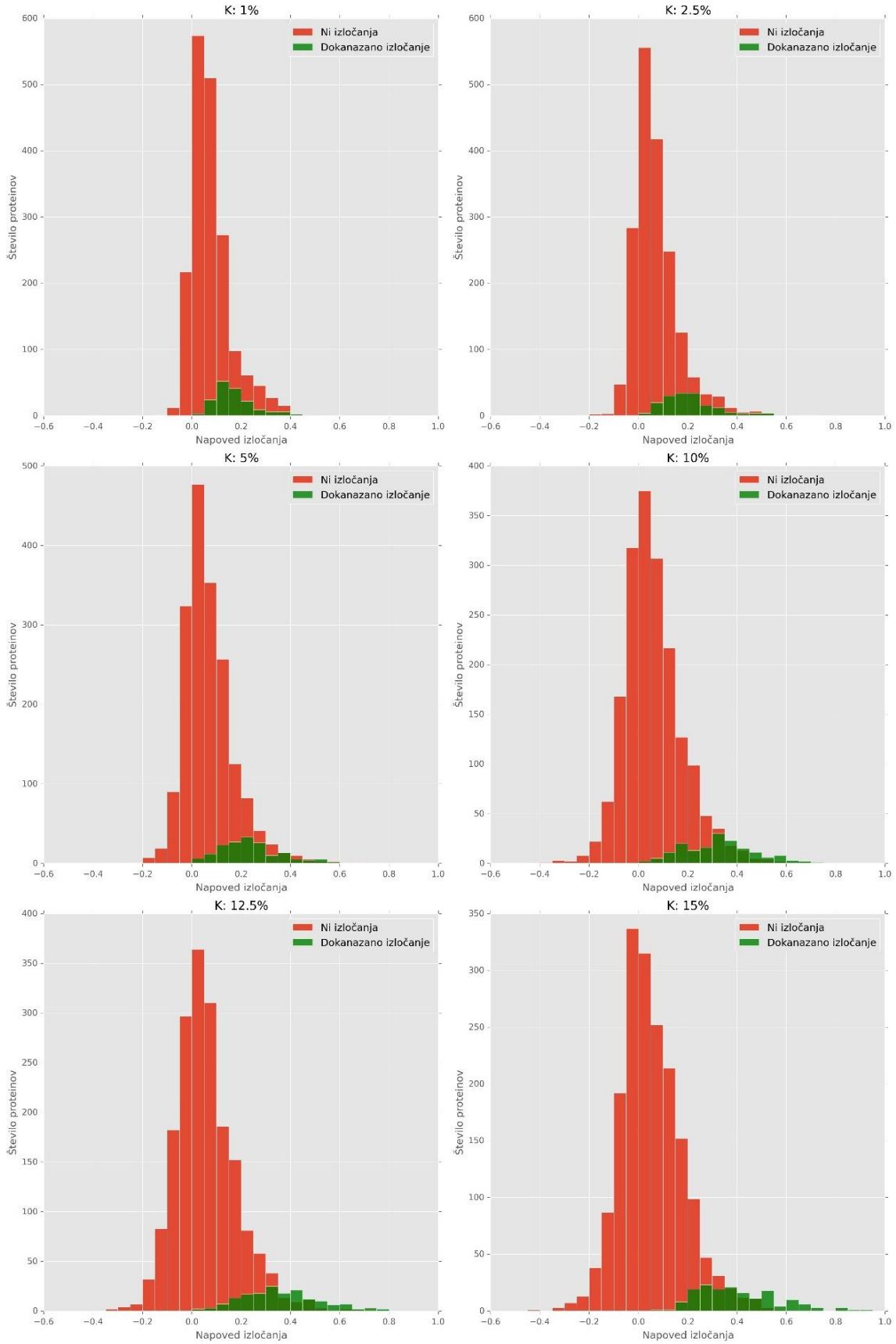
Rekonstrukcije relacij pri različnih rRF



