Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Hrvoje Mišetić

Filostratigrafska analiza bakterije *Escherichia coli*

Zagreb, 2017.

Ovaj rad izrađen je u Laboratoriju za evolucijsku genetiku Zavoda za molekularnu biologiju Instituta Ruđer Bošković pod vodstvom izv. prof. dr. sc. Tomislava Domazeta-Loše i izv. prof. dr. sc. Damjana Franjevića i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2016./2017.

SADRŽAJ

1. UVOD

1.1. Filogenija bakterija

Bakterije su uz arheje i eukariote jedna od tri domene života. Najčešće su mikrometarske veličine i različitih oblika te su jedni od prvih oblika života na Zemlji. Naseljavaju niz različitih staništa poput vode, tla, termalnih izvora i unutrašnjosti Zemljine kore. Također žive u simbiotskom i parazitskom odnosu s biljkama i životinjama (Rappé and Giovannoni, 2003).

Razumijevanje evolucijske prošlosti bakterija predstavlja dugoročni cilj mikrobiologije i bioloških znanosti općenito. Poznavanje evolucijskih odnosa između pojedinih skupina bakterija je ključno za odgovaranje na mnoga fundamentalna pitanja kao što su podrijetlo prve stanice, metabolizma, fotosinteze, procesa prijenosa informacija te uostalom i podrijetlo predačke eukariotske stanice. Za određivanje filogenije bakterija, trebaju se riješiti sljedeća dva ključna problema, a to su postavljanje dobro definiranog kriterija za identifikaciju glavnih skupina unutar bakterija i razumijevanje kako su te skupine povezane jedna s drugom te kako su se razdvojile od zajedničkog pretka (Gupta and Griffiths, 2002).

Prvotni pokušaji klasifikacije temeljeni su na sličnosti morfoloških, biokemijskih i fizioloških obilježja. Prokarioti su jedno vrijeme bili podijeljeni na dvije skupine te su prvu skupinu činile bakterije koje imaju sposobnost oksigene fotosinteze, a drugu sve ostale bakterije (Levine, 1975). U to vrijeme pitanja kako su te dvije skupine povezane ili kako su nastale od zajedničkog pretka uopće nisu ozbiljno razmatrana.

Veliki zaokret po tom pitanju nastao je spoznajom da sljedovi bioloških makromolekula u različitim vrstama sadrže ogromne količine kvalitativnih i kvantitavnih obilježja koja dolaze direktno od zajedničkog pretka (Zuckerkandl and Pauling, 1965). Na temelju sravnjenja sljedova homolognih gena/proteina, broj i vrsta promjena unutar slijeda mogu biti utvrđene između različitih vrsta. To čini temelj zaključivanja o genealoškim odnosima između vrsta baziranom na molekularnim sljedovima. Najranija primjena ovog pristupa bila je od strane Carla Woesea i njegovih kolega koji su koristeći 16S rRNA sljedove zaključili da prokariote čine dvije različite domene – arheabakterije (lat. *Archaea*) i eubakterije (lat. *Bacteria*) koje su međusobno različite i neovisnog podrijetla od univerzalnog pretka (Fox et al., 1980;

Woese, 1987). Na temelju oligonukleotidnih sljedova i obrascu grananja u 16S rRNA stablima, inicijalno je predloženo 10-12 grupa unutar bakterija (Fox et. al., 1980; Woese et al., 1985; Woese, 1987).

U vrijeme kada je ovo predloženo biološke baze su bile male i limitirane što je doprinijelo tome da te grupe budu jasno odvojene jedna od druge zbog dugih nerazgranatih grana koje su ih razdvajale unutar filogenetskog stabla. Porastom količine dostupnih sekvenciranih sljedova te nekoć prazne grane počele su se popunjavati što je uzrokovalo teže i manje precizno razlikovanje između tih skupina na temelju rRNA sljedova (Ludwig and Schleifer, 1999; Ludwig and Klenk, 2001). S vremenom ova podjela se počela širiti dodavanjem novih skupina bakterija. Neke od tih novih skupina su imale tek nekoliko vrsta ili su nekoć bile dio neke druge veće skupine. Jasno definiranih i objektivnih kriterija za određivanje skupina unutar bakterija nije bilo (Gupta and Griffiths, 2002.).

Kako su različite bakterijske skupine međusobno povezane i kako su se odvojile od zajedničkog pretka je drugi ključni problem bakterijske filogenije. Evolucijski odnosi između pojedinih skupina bakterija nisu u potpunosti objašnjeni filogenetskim stablima temeljenim na rRNA sljedovima ili sljedovima različitih proteina i ti nedostaci filogenetskih stabala su široko prihvaćeni. Nemogućnost filogenetskih stabala da u potpunosti riješe pitanje evolucijskih odnosa se javlja zbog njihove ovisnosti o nizu različitih parametara i pretpostavki kao što su pouzdanost sekvenciranih sljedova i sravnjenja, uklanjanje ili ostavljanje regija slijeda u filogenetskoj analizi, broj i rang ispitanih vrsta, redoslijed dodavanja vrsta prilikom sravnjenja, razlike u evolucijskoj brzini i sastavu baza svake vrste, korištena filogenetska metoda itd. (Woese, 1987; Lake, 1991; Gupta, 1998; Ludwig and Klenk, 2001). Ne postoje standardni kriteriji te tako promjena u samo jednom parametru može imati utjecaj na raspored grananja vrsta u filogenetskom stablu. Iako je utjecaj ovih varijabli na visoko srodne vrste malen, može vrlo jako utjecati na redoslijed grananja glavnih skupina (Ludwig and Klenk, 2001).

1.2. Istraživanje filogenije bakterija

U slučaju malenih jednostaničnih organizama poput bakterija morfološka, fiziološka i ostala obilježja nisu dovoljno informativna da bi se koristila kao filogenetski markeri (Patwardhan et al., 2014). Trenutačne metode korištene za filogenetsku klasifikaciju bakterija većinom se oslanjaju na sljedove 16S rRNA gena.

16S rRNA koja čini malu podjedinicu ribosoma bakterija smatra se zlatnim standardom za njihovu filogenetsku analizu zbog sljedećih razloga: prisutna je u svim bakterijama, sadrži oko 1500 baza što je čini dovoljno dugom za analizu, sadrži visoko očuvane regije za usporedbu vrlo udaljenih vrsta i varijabilne regije za usporedbu bližih vrsta te ima potpuno očuvane regije koje omogućuju korištenje univerzalnih početnica za detekciju PCR-om ili sekvenciranje (Kitahara and Miyazaki, 2013).

Uz sve navedeno za 16S rRNA gene smatra se da se ne prenose horizontalnim prijenosom koji je česta pojava kod bakterija te predstavljaju stvarnu filogeniju mikroorganizama. Nemogućnost horizontalnog prijenosa 16S rRNA gena objašnjena je „hipotezom kompleksnosti“ (Jain, Rivera and Lake, 1999.). Prema toj teoriji produkti informacijskih gena koji su uključeni u replikaciju DNA, transkripciju i translaciju često djeluju kao esencijalni kompleksi s više komponenata te se smatra da takvi geni uglavnom koevoluiraju skupa sa svojim vezujućim partnerom u istom kompleksu. Pri horizontalnom prijenosu gena trebalo bi doći do simultanog prijenosa više gena s različitih lokusa kako bi kompleks bio funkcionalan što postaje manje vjerojatno s porastom broja komponenti u kompleksu.

Protivno navednoj teoriji, postoje određeni dokazi horizontalnog prijenosa u prirodi i eksperimentalne potvrde. rRNA geni se transkribiraju kao operoni (*rrn* operon) i transkripti su procesirani različitim endo- ili eksoribonukleazama da se dobiju 16S rRNA, 23S rRNA, 5S rRNA i neke tRNA iako postoje neke iznimke (Lim, Furuta and Kobayashi, 2012). Određene bakterije posjeduju više kopija *rrn* operona koji se ponekada vrlo razlikuju. Tako na primjer termofilna bakterija *Thermoanaerobacter tengcongensis* sadrži četiri *rrn* operona, i 16S rRNA slijed operona *rrnC* pokazuje samo 88,4% homolgije s ostalim operonima (*rrnA, rrnaB* i *rrnD*). Gen za 16S rRNA operona *rrC* nije pseudogen jer pravilno stvara očuvanu sekundarnu strukturu 16S rRNA. Sam slijed puno je sličniji (95%) drugim vrstama

roda *Thermoanarobacter* (*T. subterraneus* SL9 i *T. keratinophilus* 2KXI) iz čega se može zaključiti da je ovaj gen bio prenesen u pretka *T. tengcongensis* i ostao kao funkcionalan gen u ovom organizmu tijekom evolucije. Slični primjeri pronađeni su još u bakterijama *Desulfotomaculum kuznestovii*, *Thermobispora bispora* R51, *Thermonospora chromogena* i *Halobacterium marismortui* (Acinas et al., 2004).

Ovi primjeri su dokaz da je horizontalni prijenos 16S rRNA u prirodi moguć i opvrgavaju teoriju kompleksnosti. Kimerni 16S rRNA geni su također pronađeni između vrlo udaljenih vrsta (Miller et al., 2015) i između bliskih linija (Wang and Zhang, 2000; Schouls, Schot and Jacobs, 2000; Eardly et al., 2005) što navodi da djelomični fragmenti gena za 16S rRNA se isto mogu prenositi horizontalnim prijenosom. U pokusima horizontalnog prijenosa gena *in-vitro* dokazano je da *E. coli* može primiti strani 16S rRNA gen sa samo 80.9% identičnosti koji inače pripada različitom filogenetskom razredu (Kitahara and Miyazaki, 2013) što isto navodi da je horizontalni prijenos tog gena ipak moguć. Filogenije dobivene na temelju tog markera moraju biti zato pažljivo analizirane i interpretirane.

Jedan od prijedloga rješavanja srodstvenih odnosa bakterija bili su i očuvani indeli (insercije ili delecije). Kada su očuvani indeli definiranog slijeda i duljine pronađeni na istoj poziciji u određenom proteinu (ili genu) kod svih članova iz jedne ili više grupa bakterija onda je najjednostavnije objašnjenje da je indel nastao samo jednom u zajedničkom pretku tih vrsta. Na temelju prisutnosti ili nedostatka zajedničkog indela, različite vrste mogu biti podijeljene u različite skupine s obzirom na posjedovanje ili nedostatak indela (Gupta 1998, 2000). Dobro definirani indeli u proteinskim sljedovima služe kao korisni markeri za evolucijske događaje budući da se očekuje da sve vrste nastale iz predačke stanice u kojoj je indel nastao će imati taj proteinski potpis, a sve one vrste koje su postojale prije te predačke stanice ili nisu nastale iz nje neće imati taj indel.

Identificiran je velik broj očuvanih indela u različitim proteinima koji služe kao markeri za određivanje redoslijeda grananja različitih skupina bakterija (Gupta, 1998; Gupta, 2000; Griffiths and Gupta, 2002). Rezultati dobiveni ovom metodom navode da ovaj pristup omogućuje pouzdan i visoko konzistentan način određivanja relativnog redoslijeda grananja bakterijskih skupina (Gupta and Griifiths, 2002).

Nemogućnost 16S rRNA ili različitih proteinskih markera da riješe evolucijske odnose bakterija je dovelo do mišljenja da je taj bitan problem praktički nerješiv. To je dovelo do rastućeg prihvaćanja ideje da se većina glavnih skupina unutar bakterija odvojila od zajedničkog pretka u sličnom trenutku (Doolittle,1999; Ludwig and Schleifer, 1999 ; Ludwig and Klenk, 2001).

Razvojem metoda sekvenciranja novonastali podaci poput čitavih bakterijskih genoma raznih vrsta revolucionarizirali su naše razumijevanje mnogih uloga bakterija. Iako postoji veliki broj sekvenciranih bakterijskih genoma, većina njih je bila odabrana za sekvenciranje na temelju svoje fiziologije. Kao rezultat toga, mogućnosti istraživanja su bile ograničene zbog visoko pristrane filogenetske distribucije (Hugenholtz, 2002; Eisen, 2000; Wu et al.,2009). Kako bi se to promijenilo pokrenut je projekt „Genomic Encyclopedia of Bacteria and Archaea“ u sklopu kojeg su sekvencirani i analizirani mikrobiološki genomi na temelju svojih evolucijskih odnosa kako bi se dobilo maksimalno moguće znanje iz postojećih mikrobioloških genomskih podataka kao i iz onih genoma koji će se tek sekvencirati (Wu et al., 2009).

Organizmi za sekvenciranje odabrani su na temelju svog položaja u filogenetskom stablu koje se temeljilo na 16S rRNA i obratila se pozornost na divergentne linije koje nemaju predstavnike sa sekvenciranim genomom iako je vrsta službeno opisana. Postavilo se pitanje je li odabir genoma za sekvenciranje na temelju 16S rRNA slijeda pouzdan način povećanja filogenetske raznolikosti sekvenciranih genoma. To se pitanje postavilo budući da je poznato da jedan gen, makar to bio i 16S rRNA gen, ne može savršeno predvidjeti filogenetske obrasce na razini genoma (Achtman and Wagner, 2008; Beiko, Doolittle and Charlebois, 2008). Kako bi se to istražilo, kreirano je „genomsko stablo“ svih dotada kompletno sekvenciranih genoma i mjerio se relativan doprinos spomenutog projekta u povećanju filogenetske raznolikosti sekvenciranih genoma koristeći mjere filogenetske različitosti (Pardi and Goldman, 2007). Dokazano je da genomi sekvencirani u sklopu ovog projekta čine 2.8-4.4 puta veću filogenetsku raznolikost nego nasumično uzorkovani genomi koji su sekvencirani ranije. Ova analiza ukazuje da iako 16S rRNA geni nisu savršeni indikatori evolucije organizama, njihovi filogentski odnosi su ispravan prediktor filogenetske raznolikosti unutar bakterijskih genoma (Wu et al., 2009.).

Pri određivanju bakterijske filogenije trebalo bi izbjegavati da se ona temelji na samo jednom genu jer to odražava evoluciju tog jednog gena koji ima svoju određenu brzinu evolucije, evolucijsku prošlost i različitu količinu filogenetskog signala (Patwardhan et al., 2014). Bolje je koristiti proteinske sljedove koji su očuvani na razini aminokiselina jer filogenetske analize temeljene na proteinskim sljedovima su manje sklone pristranosti ovisnoj o nukleotidnom sastavu koja je primjećena i u 16S rRNA (Hasegawa and Hashimoto, 1993; Loomis WF and Smith DW, 1990; Lockhart et al.,1992; Baldauf et al., 2000) koja se smatra zlatnim standardom te je to ostala zbog poteškoća u kloniranju protein kodirajućih gena u različitim vrstama.