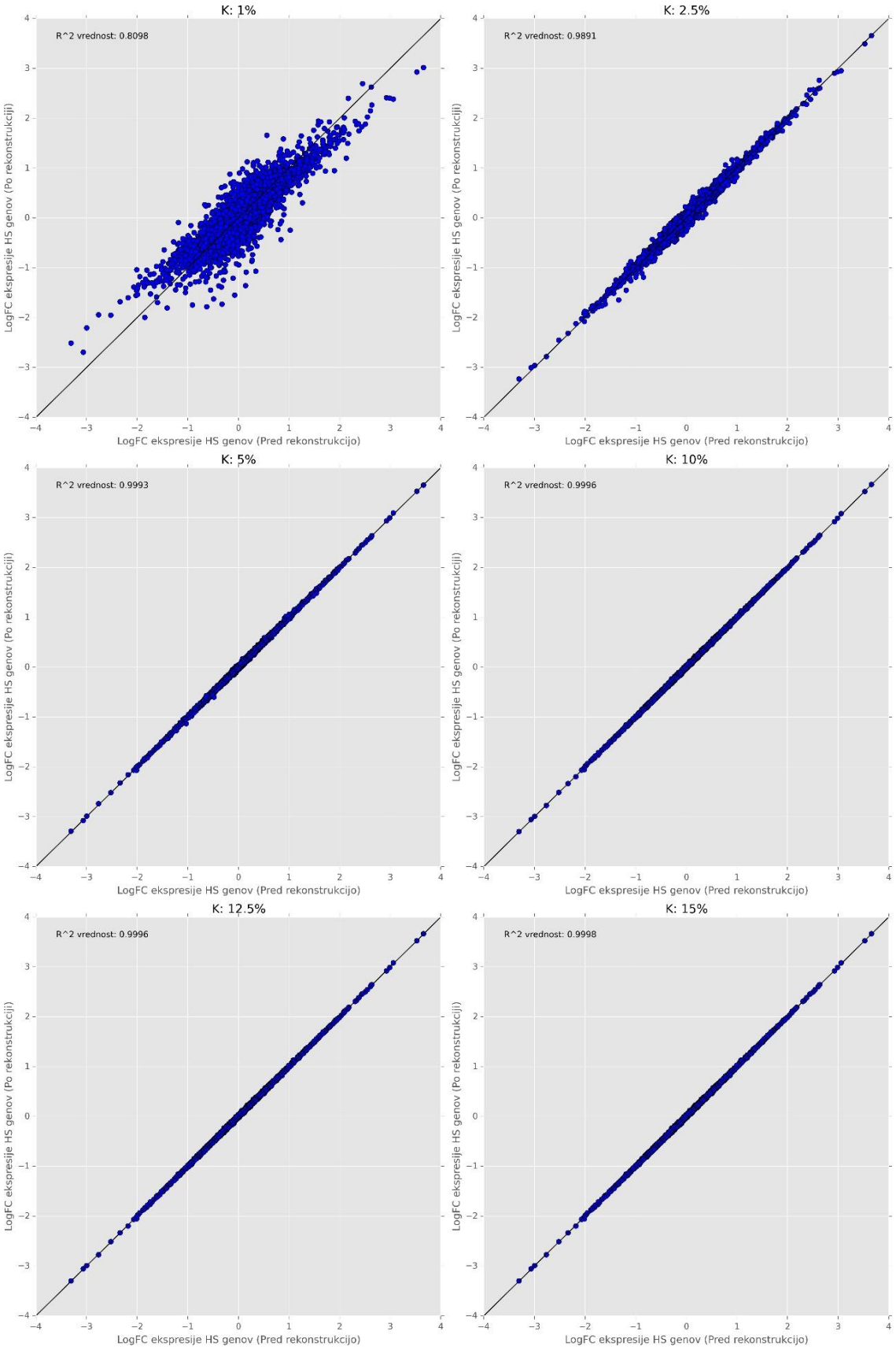
Priloga A1: Korelacija povprečnih Z-vrednosti (catRAPID) pred in po rekonstrukciji.



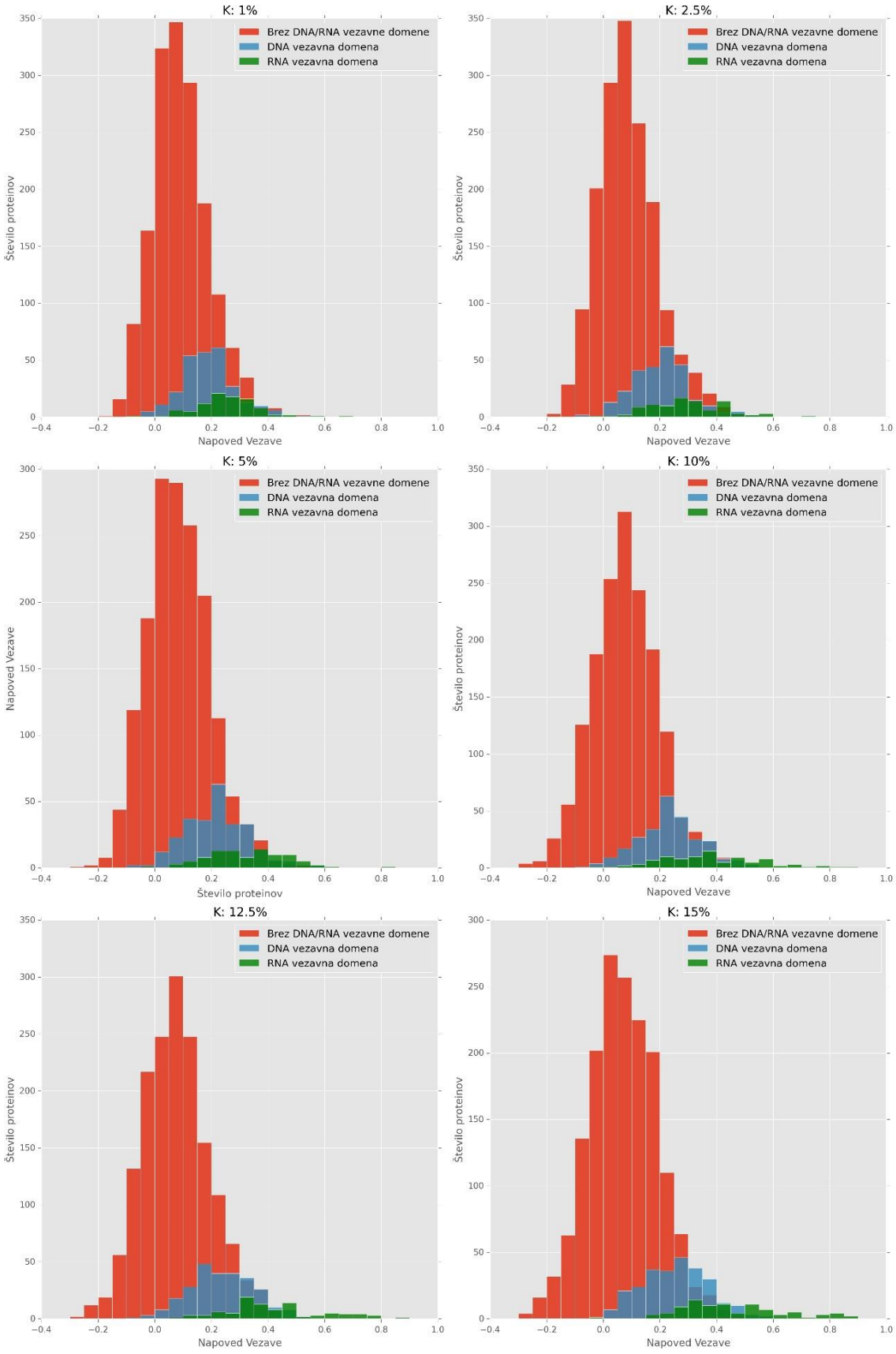
Priloga A2: Korelacija SignalP napovedi pred in po rekonstrukciji.



Priloga A3: Eksperimentalno določeno izločanje proteinov in napoved izločanja po rekonstrukciji.



Priloga A4: Korelacija diferencialnega izražanja človeških genov pred in po rekonstrukciji.



Priloga A5: Napovedovanje prisotnost RNA in DNA vezavnih domen.

PRILOGA B

Obogatitvena analiza seznama glede na catRAPID rekonstrukcijo.

10 najbolj obogatenih GO skupin za vrhnjih 20 proteinov glede na napovedano RNA vezavo (povprečna Z-vrednost catRAPID rekonstrukcije za protein) in hkratno sekrecijo. okrajšave: BP - biološki proces, MF – molekulska funkcija, CC – celični predel.