Sekvenciranje genoma je omogućilo promjenu ove situacije. Svaki kompletno sekvencirani genom sadrži sljedove za sve protein kodirajuće gene u organizmu. Sada je moguće ne samo graditi stablo na temelju željenog protein kodirajućeg gena već postoji opcija da se ulanča više genskih sljedova te se izgradi stablo na razini genoma. Sadržavajući više filogenetskih signala takva „genomska stabla“ ili „super-matrix stabla“ su puno manje sklona stohastičkim greškama od onih sagrađenih na temelju jednog gena (Jeffroy et al., 2006). Iako je ovo vrlo moćan pristup treba spomenuti da je osjetljiv na sistemske greške izazvane pristranošću sastava baza, nerealnim evolucijskim modelima i neodgovarajućim taksonomskim uzorkovanjem (Jeffroy et al., 2006; Brocchieri, 2001; Foster and Hickey, 1999).

U konačnici zasada se metoda „genomskih stabala“ ili „super-matrix stabala“ smatra najboljom. Krenulo se od vremena kada za jednostanične organizme poput bakterija morfologija, fiziologija i ostala obilježja nisu bila dovoljno informativna pa su otkriveni molekularni markeri i naposljetku sekvenciranjem genoma količina podataka i mogućih parametara je naglo porasla što će biti trend i u budućnosti. Ogroman broj novonastalih podataka o bakterijskim genomima bit će svakao izazov, ali će omogućiti da se u potpunosti razjasni njihova evolucijska prošlost.

1.3. Zadnji zajednički predak bakterija

Prve grupe koje su se pojavila u stablu života temeljenom na ribosomalnim RNA (rRNA) sljedovima su bila hipertermofilna. Što je dovelo do hipoteze da je univerzalni predak i moguće još živući organizam bio hipertermofilan. Točnije, bakterijska stabla utemeljena na rRNA prikauzuju da članovi jedinih poznatih hipertermofilnih koljena, Aquificales i Thermotogales, se odvajaju prvi odmah prije umjerenih termofila nakon kojih slijede ostale mezofilne vrste. Ovi dokazi skupa s tradicionalnim ukorijenjavanjem bakterija u univerzalnom stablu života navode da je posljednji univerzalni zajednički predak bio hipertermofilan (Stetter, 1996).

Povrh toga, činjenica da grane koje vode do hipertermofilnih bakterija su kraće od onih koje vode do mezofilnih vrsta se uzelo kao naznaka da privlačenje dugih grana (engl. long branch attraction) nije odgovorno za tako rano odvajanje i da hipertermofilne bakterije su živi fosili, tj. sadrže mnoga obilježja zadnjeg univerzalnog zajedničkog pretka. Nadalje, ova teorija je bila podupirana još takozvanom hipotezom o hipertermofilnom Edenu koja govori da je život nastao i prvotno se razvijao u vrlo vrućem okolišu sve do pojave zadnjeg zajedničkog pretka svih triju domena života (Woese, 1987; Pace, 1991; Olsen et al., 1994; Gogarten-Boekels et al., 1995; Stetter, 1996; Imai et al., 1999; Nisbet and Sleep, 2001)

Međutim, pojavila se ideja da prve grane bakterijske filogenije su ipak pogrešne zbog privlačenja dugih grana (engl. long branch attraction) budući da vanjska grupa (engl. outgroup) (ili arheje ili eukarioti) je vrlo udaljeno srodna. Primjenom Slow-Fast metode na veliki broj prokariotskih rRNA i dokazano je da je privlačenje dugih grana uzrok ranom odvajanju hipertermofilnih bakterija (Brochier and Philippe, 2002.). Uz privlačenje dugih grana spominje se i visoki udio gvanina i citozina u većini arheja i bakterija iz koljena Aquificales i Thermotogales (Galtier, Tourasse and Gouy, 1999) koji još dodatno objašnjava zašto se hipertermofilne bakterije rano odvajaju.

Kao ne-hipertermofilni predak bakterija predloženo je koljeno Planctomycetales koje ima niz osobitosti (Brochier and Philippe, 2002.). Ova skupina bakterija posebna je po tome što se svi njeni članovi dijele pupanjem i nemaju peptidoglikan u svojim staničnim stijenkama (Fuerst, 1995). Najintrigantnije obilježje ove skupine je postojanje jednostruke ili dvostruke membrane oko kromosoma koje

je prisutno kod vrsta *Gemmata* sp. i *Pirellula* sp. što je bilo uspoređivano s eukariotskom jezgrom (Fuerst, 1995).

U daljnjim istraživanjima hipoteza o ne-hipertermofilnom pretku je bila detaljnije ispitana. Kako je ta hipoteza donesena metodom višestrukog sravnjenja koje ne ukjučuje pozicije koje sadrže filogenetski šum provjerena je robustnost te metode. Utvrđeno je da broj nukleotidnih pozicija korišten u analizi nije dovoljno visok i stoga analiza Brochier i Philippe nije dovoljno robusna (Di Gulio, 2003). Nadalje, različite metode odabira pozicija bez filogenetskog šuma u rRNA sravnjenju postavljaju koljena Aquificales i Thermotogales kao prve linije divergencije, a koljeno Planctomycetales kao treću granu divergencije u filogenetskom stablu sagrađenom na temelju tih pozicija (Di Gulio, 2003). Daljnje filogenetske analize koje su uključivale konkatamere dobivene fuzijom dvadeset različitih proteina korištene su u pet različitih metoda te su još jednom potvrdili da su se hipertermofilni organizmi Thermotogales i Aquificales prvi odvojili u bakterijskoj domeni, a ne mezofilno koljeno Planctomycetales (Barion et al., 2007).

1.4. Genomska filostratigrafija

Makroevolucijski trendovi tradicionalno se proučavaju analizom fosila, komparativnom morfologijom ili evo-devo pristupima. Porastom količine dostupnih genomskih sljedova i podataka vezanih za raznovrsne rodove, moguće je dodati novu razinu analize: genomsku filostratigrafiju (Domazet-Lošo, Brajković and Tautz, 2007). Princip genomske filostratigrafije podrazumijeva da genom svake postojeće vrste sadrži dijelove slike evolucijske prošlosti, a nagli razvoj metoda sekvenciranja omogućio je akumulaciju genomskih podataka različitih organizama te su mogućnosti otkrivanja te prošlosti veće nego ikada do sada.

Genomska filostratigrafija je analitička metoda koja se temelji na modelu isprekidane evolucije proteinskih obitelji koji pretpostavlja da su proteinske obitelji nastale putem gena osnivača na raspršen način tijekom evolucijskog perioda (Domazet-Lošo, Brajković and Tautz, 2007; Domazet-Lošo and Tautz, 2008). Geni osnivači su geni koji predstavljaju evolucijske novine u proteinskom slijedu (Domazet-Lošo, Brajković and Tautz, 2007; Domazet-Lošo and Tautz, 2003; Khalturin et al., 2009), tj. nisu jednostavno duplikacije postojećih gena ili geni s opet promiješanim funkcionalnim domenama. Točnije, oni predstavljaju nove funkcionalne proteine ili proteinske domene koje prije nisu bile prisutne u genomu barem ne u novom funkcionalnom obliku. Jednom kada se takva nova funkcionalna domena pojavi, očekuje se da će sačuvati svoj primarni proteinski slijed u tolikoj mjeri da će i dalje moći biti pronađena metodama pronalaska sličnih sljedova (Domazet-Lošo and Tautz, 2003).

Na temelju toga moguće je odrediti evolucijsku pojavnost gena ili genske obitelji identifikacijom homologa u stablu života. Koristeći ovu metodu na svim genima neke vrste moguće je istraživanje obrazaca ishodišta gena na razini čitavog genoma u različitim evolucijskim periodima.

Prva filostratigrafska mapa napravljena je na genomu vinske mušice (lat. *Drosophila melanogaster*) (Domazet-Lošo, Brajković and Tautz, 2007). Filostratigrafski podaci kombinirani su s podacima o ekspresiji gena tijekom stadija embrija vinske mušice. Geni su podijeljeni u 12 grupa s obzirom na pojavu njihovog gena osnivača u filogeniji i te grupe nazivaju se genomskim filostratumima.

Također, filostratigrafska metoda je bila korištena pri određivanju evolucijskog

podrijetla gena vezanih za ljudske genetičke bolesti. Zaključeno je da je većina takvih gena već bila prisutna u eukariotskom pretku, a sljedeći put su pojavili u većem broju tijekom perioda evolucije višestaničnosti (Domazet-Lošo and Tautz, 2008).

Kombinacija filostratigrafije i podataka o genskoj ekspresiji tijekom specifičnih faza embriogeneze omogućili su određivanje kumulativnog indeksa koji reflektira evolucijsku starost transkriptoma u određenoj fazi ontogenije. Koristeći ontogeniju i razvoj odrasle jedinke zebrice (*Danio rerio*) kao modela, otkriveno je da se tijekom filotipske faze eksprimira najstariji transkriptomski skup, a da je mlađi skup gena eksprimiran tijekom ranog i kasnog razvoja tako da vjerno zrcali model pješčanog sata morfološke divergencije. Reproduktivno aktivne životinje imaju najmlađi transkriptom s velikim razlikama između muških i ženskih jedinki. Starenjem životinje eksprimiraju se sve stariji geni. Usporedbom sa sličnim podacima dobivenih od nematoda i muha potvrđeno je da se ovaj uzorak pojavljuje kroz koljeno. Ovi rezultati navode da stariji transkriptomi označavaju filotpsku fazu i da filogenetske razlike u ostalim ontogenskim stadijima koreliraju s ekspresijom novo evoluiranih gena (Domazet-Lošo and Tautz, 2010a).

Nadalje, metoda je bila primjenjena u praćenju podrijetla gena odgovornih za rak (Domazet-Lošo and Tautz, 2010b). Pronađena su dva kritična trenutka kada je pojava proteinskih domena povezanih s rakom najveća, jedan je u trenutku nastanka prve stanice, a drugi oko perioda nastanka višestaničnih životinja. Uvidjela se veza između višestaničnosti i pojave raka, a drugi maksimum u filostratigrafskoj mapi nagoviješćuje da je kompleksni selekcijski proces na više razina doveo do pojave višestaničnosti (Domazet-Lošo and Tautz, 2010b).

Osim toga metoda se primjenila u dokazivanju pojavljivanja gena na temelju *de novo* evolucijskog modela, a ne dotada šire prihvaćenog modela duplikacije i divergencije postojećih gena (Neme and Tautz, 2013).

Kombinacija genomske filostratigrafije i podataka o ekspresiji omogućila je da se dozna više o evolucijskoj prošlosti osjetilnih organa glave kralježnjaka. Suprotno tradicionalnim predviđanjima, otkriveno je da su dominantni adaptivni signali u analiziranim osjetilnim strukturama prethodili evolucijskoj pojavi kralježnjaka. Vodeći adaptivni signali u trenutku prijelaza bilateralnih životinja (engl. bilaterians) i svitkovaca sugeriraju da je vidni sustav bio je prvi osjetilni sustav koji je evoluirao, a mirisni, slušni organi i bočna pruga pokazuju snažnu vezu s precima na prijelazu

urosvitkovaca i kralježnjaka. Jedina struktura koja je uistinu inovacija kralježnjaka su derivati neuralnog grebena, trigeminus ganglion i adenohipofiza . Također, pronađeni su dokazi da su kranijalne plakode evoluirale ranije od neuralnog grebena unatoč njihovoj predloženoj embrionalnoj povezanosti (Šestak et al., 2013)

Filostratigrafski profili zebrice otkrili su da mozak kralježnjaka ima podrijetlo u svitkovcima (Šestak and Domazet-Lošo, 2014). Otkrivena su tri važna razdoblja u evolucijskoj povijesti mozga zebraste ribice. Najstariji period odgovara preadaptivnim događanjima prvih životinja i pojavu živčanog sustava na prijelazu životinja i pravih mnogostaničara. Pojava svitkovaca označava sljedeću fazu gdje je pronađen najveći ukupni adaptivni otisak u gotovo svim analiziranim dijelovima mozga. Ovi pronalasci podržavaju ideju da mozak kralježnjaka je evoluirao neovisno unutar linije proteosoma. Naposljetku, pri pojavi kralježnjaka detektiran je izražen signal iz dorzalnog telencephalona, što je u skladu s klasičnom teorijom koja smatra veliki mozak izvornom novinom kralježnjaka. Sveukupno ovi rezultati otkrivaju postupnu adaptivnu prošlost mozga kralježnjaka gdje je većina njegove organizacije već bila prisutna u pretku svitkovcu (Šestak and Domazet-Lošo, 2014) .

Iz navedenih primjera može se uvidjeti da genomska filostratigrafija samostalno i još u kombinaciji s dodatnim setovima podataka poput uzoraka genske ekspresije tijekom razvoja ima vrlo široku primjenu u istraživanju evolucijske prošlosti te je uočljivo kako je količina dostupnih podataka pogotovo genomskih rasla tako je i primjena ove metode postajala veća.

1.5. *Escherichia coli*

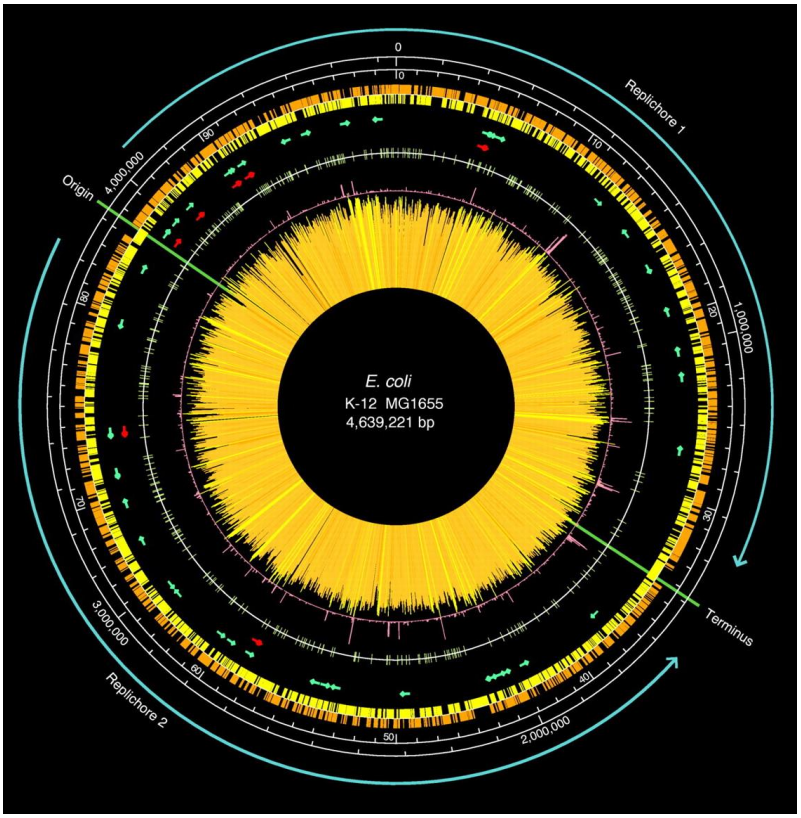
Escherichia coli je gram-negativna, fakultativno anaerobna, štapićasta, koliformna bakterija koja se obično nalazi u donjim dijelovima crijeva toplokrvnih organizama (Tenaillon et al. 2010; Singelton 1999). Bezopasni sojevi su dio normalne flore crijeva i doprinose domaćinu proizvodnjom vitamina K_2 i onemogućuju kolonizaciju patogenih bakterija u crijevu. Patogeni sojevi uzrokuju infekcije probavnog, mokraćnog, dišnog i živčanog sustava (Blattner et al., 1997).