Skupina	**Opis**	**Domena**	**Št. genov v skupini**	**Št. genov v preseku**	**p-vrednost**	**q-vrednost**
GO:0032991	Kompleks makromulekul	CC	136	6	0,0015	1
GO:0043603	Celični metabolni proces amidov	BP	120	5	0,0053	1
GO:0044267	Celični metabolni proces proteinov	BP	176	6	0,0058	1
GO:0019829	ATPazna aktivnost transporta kationov	MF	14	2	0,008	1
GO:0000041	transporta tranzicijskih kovinskih ionov	BP	14	2	0,008	1
GO:0098796	Kompleks membranskih proteinov	CC	43	3	0,008	1
GO:1901564	Metabolni proces organskih dušikovih spojin	BP	324	8	0,0091	1
GO:0004802	Transketolazna aktivnost	MF	1	1	0,0100	1
GO:0004067	Asparaginazna aktivnost	MF	1	1	0,0100	1
GO:0004345	Glukoza-6-fosfat dehidrogenazna aktivnost	MF	1	1	0,0100	1

PRILOGA C

Geni molekul mRNA, ki kažejo najboljšo vezavo z vrhnjimi 20 proteini iz prioritetnega seznama.

ID ENSEMBL gena	Opis gena
ENSG00000104687	glutathione reductase [Source:HGNC Symbol;Acc:4623]
ENSG00000100290	BCL2-interacting killer (apoptosis-inducing) [Source:HGNC Symbol;Acc:1051]
ENSG00000100122	crystallin, beta B1 [Source:HGNC Symbol;Acc:2397]
ENSG00000124164	VAMP (vesicle-associated membrane protein)-associated protein B and C [Source:HGNC Symbol;Acc:12649]
ENSG00000136160	endothelin receptor type B [Source:HGNC Symbol;Acc:3180]
ENSG00000114646	chondroitin sulfate proteoglycan 5 (neuroglycan C) [Source:HGNC Symbol;Acc:2467]
ENSG00000128791	twisted gastrulation homolog 1 (Drosophila) [Source:HGNC Symbol;Acc:12429]
ENSG00000127529	olfactory receptor, family 7, subfamily C, member 2 [Source:HGNC Symbol;Acc:8374]
ENSG00000143157	pogo transposable element with KRAB domain [Source:HGNC Symbol;Acc:18800]
ENSG00000131910	nuclear receptor subfamily 0, group B, member 2 [Source:HGNC Symbol;Acc:7961]
ENSG00000164615	calcium modulating ligand [Source:HGNC Symbol;Acc:1471]
ENSG00000118434	sperm acrosome associated 1 [Source:HGNC Symbol;Acc:14967]
ENSG00000116824	CD2 molecule [Source:HGNC Symbol;Acc:1639]
ENSG00000133265	HSPA (heat shock 70kDa) binding protein, cytoplasmic cochaperone 1 [Source:HGNC Symbol;Acc:24989]
ENSG00000117010	zinc finger protein 684 [Source:HGNC Symbol;Acc:28418]
ENSG00000162849	kinesin family member 26B [Source:HGNC Symbol;Acc:25484]
ENSG00000122481	RWD domain containing 3 [Source:HGNC Symbol;Acc:21393]
ENSG00000124209	RAB22A, member RAS oncogene family [Source:HGNC Symbol;Acc:9764]
ENSG00000033030	zinc finger, CCHC domain containing 8 [Source:HGNC Symbol;Acc:25265]
ENSG00000108106	ubiquitin-conjugating enzyme E2S [Source:HGNC Symbol;Acc:17895]
ENSG00000129270	matrix metallopeptidase 28 [Source:HGNC Symbol;Acc:14366]
ENSG00000106436	myosin, light chain 10, regulatory [Source:HGNC Symbol;Acc:29825]
ENSG00000154118	junctophilin 3 [Source:HGNC Symbol;Acc:14203]
ENSG00000164850	G protein-coupled estrogen receptor 1 [Source:HGNC Symbol;Acc:4485]
ENSG00000125084	wingless-type MMTV integration site family, member 1 [Source:HGNC Symbol;Acc:12774]
ENSG00000119402	F-box and WD repeat domain containing 2 [Source:HGNC Symbol;Acc:13608]
ENSG00000171222	SCAN domain containing 1 [Source:HGNC Symbol;Acc:10566]
ENSG00000172469	mannosidase, endo-alpha [Source:HGNC Symbol;Acc:21072]
ENSG00000167434	carbonic anhydrase IV [Source:HGNC Symbol;Acc:1375]
ENSG00000139722	vacuolar protein sorting 37 homolog B (S. cerevisiae) [Source:HGNC Symbol;Acc:25754]
ENSG00000168938	peptidylprolyl isomerase C (cyclophilin C) [Source:HGNC Symbol;Acc:9256]
ENSG00000105550	fibroblast growth factor 21 [Source:HGNC Symbol;Acc:3678]
ENSG00000125656	ClpP caseinolytic peptidase, ATP-dependent, proteolytic subunit homolog (E. coli) [Source:HGNC Symbol;Acc:2084]
ENSG00000077348	exosome component 5 [Source:HGNC Symbol;Acc:24662]
ENSG00000151292	casein kinase 1, gamma 3 [Source:HGNC Symbol;Acc:2456]
ENSG00000105370	lens intrinsic membrane protein 2, 19kDa [Source:HGNC Symbol;Acc:6610]