Najčešće je proučavan prokariotski modelni organizam i važna je vrsta u području biotehnologije i mikrobiologije gdje služi kao domaćin za rekombinantnu DNA.

1.5.1. *Escherichia coli* K-12

Soj *Escherichia coli* K-12 je izoliran iz uzorka stolice pacijenta koji se oporavljao od difterije i označan je kao K-12 1922. godine na Sveučilištu Stanford (Bachmann, 1972). Soj je bio korišten tijekom 40-ih godina prošloga stoljeća za proučavanje metabolizma dušika i bio je pohranjen u bazi ATCC (American Type Culture Collection). Dalje je korišten u eksperimentima biosinteze triptofana (Tatum and Lederberg, 1947) iako ima F+ λ+ fenotip (Lederber, 2014). Tijekom vremena od ovog soja je nastao niz drugih sojeva najčešće tretiranjem različitim agensima. Najčešće se koristi kao modelni organizam u laboratorijima.

Genom soja K-12 je kružna molekula DNA od 4,639,221 parova baza koja sadrži 4,288 protein-kodirajuća gena (organizirana u 2584 operona), sedam rRNA operona i 86 tRNA gena (Slika 1.). Iako je bio predmet intenzivnih genetičkih istraživanja velika većina tih gena je bila nepoznata. Iznimno je bogat protein kodirajućim genima budući da je prosječna udaljenost gena samo 118 parova baza. Također sadrži značajan broj pokretnih genetičkih elemenata, ponavljajućih elemenata, kriptičkih profaga i ostataka bakteriofaga te ostale dijelove neobičnog sastava koji navode na plastičnost genoma tijekom horizontalnog prijenosa (Blattner et al., 1997).



Slika 1. Sveukupna struktura *Escherichia coli* K-12 genoma. Mjesto početka i kraja replikacije je prikazano zelenim linijama s plavim linijama koje označavaju replichore (eng. replichores) 1 i 2. Skala označava koordinate i u parovima baza i u minutama. Distribucija gena je opisana na dva vanjska prstena. Narančaste oznake su geni locirani na prezentiranom lancu, a žute oznake su geni na suprotnom lancu. Crvene strelice prikazuju lokacije i smjer transkripcije rRNA gena, a tRNA geni su prikazani zelenim strelicama (Blattner et al., 1997).

1.5.2. *Escherichia coli* ISC11

Soj *Escherichia coli* ISC11 je β-laktam otporan klinički izolat. Karakteriziran je pomoću PCR-a i sekvenciranja Sangerovom metodom. Sekvenciran je pomoću IonTorrent PGM platforme. Genom je sastavljen iz neprekinutih sljedova (engl. contigs) pomoću MIRA 3.9.9. assemblera. Program RAST (Rapid Annotation using Subsystem Technology) je bio korišten za anotiranje genoma. Klinički uzorak je sakupljen u Općoj bolnici Beč, Austrija (Barišić et al., 2016).

Genom soja ISC11 sadrži 4,976,867 parova baza, 6,130 protein-kodirajuća gena i 99 nekodirajuća gena (preuzeto s http://bacteria.ensembl.org/Escherichia_coli_isc11/Info/Annotation/#about).

2. CILJ ISTRAŽIVANJA

Genomska filostratigrafija je opsežno primjenjiva metoda koja je omogućila istraživanje obrazaca genomske evolucije uključujući i evoluciju gena povezanih s bolestima, ontogeniju i *de novo* podrijetlo gena. U svim dosadašnjim publiciranim istraživanjima isključivo se primjenjivala na eukariotskim vrstama te se nije koristila na prokariotskim odnosno bakterijskim vrstama. Bakterije su jedni od prvih organizama na Zemlji i njihovi genomi posjeduju mnogo informacija o evolucijskoj prošlosti te će u budućnosti doprinijeti odgovoru na neka fundamentalna pitanja poput podrijetla prve stanice, metabolizma, fotosinteze, procesa prijenosa informacije te ancestralne eukariotske stanice. Tome ide u prilog količina dostupnih sekvenciranih bakterijskih genoma kojih je preko 30,000 prema podacima iz 2014. godine (Land et al., 2015).

Cilj ovog istraživanja je izrada filostratigrafske mape bakterije *Escherichia coli* koja je najproučavaniji i najkorišteniji prokariotski organizam u biologiji te provjera njene robusnosti.

Specifični ciljevi istraživanja:

1. Određivanje konsenzus filogenije bakterije *Escherichia coli* koja će se koristiti u izradi filostratigrafske mape
2. Izrada filostratigrafske mape sojeva *Escherichia coli* K-12 i ISC11
3. Provjera robusnosti dobivenih filostratigrafskih mapa na reduciranoj i nasumičnoj bazi
4. Provjera robusnosti dobivenih filostratigrafskih mapa s obzirom na promjenu broja hitova po filostratumu
5. Interpretacija i analiza uloge u biološkim procesima, molekularne funkcije i stanične lokalizacije proteina koji su identificirani kao evolucijski stari te onih koji su identificirani kao evolucijski mladi pomoću Gene Ontology funkcionalnih podataka

3. MATERIJALI I METODE

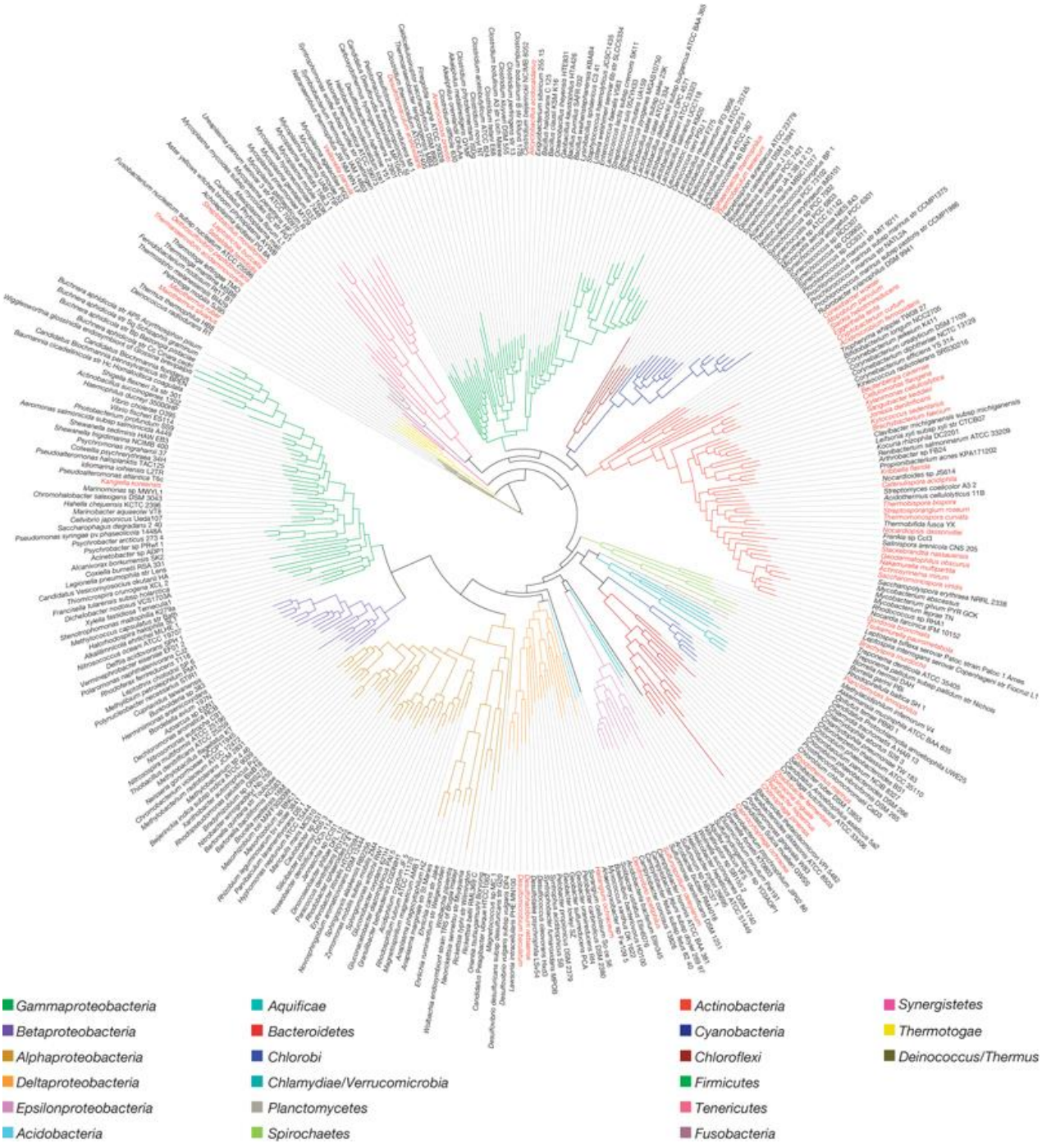
3.1. Bakterijska konsenzus filogenija

Odabir filogenije je od fundamentalne važnosti pri određivanju filogenetskog podrijetla gena (Domazet-Lošo, Brajković and Tautz, 2007). Za izradu konsenzus filogenije *E.coli* na razini koljena proučena je dostupna literatura i prema tome je određena filogenetska prošlost *E.coli*.

Kao ishodišna skupina bakterija postavljena su koljena Aquificae i Thermotogae. Na temelju 16S rRNA bakterijske filogenije koja se smatra „zlatnim standardom“ i najpouzdanijim mjerilom za ovako evolucijski udaljene skupine, ova dva koljena su određena kao najstarija. Daljnjim različitim istraživanjima je također višestruko puta dokazana teorija o hipertermofilnom pretku bakterija dok za suprotne teorije nema dovoljno dokaza ili su opovrgnute (Di Gulio, 2003; Barion et al., 2007; Boussau et al., 2008).

Između kraljevstva bakterija i koljena Proteobacteria u koje pripada bakterija *Escherichia coli* dodana su tri nova čvora koja se zovu Grupa A, B i C. U Grupu A je postavljeno koljeno Deinococcus-Thermus, u Grupu B je postavljeno koljeno Synergistetes, a u Grupu C preostala bakterijska koljena. Koljena su podijeljena prema filogenetskom stablu (Slika 2.) koje je dobiveno u sklopu projekta „Genomic Encyclopaedia of Bacteria and Archaea“. Radi se o takozvanom genomskom stablu koje je dobiveno na temelju višestrukog sravnjenja 31 proteinskog slijeda.

Unutar koljena Proteobacteria zadržani su zadani čvorovi i imena s obzirom na potvrdu filogenetske povezanosti skupina unutar pojedinih linija u nizu literaturnih navoda (Paradis et al., 2005; Hong Nhung et al., 2007; Williams et al., 2010). Filogenija je vizualizirana u programu CorelDRAW X6 i prikazana u poglavlju Rezultati.



Slika 2. Maximum-likelihood filogenetsko stablo bakterijskih koljena na temelju sravnjenja 31 visoko očuvanog protein kodirajućeg gena (Wu and Eisen, 2008). Bakterijska koljena označena su različitim bojama grana i GEBA genomi su označeni crvenom bojom u vanjskom krugu s imenima vrsta (Wu et al*., 2009*)

3.2. Prikupljanje podataka

3.2.1. Proteomi *E. coli*

U radu su korištena dva različita proteoma *E.coli* jedan soja K-12, a drugi soja ISC11. Oba proteoma su preuzeta iz baze proteoma unutar baze UniProtKB (http://www.uniprot.org/proteomes/) u FASTA formatu. U njoj su pohranjeni kao referentni proteomi što znači da su ručno i pomoću algoritma odabrani prema nizu kriterija od mnogo ostalih proteoma kao oni koji predstavljaju presjek taksonomske različitosti koja može biti pronađena unutar baze.

Proteom soja K-12 (http://www.uniprot.org/proteomes/UP000000625) spremljen je u bazi pod Proteome ID-om UP000000625, a proteom soja ISC11 (http://www.uniprot.org/proteomes/UP000019194) pod Proteome ID-om UP000019194.

3.2.2. Proteomske baze

Korištene su tri različite baze proteoma: sveobuhvatna, reducirana i nasumična. Sveobuhvatna baza sadrži eukariotske proteome preuzete s poslužitelja Ensembl i bakterijske proteome preuzete s poslužitelja Ensembl, JGI i Broad (Tablica 1. i Tablica 4.). Od sveobuhvatne baze su kreirane reducirana i nasumična proteomska baza. Za reduciranu bazu odabrano je 10 proteoma po filostratumu na principu što šire filogenetske pokrivenosti bočnih ogranaka (Tablica 2. i Tablica 5.), a za nasumičnu bazu odabrano je 10 proteoma po filostratumu slučajnim uzorkovanjem u Excelu (Tablica 3. i Tablica 6). Proteomi iz 12. filostratuma reducirane i nasumične baze oba soja su isti kao oni u sveobuhvatnoj bazi budući da ih je manje od 10 pa se nije moglo provoditi uzorkovanje. Odabrani proteomi u FASTA formatu su uzeti iz sveobuhvatne baze te pomoću alata iz paketa BLAST+ kreirane su formatirane proteinske baze za BLAST.

Tablica 1. Sažetak sveobuhvatne BLAST baze koja je korištena za izradu filostratigrafske mape *E.coli* K-12

Redni broj filostratuma	Tax ID filostratuma	Broj proteoma	Broj proteina	Naziv filostratuma
1	131567	1814	25030270	root : cellular_organisms
2	2	77	193249	cellular_organisms : Bacteria
3	1708685	47	142569	Bacteria : GroupA
4	1708686	17	45891	GroupA : GroupB
5	1708687	13369	44895661	GroupB : GroupC
6	1224	3093	11761731	GroupC : Proteobacteria
7	1236	2719	11602167	Proteobacteria : Gammaproteobacteria
9	543	2079	9893420	Enterobacteriales : Enterobacteriaceae
10	561	19	79099	Enterobacteriaceae : Escherichia
11	562	1981	10134377	Escherichia : Escherichia_coli
12	1432557	8	32292	Escherichia_coli : Escherichia_coli_K-12

Tablica 2. Sažetak reducirane BLAST baze koja je korištena za izradu filostratigrafske mape *E.coli* K-12

Redni broj filostratuma	Tax ID filostratuma	Broj proteoma	Broj proteina	Naziv filostratuma
1	131567	10	172680	root : cellular_organisms
2	2	10	31318	cellular_organisms : Bacteria
3	1708685	10	30533	Bacteria : GroupA
4	1708686	10	19807	GroupA : GroupB
5	1708687	10	40813	GroupB : GroupC
6	1224	10	43927	GroupC : Proteobacteria
7	1236	10	28645	Proteobacteria : Gammaproteobacteria
9	543	10	54774	Enterobacteriales : Enterobacteriaceae
10	561	10	44153	Enterobacteriaceae : Escherichia
11	562	10	50110	Escherichia : Escherichia_coli
12	83333	8	32292	Escherichia_coli : Escherichia_coli_K12

Tablica 3. Sažetak nasumične BLAST baze koja je korištena za izradu filostratigrafske mape *E.coli* K-12

Redni broj filostratuma	Tax ID filostratuma	Broj proteoma	Broj proteina	Naziv filostratuma
1	131567	10	153794	root : cellular_organisms
2	2	10	18069	cellular_organisms : Bacteria
3	1708685	10	28975	Bacteria : GroupA
4	1708686	10	21188	GroupA : GroupB
5	1708687	10	42956	GroupB : GroupC
6	1224	10	26418	GroupC : Proteobacteria
7	1236	10	56974	Proteobacteria : Gammaproteobacteria
9	543	10	48671	Enterobacteriales : Enterobacteriaceae
10	561	10	41034	Enterobacteriaceae : Escherichia
11	562	10	51644	Escherichia : Escherichia_coli
12	83333	8	32292	Escherichia_coli : Escherichia_coli_K12

Tablica 4. Sažetak sveobuhvatne BLAST baze koja je korištena za izradu filostratigrafske mape *E.coli* ISC11

Redni broj filostratuma	Tax ID filostratuma	Broj proteoma	Broj proteina	Naziv filostratuma
1	131567	1814	25042560	root : cellular_organisms
2	2	77	193894	cellular_organisms : Bacteria
3	1708685	47	144659	Bacteria : GroupA
4	1708686	17	34980	GroupA : GroupB
5	1708687	13369	44898601	GroupB : GroupC
6	1224	3093	11762555	GroupC : Proteobacteria
7	1236	2719	11605975	Proteobacteria : Gammaproteobacteria
9	543	2079	9897789	Enterobacteriales : Enterobacteriaceae
10	561	19	81616	Enterobacteriaceae : Escherichia
11	562	1988	10165582	Escherichia : Escherichia_coli
12	1432557	1	6140	Escherichia_coli : Escherichia_coli_ISC11

Tablica 5. Sažetak reducirane BLAST baze koja je korištena za izradu filostratigrafske mape *E.coli* ISC11

Redni broj filostratuma	Tax ID filostratuma	Broj proteoma	Broj proteina	Naziv filostratuma
1	131567	10	172680	root : cellular_organisms
2	2	10	31318	cellular_organisms : Bacteria
3	1708685	10	30533	Bacteria : GroupA
4	1708686	10	19807	GroupA : GroupB
5	1708687	10	40813	GroupB : GroupC
6	1224	10	43927	GroupC : Proteobacteria
7	1236	10	28645	Proteobacteria : Gammaproteobacteria
9	543	10	54774	Enterobacteriales : Enterobacteriaceae
10	561	10	44153	Enterobacteriaceae : Escherichia
11	562	10	50110	Escherichia : Escherichia_coli
12	1432557	1	6140	Escherichia_coli : Escherichia_coli_ISC11

Tablica 6. Sažetak nasumične BLAST baze koja je korištena za izradu filostratigrafske mape *E.coli* ISC11

Redni broj filostratuma	Tax ID filostratuma	Broj proteoma	Broj proteina	Naziv filostratuma
1	131567	10	153794	root : cellular_organisms
2	2	10	18069	cellular_organisms : Bacteria
3	1708685	10	28975	Bacteria : GroupA
4	1708686	10	21188	GroupA : GroupB
5	1708687	10	42956	GroupB : GroupC
6	1224	10	26418	GroupC : Proteobacteria
7	1236	10	56974	Proteobacteria : Gammaproteobacteria
9	543	10	48671	Enterobacteriales : Enterobacteriaceae
10	561	10	41034	Enterobacteriaceae : Escherichia
11	562	10	51644	Escherichia : Escherichia_coli
12	1432557	1	6140	Escherichia_coli : Escherichia_coli_ISC11

3.3. Priprema skripti za pokretanje programa BLAST

Skripte za pokretanje programa BLAST su pripremljene pomoću programa PsiBlastHelper (preuzeto s https://github.com/msestak/PsiBlastHelper). PsiBlastHelper je modulino koji cijepa FASTA input na manje dijelove za visoko protočni (engl. high throughput) BLAST, PSI-BLAST i HMMER te kreira SGE i HTCondor skripte koje pokreću te poslove na klasteru.

Kao input su potrebni proteom u FASTA formatu i lokacija baze i zadati parametre: naziv podskupova (engl. chunks) proteoma, broj proteinskih sljedova po podskupu, broj BLAST poslova u svakom skriptu i lokaciju gdje se nalazi BLAST. U ovom specifičnom slučaju zadano je da bude 50 proteinskih sljedova u svakom podskupu i 6 BLAST poslova u svakom skriptu. Također se zadaje granični broj aminokiselina prema kojem se proteinski sljedovi koji imaju više aminokiselina od tog graničnog broja tretiraju kao iznimno veliki i odvojeno se procesiraju jedan po jedan zato jer BLAST radi vrlo sporo s dugim sljedovima. Kao granični broj za tu opciju određeno je 3000 aminokiselina iako su svi proteinski sljedovi bili kraći od te duljine. Svaki skript radi više BLAST poslova zato da se smanji broj kopija BLAST baze koja može voditi do propalih poslova ako nestane mjesta na serveru tako da jedan skript predstavlja jednu kopiju bazu ,ali više BLAST poslova.

Pripremljene su skripte za pokretanje programa BLAST za proteome *E.coli* sojeva K-12 i ISC11 za pretraživanje sveobuhvatne, reducirane i nasumične baze.

3.4. Pretraživanje i sravnjenje upotrebom programa BLAST

Kako bi proteinske sljedove iz određenog proteoma podijelili po međučvorovima filogenetskog stabla korišten je alat BLAST (Basic Local Alignment Search Tool). BLAST je algoritam za uspoređivanje primarnih bioloških sljedova poput aminokiselinskih ili nukleotidnih sljedova. Pomoću njega uspoređuje se željeni slijed s bazom sljedova i identificiraju se oni sljedovi u bazi koji ima najveću sličnost sa željenim slijedom prema određenim kriterijima.

Program je napisan tako da se uspoređuju proteinski ili DNA sljedovi s proteinskom ili DNA bazom u bilo kojoj kombinaciji s time da DNA sljedovi često prolaze kroz konceptualnu translaciju prije nego su usporedbe provedene.

Radi se se o heurističkom algoritmu koji u prvom koraku slijed s kojim pretražuje bazu cijepa na tzv. riječi. Zadana veličina riječi je 3 aminokiseline za proteinske sljedove, a za nukleotidne sljedove to je 11 nukleotida. Ako pretražujemo bazu na primjer sa proteinskim slijedom VPDSI dobit ćemo riječi VPD, PDS, DSI. Nakon toga svaka riječ se boduje prema supstitucijskoj matrici. Najčešće korištena je supstitucijska matrica BLOSUM62 i odabir matrice ovisi o tome što želimo pronaći. Od dobivene liste riječi zadržavaju se samo one riječi koje imaju broj bodova iznad zadanog praga. Te riječi se potom koriste za pretraživanje baze i uspoređuju se sa sljedovima baze u potrazi za identičnim nizom aminokiselina ili nukleotida. Nakon sparivanja identičnih riječi provodi se sravnjenje između pretražujućeg slijeda i onoga iz baze u oba smjera počevši od mjesta gdje je pronađena riječ. Produljenje ne prestaje sve dok ukupan broj bodova ne počinje opadati. Tako produljene inicijalne riječi se nazivaju visoko bodovni segmentski parovi (engl. high-scoring segment pair). Od tih parova zadržavaju se samo oni čiji broj bodova je veći od empirijski određene bodovne granice S te se statistički određuje značaj svakog rezultata koji je preostao. Statistička značajnost svakog para se ispituje eksplotacijom distribucije Gumbelove ekstremne vrijednosti. Prema toj distribuciji, vjerojatnost da će određeni broj bodova S biti jednak ili veći od x je određen jednadžbom:

$$p(S \geq x) = 1-\exp(-e^{-\lambda(x-\mu)})$$

u kojoj je $\mu = [\log(K m' n')] / \lambda$. Statistički parametri λ i K procjenjuju se namještanjem distribucije bodova lokalnih sravnjenja bez praznina slijeda kojim smo pretraživali bazu i mnogih izmješanih verzija slijeda iz baze na distribuciju Gumbelove ekstremne vrijednosti. Oba parametra ovise o supstitucijskoj matrici, bodovanju praznina i sastavu slijeda. Parametri *m'* i *n'* označavaju efektivnu duljinu slijeda kojim smo pretraživali bazu i duljinu svih sljedova u bazi. Stvarna duljina je skraćena na efektivnu duljinu kako bi se kompenzirao efekt rubova (engl. edge effect) (sravnjenje koje počinje blizu kraja slijeda vjerojatno neće imati dovoljno slijeda da se napravi optimalno sravnjenje. Parametri *m'* i *n'* računaju se prema jednadžbama:

$$m' \approx m - (\ln Kmn)/H$$

$$n' \approx n - (\ln Kmn)/H$$

parametar H predstavlja prosječni očekivani broj bodova po paru sravnjenih aminokiselina u sravnjenju dvaju nasumičnih sljedova.
Očekivana vrijednost E (engl. expect score) dobivenog rezultata predstavlja broj puta koliko bi neki nepovezani slijed iz baze dobio broj bodova S veći od x čistom slučajnošću. Vrijednost E dobivena pretraživanjem baze veličine D je određena kao:

$$E \approx 1 - e^{-p(s>x)D}$$

Ova vrijednost koja procjenjuje značaj vrijednosti visoko bodovnih segmentskih parova za lokalno sravnjenje bez praznina koje se nalazi u BLAST rezultatima. Kada se dva ili više visoko bodovna segmentska para nalaze u istom slijedu iz baze onda mogu činiti veće sravnjenje. Jednadžbe prikazane ovdje se modificiraju ako se pojedini visoko bodovni parovi kombiniraju prilikom stvaranja sravnjenja s prazninama zbog varijacije u statističkim parametrima i uključuje korištenje određenih penala za praznine (Mount, 2004).

U sklopu ovo rada napravljene su BLAST pretrage proteoma *E.coli* soja K-12 i soja ISC11 s kompletnom, reduciranom i nasumičnom proteinskom bazom. Koristila se BLAST verzija 2.6.0+. Zadan je E-value 0.001 i supstitucijska matrica BLOSUM62.

3.5. Analiza BLAST rezultata

BLAST rezultati dobiveni pretraživanjem proteinskih sljedova iz proteoma *E.coli* sojeva K-12 i ISC11 sa sveobuhvatnom bazom su analizirani pomoću programa BlastoutAnalyze (preuzeto s https://github.com/msestak/BlastoutAnalyze).

To je modulino koji analizira BLAST rezultate i bazu. Od više mogućih opcija koje nudi ovaj program korištena je redukcija BLAST rezultata s obzirom na broj hitova u filostratumu. Spomenuta opcija uklanja pojedine rezultate iz sveukupnih BLAST rezultata ako je broj različitih organizama iz baze s kojim imaju rezultat unutar jednog filostratuma manji ili jednak zadanoj granici. To znači ako neki proteinski slijed ima npr. tri rezultata u trećem filostratumu, a zadana granica je 3 onda će taj rezultat biti uklonjen dok će oni koji imaju 4 ili više rezultata biti zadržani. Za ovu opcije su potrebni BLAST rezultati i datoteka analize (engl. analyze file). U datoteci se nalazi distribucija organizama po filostratumima koja je potrebna kako bi se znalo iz kojeg filostratuma je koji organizam. Redukcija BLAST rezultata s obzirom na broj hitova u filostratumu je napravljena s graničnim vrijednostima 2,3,5,7,9,10 i 15.

3.6. Izrada filostratigrafskih mapa

Filostratigrafske mape izrađene su pomoću paketa PhyloToolKit (Bakarić, neobjavljeno). Ovaj paket sadrži kolekciju C++ alata skupa s pripadajućim Perl programima dizajniranim za brzu i efikasnu računalnu analizu dobitka i gubitka gena te genskih obitelji.

Prvi korišteni program je bio PhyloStrat. PhyloStrat je implementacija genomske stratifikacije temeljene na filogeniji koju su predložili autori Tomislav Domazeta-Lošo, Josip Brajković i Diethard Tautz 2007. godine. Na temelju BLAST rezultata primjenjujući Dollovu parsimoniju određuje se obrazac prisutnosti odnosno odsutnosti gena u filostratumu određene vrste. Program određuje točku podrijetla za svaki gen koji je korišten u analizi. Program kao input treba dvije datoteke i taksonomski broj (engl. Taxonomy ID) vrste koja se analizira. Prva datoteka su BLAST rezultati u obliku tab formatirane tablice, a druga datoteka su čvorovi filogenetskog stabla vrste čija mapa se izrađuje. Ta datoteka je lista taksonomskih brojeva podijeljenih u dva stupca, u lijevom stupcu je dječji čvor (engl. child node) dok se u desnom stupcu nalazi pripadajući roditeljski čvor (parent node). Rezultat ovog programa je podijeljen u tri dijela. Prvi dio Taxonomy id sadrži taksonomski broj vrste koja je korištena u analizi, drugio dio Phylogeny used prikazuje informacije o čvorovima korištenim u stratifikacijskoj analizi, posljednji treći dio Phylostratigraphy map prikazuje rezultate stratifikacije.

Drugi korišteni program je bio AddNames. Ovaj program daje ime svakom filostratumu. Program traži dvije input datoteke. Prva potrebna datoteka je output programa PhyloStrat i tablica s imenima filostratuma. Rezultat ovog programa je isti kao programa PhyloStrat samo što svaki filostrat ovdje još ima ime.

Treći korišteni program je bio MapSummary. Pomoću njega se se dobije sumarna statistika podataka koji se nalaze u filostratigrafskoj mapi.Kao input koristi rezultat koji je dobiven programom AddNames. Rezultat ovog programa je tablica s pet stupaca. S lijeva na desno stupci sadrže: redni broj filostratuma, taksonomski broj filostratuma, ime filostratuma, broj gena unutar filostratuma i postotak gena unutar filostratuma.

U sklopu ovog rada napravljene su sveukupno 22 filostratigrafske mape bakterije *E.coli.* Za oba soja su napravljene mape na temelju BLAST rezultata

dobivenih koristeći kompletnu, reduciranu i nasumičnu bazu. BLAST rezultati dobiveni korištenjem kompletne baze koji su potom analizirani s različitim graničnim vrijednostima su isto stratificirani.