… se nadaljuje

… Nadaljevanje priloge C: Geni molekul mRNA, ki kažejo najboljšo vezavo s vrhnjih 20 proteini iz prioritetnega seznama.

ID ENSEMBL gena	Opis gena
ENSG00000128310	galanin receptor 3 [Source:HGNC Symbol;Acc:4134]
ENSG00000126953	translocase of inner mitochondrial membrane 8 homolog A (yeast) [Source:HGNC Symbol;Acc:11817]
ENSG00000172238	atonal homolog 1 (Drosophila) [Source:HGNC Symbol;Acc:797]
ENSG00000123870	
ENSG00000170956	carcinoembryonic antigen-related cell adhesion molecule 3 [Source:HGNC Symbol;Acc:1815]
ENSG00000107859	paired-like homeodomain 3 [Source:HGNC Symbol;Acc:9006]
ENSG00000138175	ADP-ribosylation factor-like 3 [Source:HGNC Symbol;Acc:694]
ENSG00000110848	CD69 molecule [Source:HGNC Symbol;Acc:1694]
ENSG00000117091	CD48 molecule [Source:HGNC Symbol;Acc:1683]
ENSG00000165970	solute carrier family 6 (neurotransmitter transporter, glycine), member 5 [Source:HGNC Symbol;Acc:11051]
ENSG00000099381	SET domain containing 1A [Source:HGNC Symbol;Acc:29010]
ENSG00000089116	LIM homeobox 5 [Source:HGNC Symbol;Acc:14216]
ENSG00000165502	ribosomal protein L36a-like [Source:HGNC Symbol;Acc:10346]
ENSG00000127364	taste receptor, type 2, member 4 [Source:HGNC Symbol;Acc:14911]
ENSG00000122971	acyl-CoA dehydrogenase, C-2 to C-3 short chain [Source:HGNC Symbol;Acc:90]
ENSG00000135222	casein beta [Source:HGNC Symbol;Acc:2447]
ENSG00000164933	solute carrier family 25, member 32 [Source:HGNC Symbol;Acc:29683]
ENSG00000077616	N-acetylated alpha-linked acidic dipeptidase 2 [Source:HGNC Symbol;Acc:14526]
ENSG00000117569	polypyrimidine tract binding protein 2 [Source:HGNC Symbol;Acc:17662]
ENSG00000172156	chemokine (C-C motif) ligand 11 [Source:HGNC Symbol;Acc:10610]
ENSG00000119121	transient receptor potential cation channel, subfamily M, member 6 [Source:HGNC Symbol;Acc:17995]
ENSG00000025039	Ras-related GTP binding D [Source:HGNC Symbol;Acc:19903]
ENSG00000066422	zinc finger and BTB domain containing 11 [Source:HGNC Symbol;Acc:16740]
ENSG00000169951	zinc finger protein 764 [Source:HGNC Symbol;Acc:28200]
ENSG00000119537	3-ketodihydrosphingosine reductase [Source:HGNC Symbol;Acc:4021]
ENSG00000109927	tectorin alpha [Source:HGNC Symbol;Acc:11720]
ENSG00000123505	adenosylmethionine decarboxylase 1 [Source:HGNC Symbol;Acc:457]
ENSG00000105880	distal-less homeobox 5 [Source:HGNC Symbol;Acc:2918]
ENSG00000164106	stimulator of chondrogenesis 1 [Source:HGNC Symbol;Acc:17036]
ENSG00000171119	neurturin [Source:HGNC Symbol;Acc:8007]
ENSG00000138071	ARP2 actin-related protein 2 homolog (yeast) [Source:HGNC Symbol;Acc:169]
ENSG00000089199	chromogranin B (secretogranin 1) [Source:HGNC Symbol;Acc:1930]
ENSG00000114315	hairy and enhancer of split 1, (Drosophila) [Source:HGNC Symbol;Acc:5192]
ENSG00000142634	EF-hand domain family, member D2 [Source:HGNC Symbol;Acc:28670]
ENSG00000129355	cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4) [Source:HGNC Symbol;Acc:1790]
ENSG00000139233	
ENSG00000171794	undifferentiated embryonic cell transcription factor 1 [Source:HGNC Symbol;Acc:12634]
ENSG00000164326	CART prepropeptide [Source:HGNC Symbol;Acc:24323]