3.7. Analiza rezultata iz pojedinih filostratuma prema rječniku Gene Ontology (GO)

Gene Ontology projekt omogućio je stvaranje seta hijerarhijski organiziranog vokabulara koji opisuje gene i genske produkte. Konzorcij GO uključuje mnoge baze podataka koje su usvojile anotaciju temeljenu na kontroliranom vokabularu, a među njima je i baza podataka o proteomu *E.coli* soja K-12 (Kaminuma et al., 2008).

Korištenjem GO Annotations alata na mrežnim stranicama Gene Ontology Consortium-a (http://www.geneontology.org/) provedena je pretraga anotacije za proteine iz prvog filostratuma i mlađih filostratuma filostratigrafskih mapa soja K-12 koje su dobivene na sveobuhvatnoj, reducirano i nasumičnoj bazi te nakon podizanja broja pogodaka po filostratumu na vrijednost 15. Za pretraživanje baze korištene su oznake proteina iz baze UniProtKB. Analizirane su molekularna funkcija, stanična lokalizacija i uloga u biološkim procesima.

4. REZULTATI

4.1. Filostratigrafska mapa soja *Escherichia coli* K-12

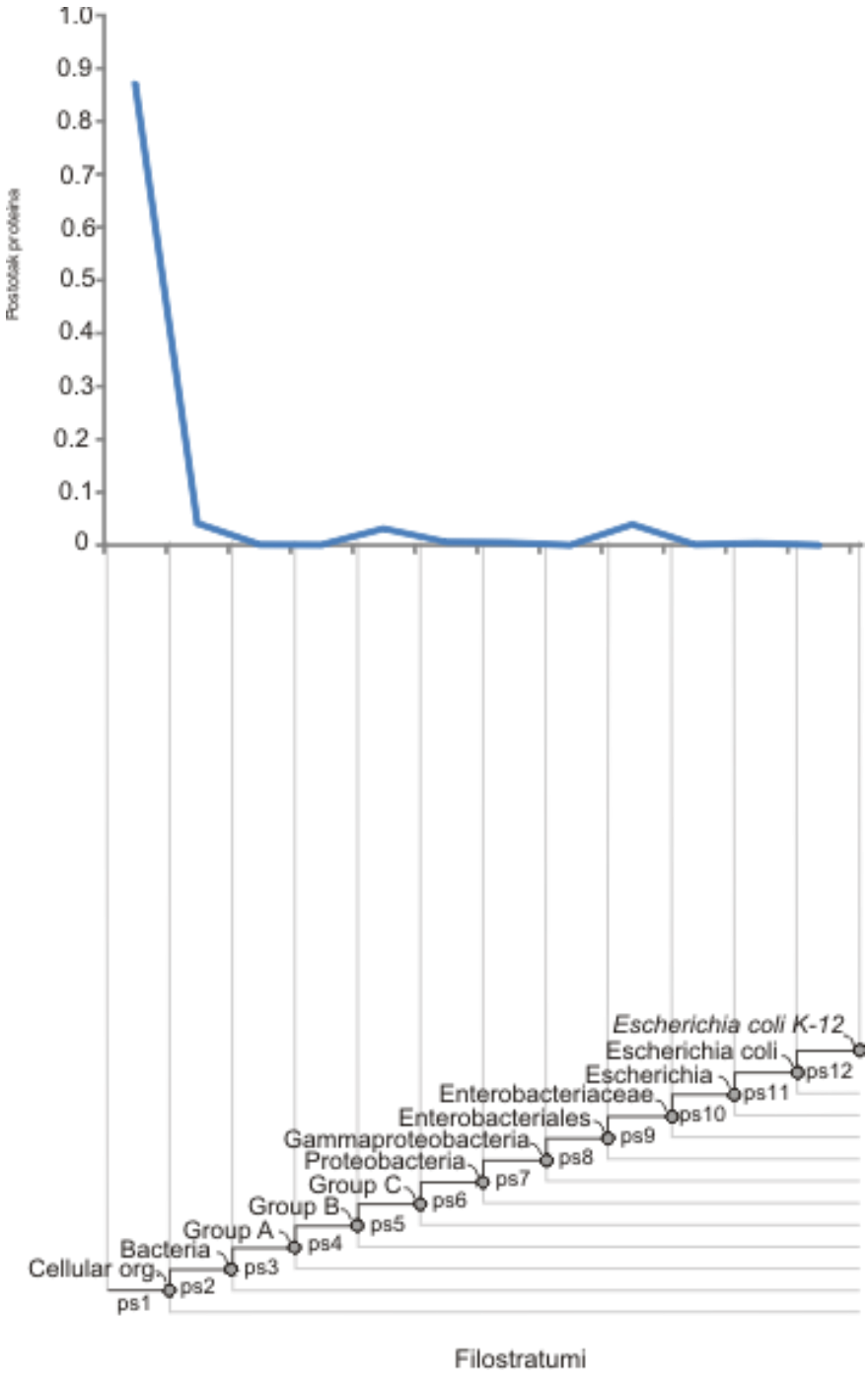
Filogenija bakterija izrađena na temelju novije literature prikazana je u sklopu Slika 1. i 2. Važno je napomenuti kako je teorijski određeno 12 filostratuma no međutim u dva filostratuma 8. i 9. se nalaze isti organizmi. U 8. filostratumu se nalazi red Enterobacteriales, a u 9. filostratumu porodica Enterobacteriaceae koja je jedina porodica tog reda tako da red Enterobacteriales čine samo organizmi iz porodice Enterobacteriaceae. Zbog toga u svim daljnjim rezultatima nema 8. filostratuma. Analizu programom BLAST nije uspjelo proći 0.627% proteinskih sljedova *E.coli* K-12. Ostatak proteinskih sljedova, njih 4279, raspoređeni su u 11 filostratuma pomoću metode genomske filostratigrafije (Slika 3.). Nazivi filostratuma i pridružene bočne skupine organizama koje se u navedenom filostratumu odvajaju od evolucijske linije *E.coli* K-12 prikazane su u Tablici 7. Daleko najviše proteinskih sljedova nalazi se u prvom filostratumu što se može djelomično opravdati time jer je taj evolucijski period bio iznimno dug i time što prvi filostratum ima puno više predstavnika od ostalih filostratuma u sveobuhvatnoj proteinskoj bazi koja je bila korištena. Tri manja maksimuma mogu se primjetiti u 2., 5. i 9. filostratumu dok su signali u ostalim filostratumima minimalni. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima prikazan je u Tablici 8.

Tablica 7. U tablici su pobrojeni i opisani filostratumi te pridružene skupine organizama koje se u navedenom filostratumu odvajaju od evolucijske linije soja *E.coli* K-12, tzv. bočne skupine. Filostratumi predstavljaju razine filogenetskog stabla soja do kojih možemo pratiti podrijetlo nekog slijeda metodom filostratigrafije.

Broj filostratuma	Naziv filostratuma	Bočne skupine
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama
Ps2	Staničan organizam - Bacteria	Eukaryota, Archaea
Ps3	Bacteria – Grupa A	Aquificeae, Thermotogae
Ps4	Grupa A – Grupa B	Deinococcus-Thermus
Ps5	Grupa B – Grupa C	Synergistetes
Ps6	Grupa C - Proteobacteria	Chlamydiae/Verrucomicrobia, Planctomycetes, Bacteroidetes/Chlorobi group, Spirochaetes, Firmicutes Fibrobacteres/Acidobacteria, Nitrospinae Deferribacteres, Nitrospirae, Elusimicrobia, Tenericutes
Ps7	Proteobacteria - Gammaproteobacteria	Proteobacteria
Ps8	Gammaproteobacteria - Enterobacteriales	Gammaproteobacteria
Ps9	Enterobacteriales - Enterobacteriaceae	Ostali Enterobacteriales
Ps10	Enterobacteriaceae - Escherichia	Ostali Enterobacteriaceae
Ps11	Escherichia – Escherichia coli	Ostali Escherichia
Ps12	Escherichia coli – Escherichia coli K 12	Ostali Escherichia coli

Tablica 8. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon filostratigrafije sa sveobuhvatnom proteinskom bazom eukariota, arheja i bakterija.

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3723	87,01%
Ps2	Staničan organizam - Bacteria	176	4,11%
Ps3	Bacteria – Grupa A	5	0,12%
Ps4	Grupa A – Grupa B	4	0,09%
Ps5	Grupa B – Grupa C	134	3,13%
Ps6	Grupa C - Proteobacteria	26	0,61%
Ps7	Proteobacteria - Gammaproteobacteria	21	0,49%
Ps9	Enterobacteriales - Enterobacteriaceae	168	3,93%
Ps10	Enterobacteriaceae - Escherichia	5	0,12%
Ps11	Escherichia – Escherichia coli	17	0,40%
Ps12	Escherichia coli – Escherichia coli K 12	0	0,0%



Slika 3. Filostratigrafska mapa *E.coli* soja K-12 rađena na bazi eukariota, arheja i bakterija i pripadajuća konsenzus filogenija. Mapa prikazuje preraspodjelu proteinskih sljedova soja *E.coli* K-12 po pojedinim filostratumima. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* K-12.

4.2. Filostratigrafska mapa soja *Escherichia coli* ISC11

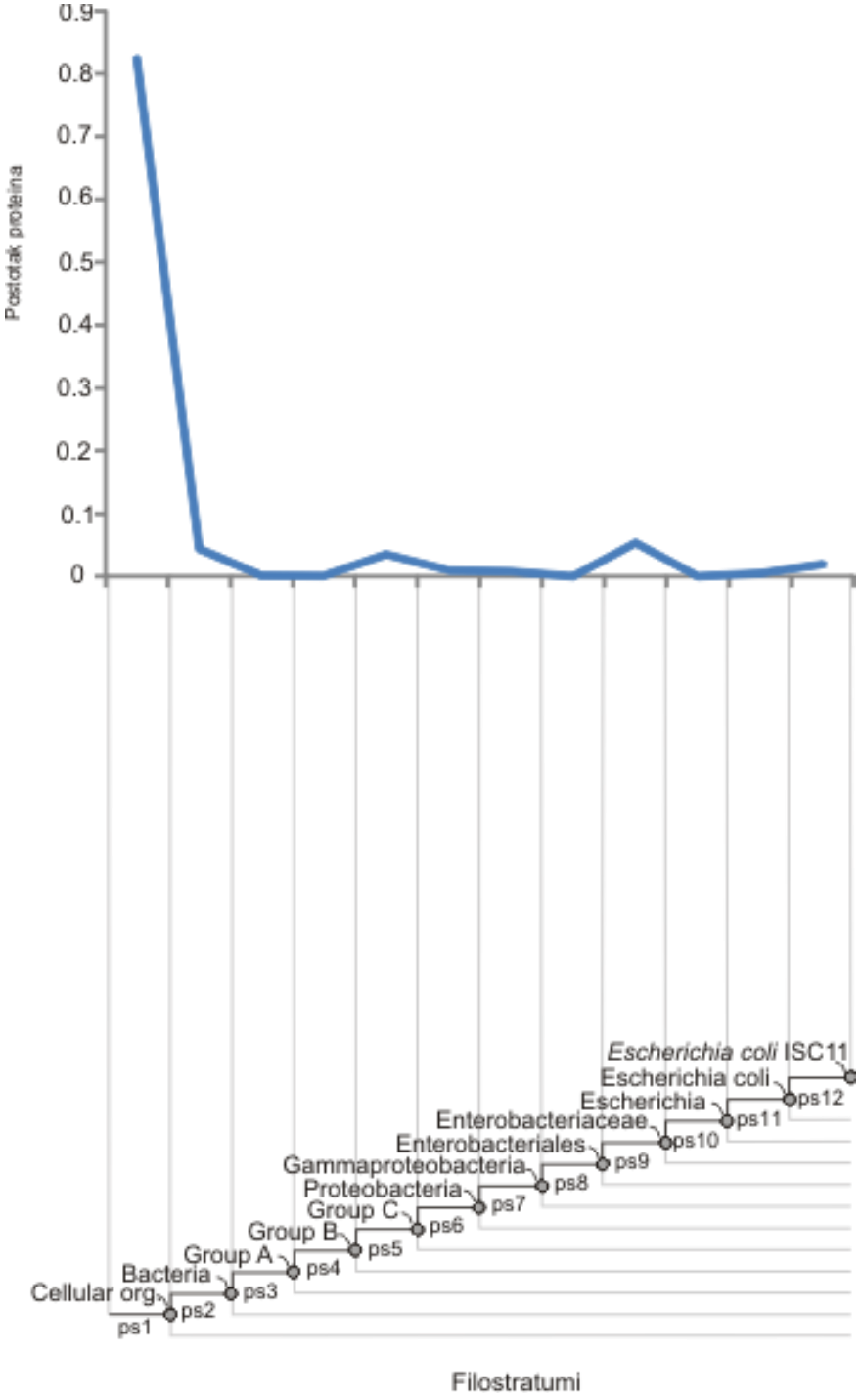
Analizu programom BLAST nije uspjelo proći 2.34% proteinskih sljedova *E.coli* ISC11. Ostatkak proteinskih sljedova, njih 5986, preraspoređeni su u 11 filostratuma pomoću metode genomske filostratigrafije (Slika 4.). Nazivi filostratuma i pridružene bočne skupine organizama koje se u navedenom filostratumu odvajaju od evolucijske linije *E.coli* ISC11 prikazane su u Tablici 9. Najviše proteinskih sljedova se nalazi u prvom filostratumu iz istih razloga koji su navedeni i za soj K-12. Ovdje je postotak proteinskih sljedova u prvom filostratumu nešto manji, ali je i dalje jako visok. Manji maksimumi mogu se primjetiti u 2.,5.,9. i 12. filostratumu dok u ostalim filostratumima signali su vrlo mali. Proteom soja ISC11 sadrži 1824 proteinskih sljedova više od soja K-12 i taj soj je klinički izolat dok je soj K-12 laboratorijski soj međutim distribucija proteinskih sljedova u filostratigrafskoj mapi je ista.

Tablica 9. U tablici su pobrojeni i opisani filostratumi te pridružene skupine organizama koje se u navedenom filostratumu odvajaju od evolucijske linije soja *E.coli* ISC11, tzv. bočne skupine. Filostratumi predstavljaju razine filogenetskog stabla soja do kojih možemo pratiti podrijetlo nekog slijeda metodom filostratigrafije.

Broj filostratuma	Naziv filostratuma	Bočne skupine
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama
Ps2	Staničan organizam - Bacteria	Eukaryota, Archaea
Ps3	Bacteria – Grupa A	Aquificeae, Thermotogae
Ps4	Grupa A – Grupa B	Deinococcus-Thermus
Ps5	Grupa B – Grupa C	Synergistetes
Ps6	Grupa C - Proteobacteria	Chlamydiae/Verrucomicrobia, Planctomycetes, Bacteroidetes/Chlorobi group, Spirochaetes, Firmicutes Fibrobacteres/Acidobacteria, Nitrospinae Deferribacteres, Nitrospirae, Elusimicrobia, Tenericutes
Ps7	Proteobacteria - Gammaproteobacteria	Proteobacteria
Ps8	Gammaproteobacteria - Enterobacteriales	Gammaproteobacteria
Ps9	Enterobacteriales - Enterobacteriaceae	Ostali Enterobacteriales
Ps10	Enterobacteriaceae - Escherichia	Ostali Enterobacteriaceae
Ps11	Escherichia – Escherichia coli	Ostali Escherichia
Ps12	Escherichia coli – Escherichia coli ISC11	Ostali Escherichia coli

Tablica 10. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon filostratigrafije sa sveobuhvatnom proteinskom bazom eukariota, arheja i bakterija.

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	4927	82,31%
Ps2	Staničan organizam - Bacteria	262	4,38%
Ps3	Bacteria – Grupa A	8	0,13%
Ps4	Grupa A – Grupa B	6	0,10%
Ps5	Grupa B – Grupa C	210	3,51%
Ps6	Grupa C - Proteobacteria	57	0,95%
Ps7	Proteobacteria - Gammaproteobacteria	49	0,82%
Ps9	Enterobacteriales - Enterobacteriaceae	322	5,38%
Ps10	Enterobacteriaceae - Escherichia	1	0,02%
Ps11	Escherichia – Escherichia coli	29	0,48%
Ps12	Escherichia coli – Escherichia coli ISC11	115	1,92%



Slika 4. Filostratigrafska mapa *E.coli* soja ISC11 rađena na bazi eukariota, arheja i bakterija i pripadajuća konsenzus filogenija. Mapa prikazuje preraspodjelu proteinskih sljedova soja *E.coli* ISC11 po pojedinim filostratumima. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* ISC11.

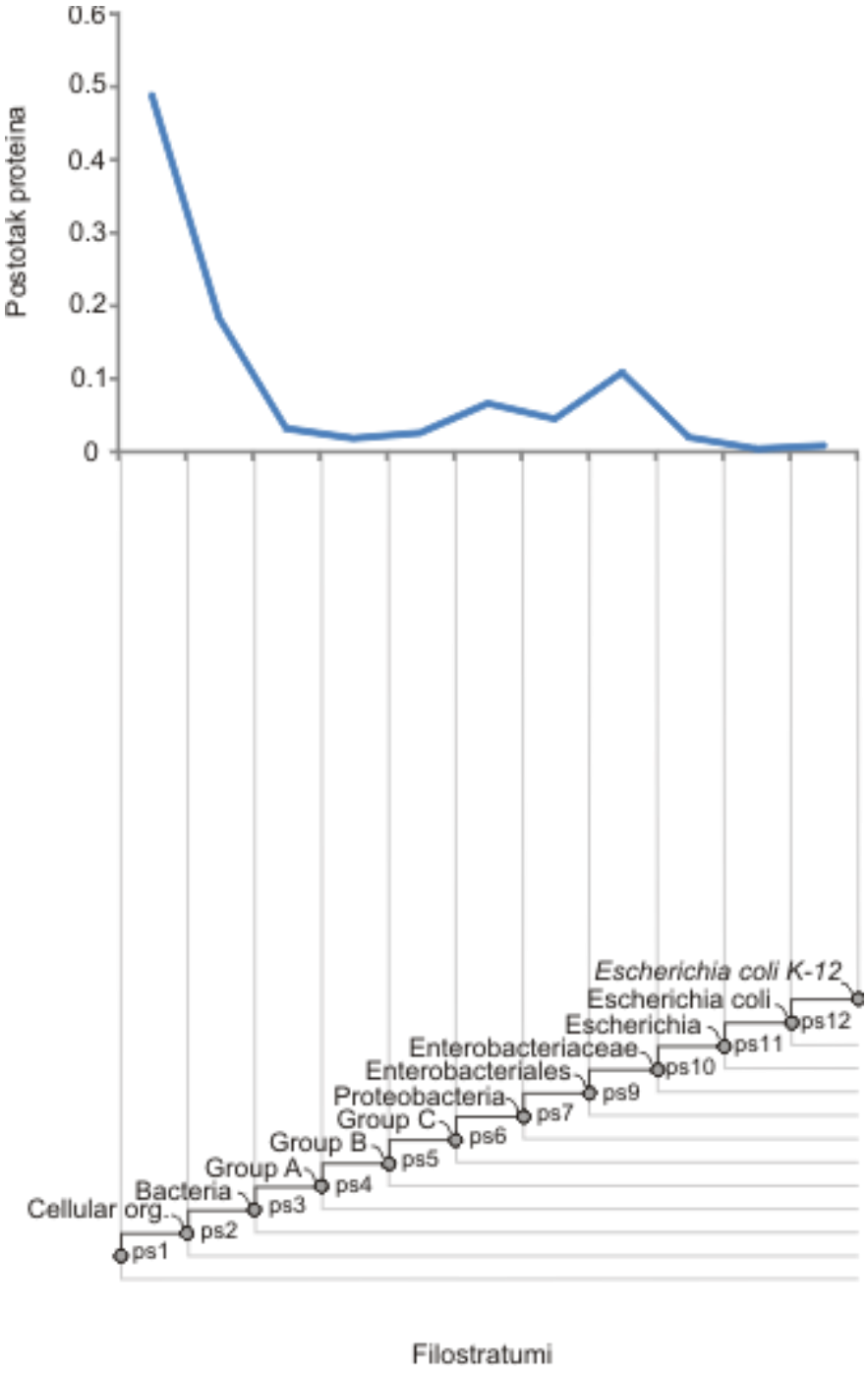
4.3. Provjera robusnosti filostratigrafije na reduciranoj i nasumičnoj bazi

4.3.1. Provjera robusnosti filostratigrafije soja *E.coli* K-12 na reduciranoj bazi

Filostratigrafska mapa *E.coli* soja K-12 izrađena nakon pretraživanja programom BLAST s reduciranom bazom prikazana je na Slici 5. Analizu programom BLAST nije uspjelo proći 0.55% proteinskih sljedova *E.coli* soja K-12. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima prikazan je u Tablici 11. Na temelju usporedbe s mapom koja je dobivena korištenjem sveobuhvatne baze (Slika 5.) može se uočiti da je postotak proteinskih sljedova u prvom filostratumu puno manji te da su se proteinski sljedovi razmjestili u ostale filostratume, a najviše u 2.,6,7. i 9. Ovakvi rezultati su očekivani budući da sada svi filostratumi imaju jednak broj predstavnika.

Tablica 11. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon filostratigrafije s reduciranom proteinskom bazom eukariota, arheja i bakterija.

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2090	48,81%
Ps2	Staničan organizam - Bacteria	783	18,29%
Ps3	Bacteria – Grupa A	137	3,20%
Ps4	Grupa A – Grupa B	80	1,87%
Ps5	Grupa B – Grupa C	112	2,62%
Ps6	Grupa C - Proteobacteria	283	6,61%
Ps7	Proteobacteria - Gammaproteobacteria	193	4,51%
Ps9	Enterobacteriales - Enterobacteriaceae	465	10,86%
Ps10	Enterobacteriaceae - Escherichia	84	1,96%
Ps11	Escherichia – Escherichia coli	18	0,42%
Ps12	Escherichia coli – Escherichia coli K 12	37	0,86%



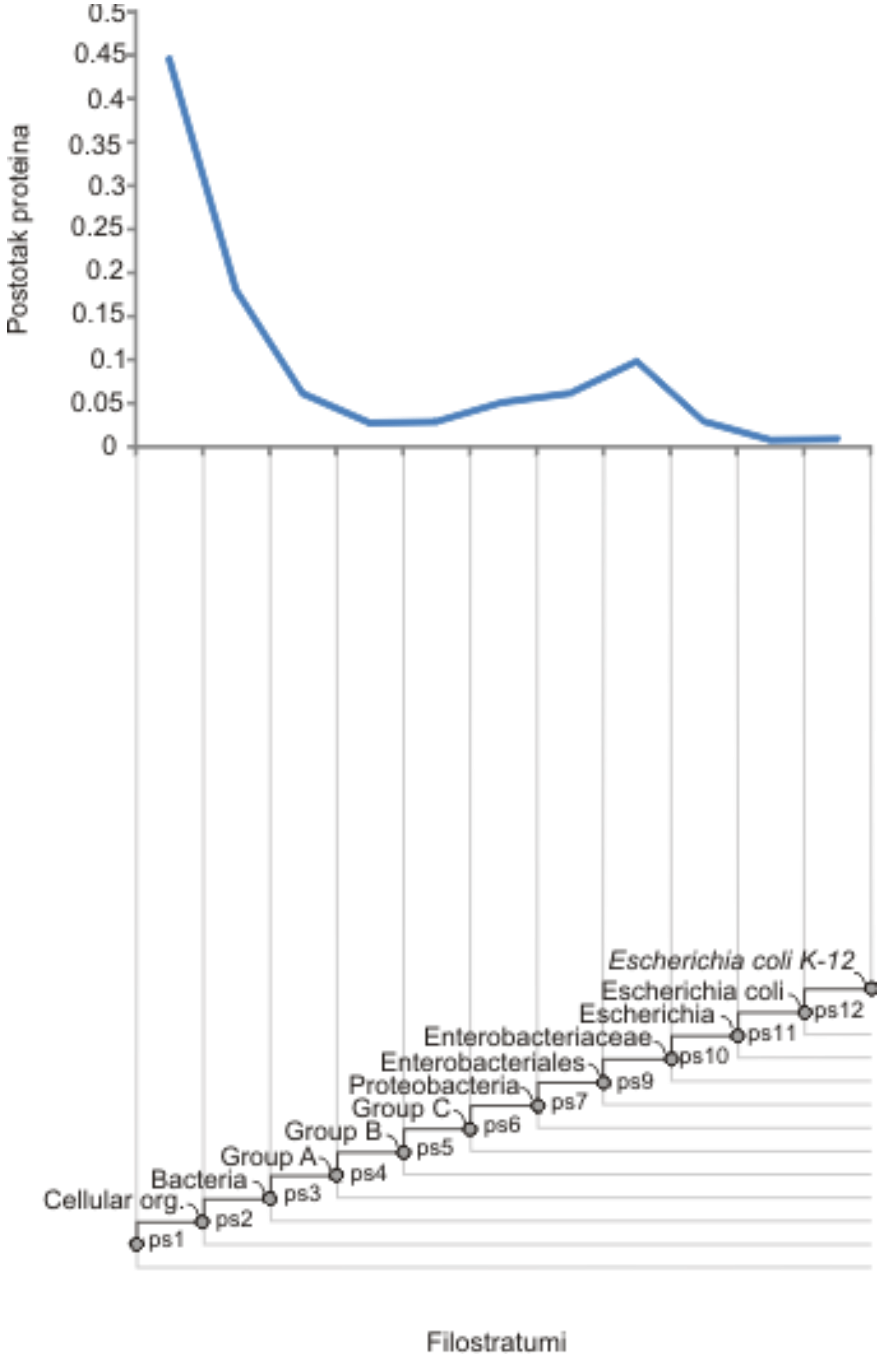
Slika 5. Filostratigrafska mapa *E.coli* soja K-12 rađena na reduciranoj bazi eukariota, arheja i bakterija i pripadajuća konsenzus filogenija. Mapa prikazuje preraspodjelu proteinskih sljedova soja *E.coli* K-12 po pojedinim filostratumima. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* K-12.

4.3.2. Provjera robusnosti filostratigrafije soja *E.coli* K-12 na nasumičnoj bazi

Filostratigrafska mapa *E.coli* soja K-12 izrađena nakon pretraživanja programom BLAST s nasumičnom bazom prikazana je na slici 6. Analizu programom BLAST nije uspjelo proći 0.58% proteinskih sljedova. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima prikazan je u Tablici 12. Može se vidjeti da kao i u rezultatima filostratigrafske mape s reduciranom bazom (Slika 5.) se broj proteinskih sljedova u prvom filostratumu jako smanjio i da su se proteinski sljedovi rasporedili po ostalim filostratumima. Na rezultatima s reduciranom i nasumičnom bazom uočljivo da je nakon pada broja proteinskih sljedova na prijelazu iz prvog u drugi filostratum broj proteinskih sljedova počinje rasti od 5. filostratuma pa sve do 9. filostratuma.

Tablica 12. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon filostratigrafije s nasumičnom proteinskom bazom eukariota, arheja i bakterija.

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	1909	44,59%
Ps2	Staničan organizam - Bacteria	771	18,01%
Ps3	Bacteria – Grupa A	262	6,12%
Ps4	Grupa A – Grupa B	118	2,76%
Ps5	Grupa B – Grupa C	123	2,87%
Ps6	Grupa C - Proteobacteria	219	5,12%
Ps7	Proteobacteria - Gammaproteobacteria	263	6,14%
Ps9	Enterobacteriales - Enterobacteriaceae	420	9,81%
Ps10	Enterobacteriaceae - Escherichia	124	2,90%
Ps11	Escherichia – Escherichia coli	34	0,79%
Ps12	Escherichia coli – Escherichia coli K 12	38	0,89%