… se nadaljuje

… Nadaljevanje priloge C: Geni molekul mRNA, ki kažejo najboljšo vezavo s vrhnjih 20 proteini iz prioritetnega seznama.

ID ENSEMBL gena	Opis gena
ENSG00000136122	bora, aurora kinase A activator [Source:HGNC Symbol;Acc:24724]
ENSG00000101079	NDRG family member 3 [Source:HGNC Symbol;Acc:14462]
ENSG00000096092	transmembrane protein 14A [Source:HGNC Symbol;Acc:21076]
ENSG00000145545	steroid-5-alpha-reductase, alpha polypeptide 1 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 1) [Source:HGNC Symbol;Acc:11284]
ENSG00000167136	endonuclease G [Source:HGNC Symbol;Acc:3346]
ENSG00000128708	histone acetyltransferase 1 [Source:HGNC Symbol;Acc:4821]
ENSG00000118412	
ENSG00000132703	amyloid P component, serum [Source:HGNC Symbol;Acc:584]
ENSG00000075429	calcium channel, voltage-dependent, gamma subunit 5 [Source:HGNC Symbol;Acc:1409]
ENSG00000158481	CD1c molecule [Source:HGNC Symbol;Acc:1636]
ENSG00000166170	BCL2-associated athanogene 5 [Source:HGNC Symbol;Acc:941]
ENSG00000106733	chromosome 9 open reading frame 95 [Source:HGNC Symbol;Acc:26057]
ENSG00000116678	leptin receptor [Source:HGNC Symbol;Acc:6554]
ENSG00000130958	solute carrier family 35, member D2 [Source:HGNC Symbol;Acc:20799]
ENSG00000122862	serglycin [Source:HGNC Symbol;Acc:9361]
ENSG00000147885	interferon, alpha 16 [Source:HGNC Symbol;Acc:5421]
ENSG00000137944	cysteine conjugate-beta lyase 2 [Source:HGNC Symbol;Acc:33238]
ENSG00000125531	
ENSG00000073861	T-box 21 [Source:HGNC Symbol;Acc:11599]
ENSG00000099785	membrane-associated ring finger (C3HC4) 2, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:28038]
ENSG00000105171	processing of precursor 4, ribonuclease P/MRP subunit (S. cerevisiae) [Source:HGNC Symbol;Acc:30081]
ENSG00000077498	tyrosinase (oculocutaneous albinism IA) [Source:HGNC Symbol;Acc:12442]
ENSG00000101977	MCF.2 cell line derived transforming sequence [Source:HGNC Symbol;Acc:6940]
ENSG00000102805	ceroid-lipofuscinosis, neuronal 5 [Source:HGNC Symbol;Acc:2076]
ENSG00000070190	dual adaptor of phosphotyrosine and 3-phosphoinositides [Source:HGNC Symbol;Acc:16500]
ENSG00000144120	transmembrane protein 177 [Source:HGNC Symbol;Acc:28143]
ENSG00000040731	cadherin 10, type 2 (T2-cadherin) [Source:HGNC Symbol;Acc:1749]
ENSG00000124523	sirtuin 5 [Source:HGNC Symbol;Acc:14933]
ENSG00000136451	vascular endothelial zinc finger 1 [Source:HGNC Symbol;Acc:12949]
ENSG00000065883	cyclin-dependent kinase 13 [Source:HGNC Symbol;Acc:1733]
ENSG00000173862	