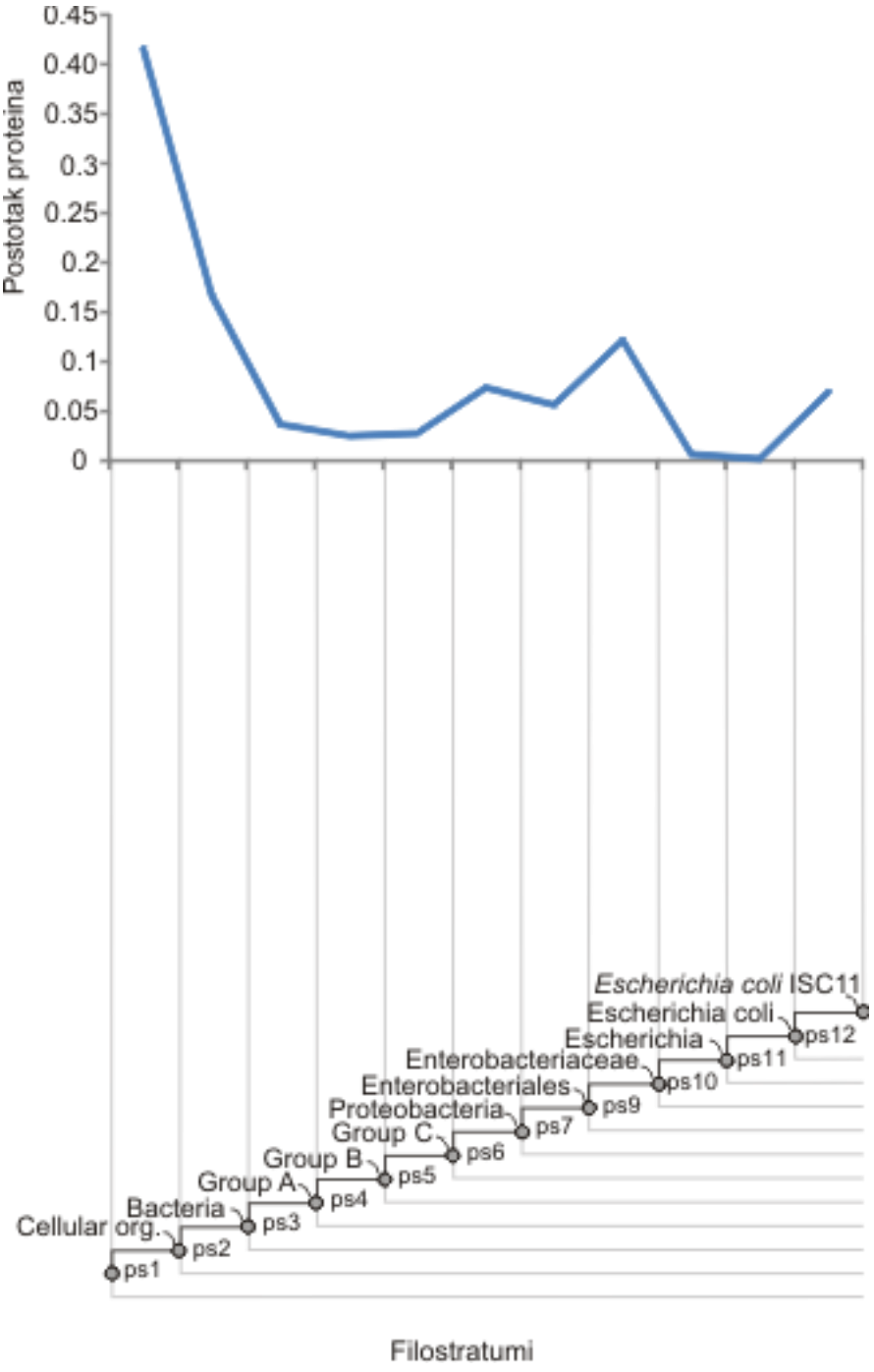
Slika 6. Filostratigrafska mapa *E.coli* soja K-12 rađena na nasumičnoj bazi eukariota, arheja i bakterija i pripadajuća konsenzus filogenija. Mapa prikazuje preraspodjelu proteinskih sljedova soja *E.coli* K-12 po pojedinim filostratumima. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* K-12.

4.3.3. Provjera robusnosti filostratigrafije soja *E.coli* ISC11 na reduciranoj bazi

Filostratigrafska mapa *E.coli* soja ISC11 izrađena je nakon pretraživanja programom BLAST s reduciranom bazom prikazana je na Slici 7. Analizu programom BLAST nije uspjelo proći 0,09% proteinskih sljedova. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima prikazan je u Tablici 13. U usporedbi s filostratigrafskom mapom dobivenom korištenjem sveobuhvatne baze može se uočiti puno manje proteinskih sljedova u prvom filostratumu te posebno povećanje broja proteinskih sljedova u 12. filostratumu. Značajan signal je prisutan u 2., 6., 7. i 9. filostratumu.

Tablica 13. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon filostratigrafije s reduciranom proteinskom bazom eukariota, arheja i bakterija.

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2545	41,56%
Ps2	Staničan organizam - Bacteria	1018	16,62%
Ps3	Bacteria – Grupa A	225	3,67%
Ps4	Grupa A – Grupa B	154	2,51%
Ps5	Grupa B – Grupa C	169	2,76%
Ps6	Grupa C - Proteobacteria	451	7,36%
Ps7	Proteobacteria - Gammaproteobacteria	345	5,63%
Ps9	Enterobacteriales - Enterobacteriaceae	743	12,13%
Ps10	Enterobacteriaceae - Escherichia	41	0,67%
Ps11	Escherichia – Escherichia coli	11	0,18%
Ps12	Escherichia coli – Escherichia coli ISC11	422	6,89%



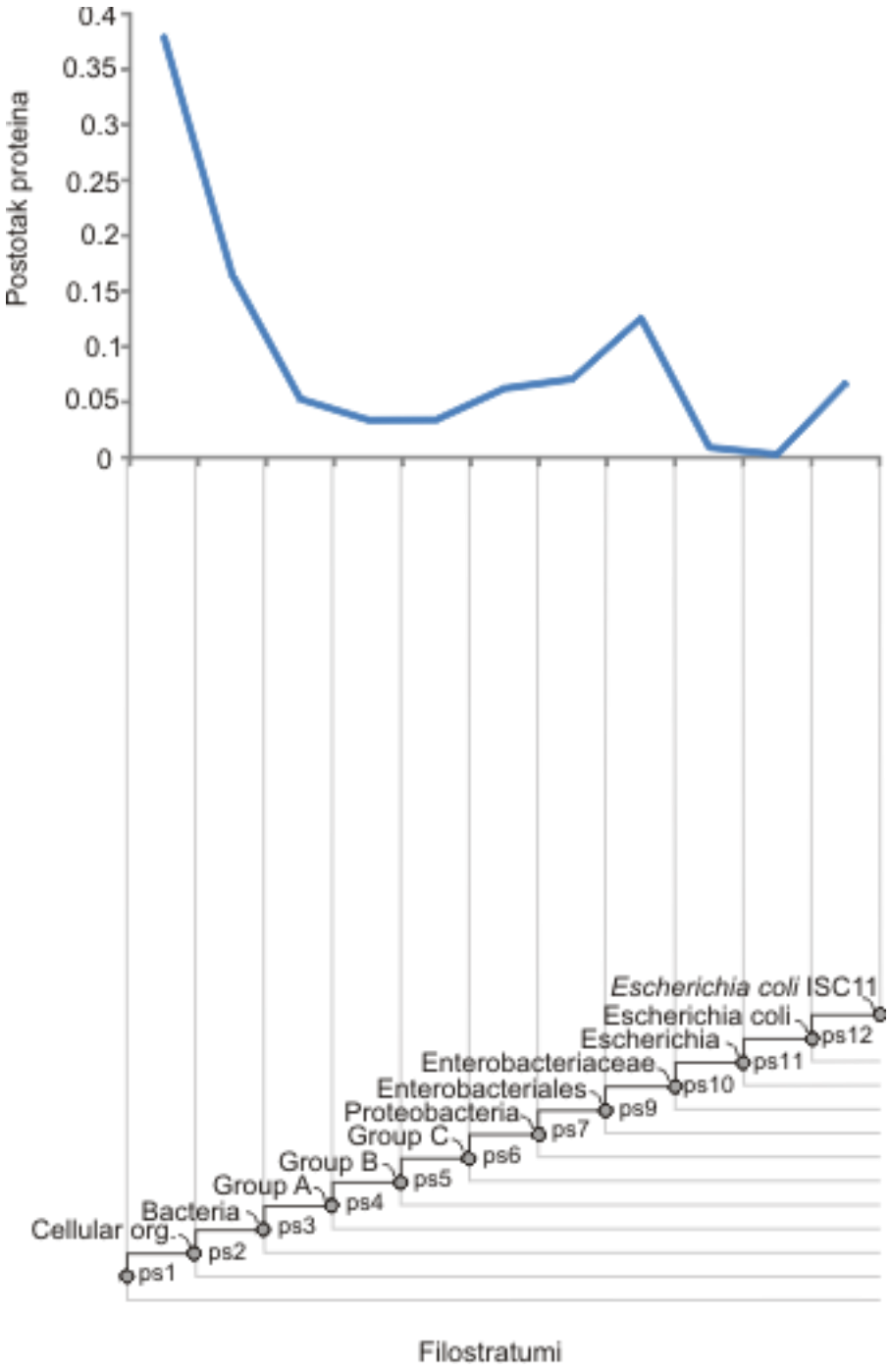
Slika 7. Filostratigrafska mapa *E.coli* soja ISC11 rađena na reduciranoj bazi eukariota, arheja i bakterija i pripadajuća konsenzus filogenija. Mapa prikazuje preraspodjelu proteinskih sljedova soja *E.coli* ISC11 po pojedinim filostratumima. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* ISC11.

4.3.4. Provjera robusnosti filostratigrafije soja *E.coli* ISC11 na nasumičnoj bazi

Filostratigrafska mapa *E.coli* soja ISC11 izrađena je nakon pretraživanja programom BLAST s nasumičnom bazom prikazana je na Slici 8. Analizu programom BLAST nije uspjelo proći 0,09% proteinskih sljedova. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima prikazan je u Tablici 14. U usporedbi s filostratigrafskom mapom dobivenom korištenjem reducirane baze može se uočiti još manje proteinskih sljedova u prvom filostratumu. Kao i u rezultatima s reduciranom bazom značajan signal je prisutan u najmlađem 12. filostratumu te u 6., 7. i 9. filostratumu.

Tablica 14. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon filostratigrafije s nasumičnom proteinskom bazom eukariota, arheja i bakterija.

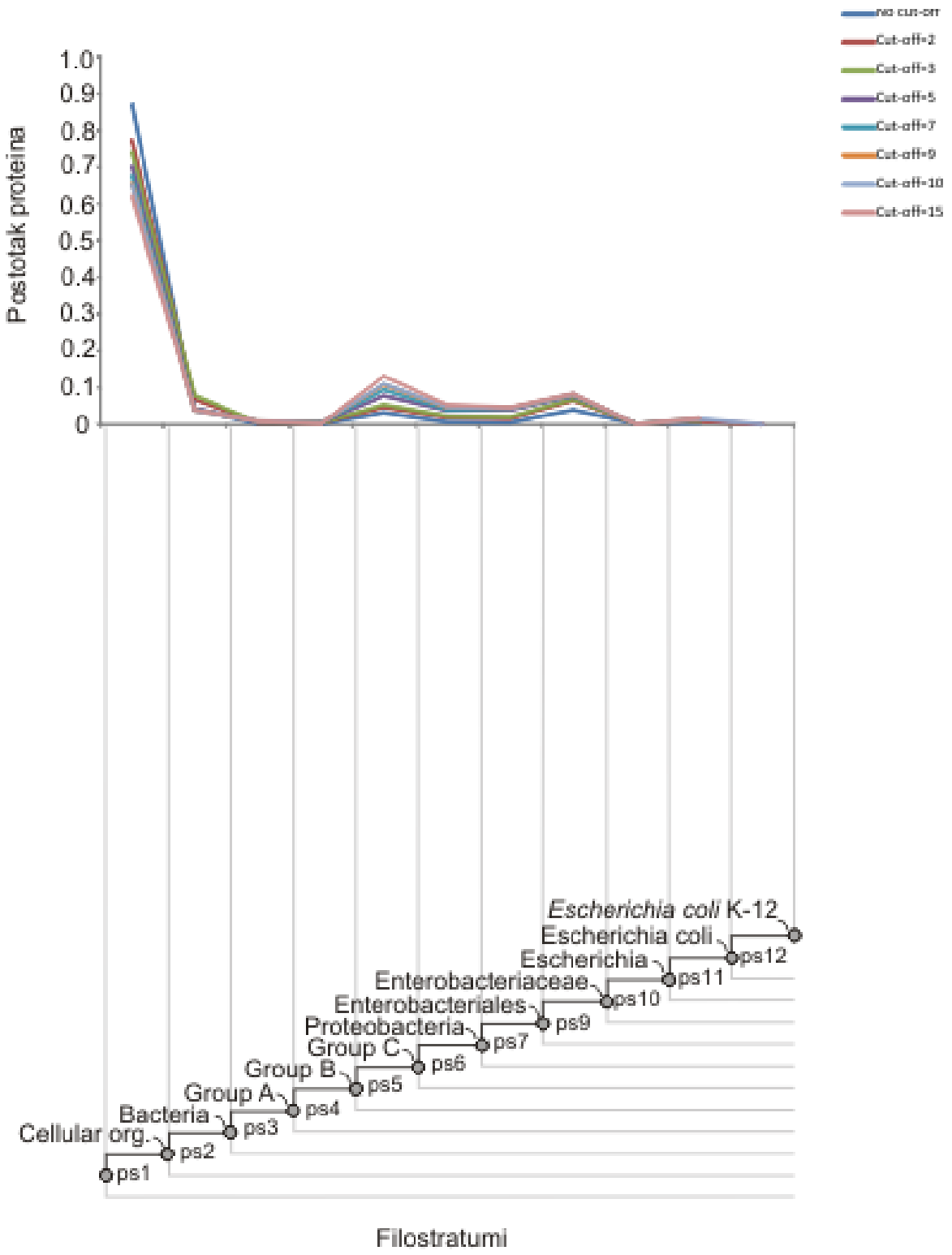
Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2319	37,87%
Ps2	Staničan organizam - Bacteria	1007	16,44%
Ps3	Bacteria – Grupa A	323	5,27%
Ps4	Grupa A – Grupa B	206	3,36%
Ps5	Grupa B – Grupa C	208	3,40%
Ps6	Grupa C - Proteobacteria	381	6,22%
Ps7	Proteobacteria - Gammaproteobacteria	433	7,07%
Ps9	Enterobacteriales - Enterobacteriaceae	769	12,56%
Ps10	Enterobacteriaceae - Escherichia	56	0,91%
Ps11	Escherichia – Escherichia coli	16	0,26%
Ps12	Escherichia coli – Escherichia coli ISC11	406	6,63%



Slika 8. Filostratigrafska mapa *E.coli* soja ISC11 rađena na nasumičnoj bazi eukariota, arheja i bakterija i pripadajuća konsenzus filogenija. Mapa prikazuje preraspodjelu proteinskih sljedova soja *E.coli* ISC11 po pojedinim filostratumima. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* ISC11.

4.4. Provjera robusnosti filostratigrafije soja *E.coli* K-12 promjenom broja hitova po filostratumu

Rezultati pretraživanja sveobuhvatne baze s proteinskim sljedovim iz proteoma *E.coli* K-12 pomoću programa BLAST su analizirani na način da su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak zadanoj graničnoj vrijednosti. Analize su napravljene s graničnim vrijednostima 2,3,7,9,10 i 15 te su ti rezultati stratificirani i prikazani na Slici 9. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima za svaku zadanu graničnu vrijednost prikazan je u Tablicima od rednog broja 15 do 21. Povećanjem granične vrijednosti postotak proteinskih sljedova u prvom filostratumu pada, a postotak proteinskih sljedova u mlađim filostratumima raste.



Slika 9. Odnosi filostratigrafskih mapa soja *E.coli* K-12 izrađenih na sveobuhvatnoj proteinskoj bazi eukariota, arheja i bakterija s različitim graničnim vrijednostima iznad kojih proteinski slijed može biti smješten u pripadajući filostratum. Prikazani su rezultati bez granične vrijednosti te rezultati s graničnim vrijednostima 2,3,5,7,9,10 i 15. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* K-12.

Tablica 15. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 2

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3301	77,14%
Ps2	Staničan organizam – Bacteria	289	6,75%
Ps3	Bacteria – Grupa A	14	0,33%
Ps4	Grupa A – Grupa B	14	0,33%
Ps5	Grupa B – Grupa C	197	4,60%
Ps6	Grupa C – Proteobacteria	79	1,85%
Ps7	Proteobacteria – Gammaproteobacteria	66	1,54%
Ps9	Enterobacteriales – Enterobacteriaceae	277	6,47%
Ps10	Enterobacteriaceae – Escherichia	4	0,09%
Ps11	Escherichia – Escherichia coli	37	0,86%
Ps12	Escherichia coli – Escherichia coli K 12	1	0,02%

Tablica 16. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 3

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3167	74,03%
Ps2	Staničan organizam - Bacteria	336	7,85%
Ps3	Bacteria – Grupa A	19	0,44%
Ps4	Grupa A – Grupa B	14	0,33%
Ps5	Grupa B – Grupa C	227	5,31%
Ps6	Grupa C - Proteobacteria	95	2,22%
Ps7	Proteobacteria - Gammaproteobacteria	85	1,99%
Ps9	Enterobacteriales - Enterobacteriaceae	291	6,80%
Ps10	Enterobacteriaceae - Escherichia	4	0,09%
Ps11	Escherichia – Escherichia coli	40	0,94%
Ps12	Escherichia coli – Escherichia coli K 12	0	0,0%

Tablica 17. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 5

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3005	70,28%
Ps2	Staničan organizam - Bacteria	184	4,30%
Ps3	Bacteria – Grupa A	25	0,58%
Ps4	Grupa A – Grupa B	21	0,49%
Ps5	Grupa B – Grupa C	332	7,76%
Ps6	Grupa C - Proteobacteria	162	3,79%
Ps7	Proteobacteria - Gammaproteobacteria	165	3,86%
Ps9	Enterobacteriales - Enterobacteriaceae	330	7,72%
Ps10	Enterobacteriaceae - Escherichia	3	0,07%
Ps11	Escherichia – Escherichia coli	49	1,15%
Ps12	Escherichia coli – Escherichia coli K 12	0	0,0%

Tablica 18. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 7

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2896	67,74%
Ps2	Staničan organizam - Bacteria	171	4,00%
Ps3	Bacteria – Grupa A	33	0,77%
Ps4	Grupa A – Grupa B	19	0,44%
Ps5	Grupa B – Grupa C	400	9,36%
Ps6	Grupa C - Proteobacteria	170	3,98%
Ps7	Proteobacteria - Gammaproteobacteria	184	4,30%
Ps9	Enterobacteriales - Enterobacteriaceae	336	7,86%
Ps10	Enterobacteriaceae - Escherichia	4	0,09%
Ps11	Escherichia – Escherichia coli	62	1,45%
Ps12	Escherichia coli – Escherichia coli K 12	0	0,0%

Tablica 19. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 9

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2805	65,68%
Ps2	Staničan organizam - Bacteria	159	3,72%
Ps3	Bacteria – Grupa A	39	0,91%
Ps4	Grupa A – Grupa B	17	0,40%
Ps5	Grupa B – Grupa C	456	10,68%
Ps6	Grupa C - Proteobacteria	194	4,54%
Ps7	Proteobacteria - Gammaproteobacteria	189	4,43%
Ps9	Enterobacteriales - Enterobacteriaceae	344	8,05%
Ps10	Enterobacteriaceae - Escherichia	4	0,09%
Ps11	Escherichia – Escherichia coli	64	1,50%
Ps12	Escherichia coli – Escherichia coli K 12	0	0,0%

Tablica 20. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 10

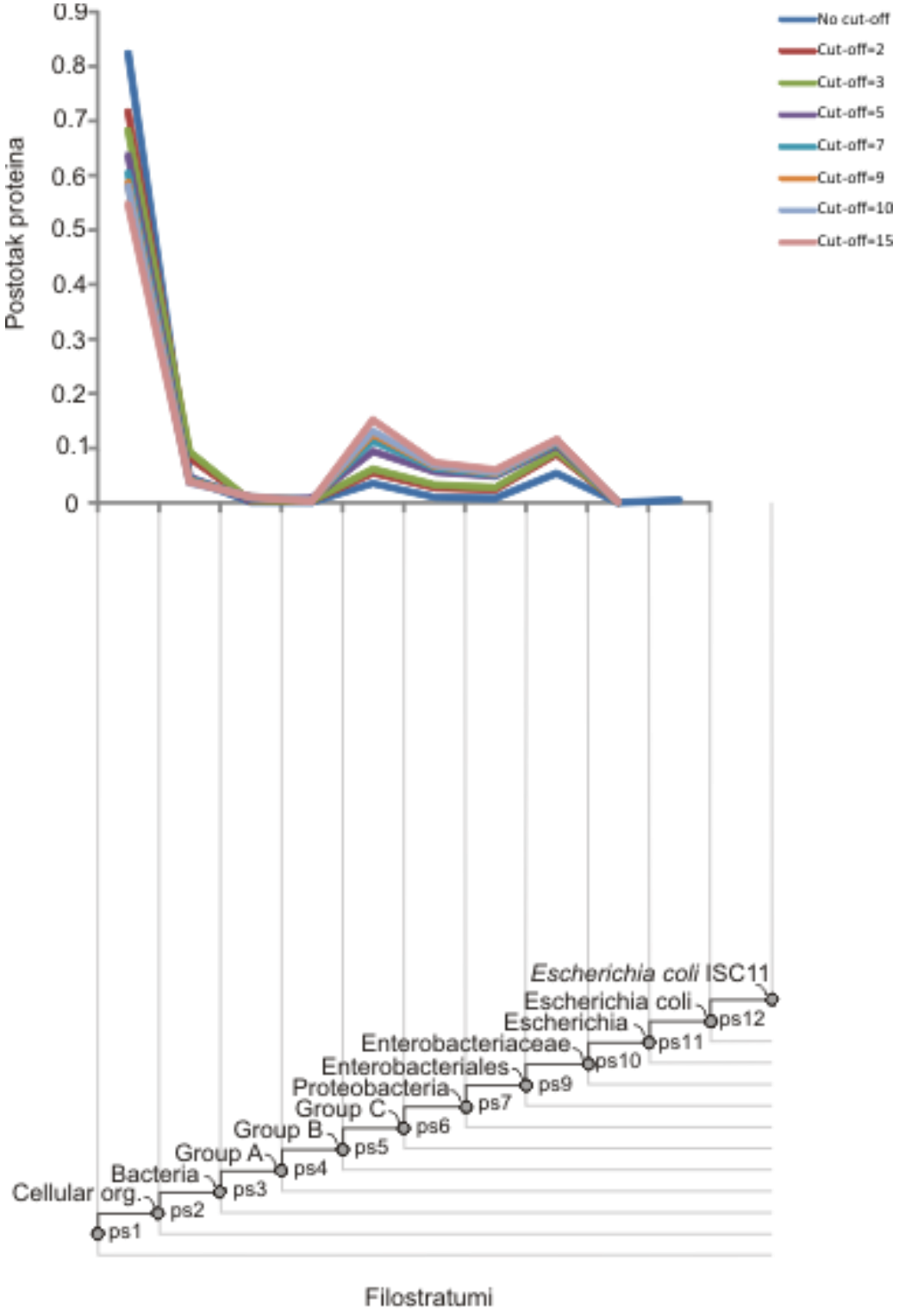
Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2783	65,21%
Ps2	Staničan organizam - Bacteria	157	3,68%
Ps3	Bacteria – Grupa A	40	0,94%
Ps4	Grupa A – Grupa B	15	0,35%
Ps5	Grupa B – Grupa C	466	10,92%
Ps6	Grupa C - Proteobacteria	202	4,73%
Ps7	Proteobacteria - Gammaproteobacteria	187	4,38%
Ps9	Enterobacteriales - Enterobacteriaceae	350	8,20%
Ps10	Enterobacteriaceae - Escherichia	3	0,07%
Ps11	Escherichia – Escherichia coli	64	1,50%
Ps12	Escherichia coli – Escherichia coli K 12	1	0,02%

Tablica 21. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* K-12 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 15

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	2640	62,02%
Ps2	Staničan organizam - Bacteria	168	3,95%
Ps3	Bacteria – Grupa A	36	0,85%
Ps4	Grupa A – Grupa B	7	0,16%
Ps5	Grupa B – Grupa C	557	13,08%
Ps6	Grupa C - Proteobacteria	226	5,31%
Ps7	Proteobacteria - Gammaproteobacteria	196	4,60%
Ps9	Enterobacteriales - Enterobacteriaceae	355	8,34%
Ps10	Enterobacteriaceae - Escherichia	2	0,05%
Ps11	Escherichia – Escherichia coli	70	1,64%
Ps12	Escherichia coli – Escherichia coli K 12	0	0,0%

4.5. Provjera robusnosti filostratigrafije soja *E.coli* ISC11 promjenom broja hitova po filostratumu

Rezultati pretraživanja sveobuhvatne baze s proteinskim sljedovim iz proteoma *E.coli* ISC11 pomoću programa BLAST su analizirani na način da su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak zadanoj graničnoj vrijednosti. Analize su napravljene s graničnim vrijednostima 2,3,7,9,10 i 15 te su ti rezultati stratificirani i prikazani na Slici 10. Broj i postotak proteinskih sljedova preraspoređenih po filostratumima za svaku zadanu graničnu vrijednost prikazan je u Tablicima od rednog broja 22. do 28. Povećanjem granične vrijednosti postotak proteinskih sljedova u prvom filostratumu i ovdje pada, a postotak proteinskih sljedova u mlađim filostratumima raste. Najveće povećanje proteinskih sljedova je u 5.,6.,7. i 9. filostratumu. Zanimljivo je da su u 10. i 12. filostratumu povećanjem kriterija proteinski sljedovi nestali te su ti filostratumi ostali prazni.



Slika 10. Odnosi filostratigrafskih mapa soja *E.coli* ISC11 izrađenih na sveobuhvatnoj proteinskoj bazi eukariota, arheja i bakterija s različitim graničnim vrijednostima iznad kojih proteinski slijed može biti smješten u pripadajući filostratum. Prikazani su rezultati bez granične vrijednosti te rezultati s graničnim vrijednostima 2,3,5,7,9,10 i 15. Čvorovi predstavljaju skupine koje sadrže hipotetskog pretka *E.coli* ISC11.

Tablica 22. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 2

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	4145	71,60%
Ps2	Staničan organizam - Bacteria	479	8,27%
Ps3	Bacteria – Grupa A	16	0,28%
Ps4	Grupa A – Grupa B	20	0,35%
Ps5	Grupa B – Grupa C	315	5,44%
Ps6	Grupa C - Proteobacteria	158	2,73%
Ps7	Proteobacteria - Gammaproteobacteria	135	2,33%
Ps9	Enterobacteriales - Enterobacteriaceae	513	8,86%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	8	0,14%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

Tablica 23. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 3

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3934	68,19%
Ps2	Staničan organizam - Bacteria	543	9,41%
Ps3	Bacteria – Grupa A	23	0,40%
Ps4	Grupa A – Grupa B	26	0,45%
Ps5	Grupa B – Grupa C	353	6,12%
Ps6	Grupa C - Proteobacteria	187	3,24%
Ps7	Proteobacteria - Gammaproteobacteria	160	2,77%
Ps9	Enterobacteriales - Enterobacteriaceae	537	9,31%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	6	0,10%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

Tablica 24. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 5

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3655	63,62%
Ps2	Staničan organizam - Bacteria	267	4,65%
Ps3	Bacteria – Grupa A	30	0,52%
Ps4	Grupa A – Grupa B	46	0,80%
Ps5	Grupa B – Grupa C	538	9,36%
Ps6	Grupa C - Proteobacteria	325	5,66%
Ps7	Proteobacteria - Gammaproteobacteria	281	4,89%
Ps9	Enterobacteriales - Enterobacteriaceae	600	10,44%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	3	0,05%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

Tablica 25. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 7

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3468	60,45%
Ps2	Staničan organizam - Bacteria	246	4,29%
Ps3	Bacteria – Grupa A	52	0,91%
Ps4	Grupa A – Grupa B	27	0,47%
Ps5	Grupa B – Grupa C	650	11,33%
Ps6	Grupa C - Proteobacteria	367	6,40%
Ps7	Proteobacteria - Gammaproteobacteria	298	5,19%
Ps9	Enterobacteriales - Enterobacteriaceae	626	10,91%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	3	0,05%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

Tablica 26. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 9

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3358	58,62%
Ps2	Staničan organizam - Bacteria	224	3,91%
Ps3	Bacteria – Grupa A	59	1,03%
Ps4	Grupa A – Grupa B	23	0,40%
Ps5	Grupa B – Grupa C	721	12,59%
Ps6	Grupa C - Proteobacteria	386	6,74%
Ps7	Proteobacteria - Gammaproteobacteria	311	5,43%
Ps9	Enterobacteriales - Enterobacteriaceae	641	11,19%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	5	0,09%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

Tablica 27. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 10

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3318	57,94%
Ps2	Staničan organizam - Bacteria	208	3,63%
Ps3	Bacteria – Grupa A	64	1,12%
Ps4	Grupa A – Grupa B	21	0,37%
Ps5	Grupa B – Grupa C	752	13,13%
Ps6	Grupa C - Proteobacteria	392	6,84%
Ps7	Proteobacteria - Gammaproteobacteria	318	5,55%
Ps9	Enterobacteriales - Enterobacteriaceae	649	11,33%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	5	0,09%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

Tablica 28. Tablica prikazuje broj i postotak proteinskih sljedova soja *E.coli* ISC11 preraspodijeljenih po pojedinim filostratumima nakon što su u filostratigrafiji korišteni rezultati programa BLAST u kojem su uklonjeni oni rezultati koji imaju broj hitova s različitim organizmima iz nekog filostratuma ispod ili jednak graničnoj vrijednosti 15

Broj filostratuma	Naziv filostratuma	Broj proteinskih sljedova soja	Postotak ukupnog broja proteinskih sljedova soja
Ps1	Organizmi nastali prije zadnjeg zajedničkog pretka staničnih organizama – staničan organizam	3123	54,67%
Ps2	Staničan organizam - Bacteria	218	3,83%
Ps3	Bacteria – Grupa A	60	1,05%
Ps4	Grupa A – Grupa B	14	0,25%
Ps5	Grupa B – Grupa C	865	15,14%
Ps6	Grupa C - Proteobacteria	418	7,32%
Ps7	Proteobacteria - Gammaproteobacteria	344	6,02%
Ps9	Enterobacteriales - Enterobacteriaceae	663	11,61%
Ps10	Enterobacteriaceae - Escherichia	0	0,0%
Ps11	Escherichia – Escherichia coli	7	0,12%
Ps12	Escherichia coli – Escherichia coli ISC11	0	0,0%

4.7. Analiza proteina iz pojedinih filostratuma prema rječniku Gene Ontology

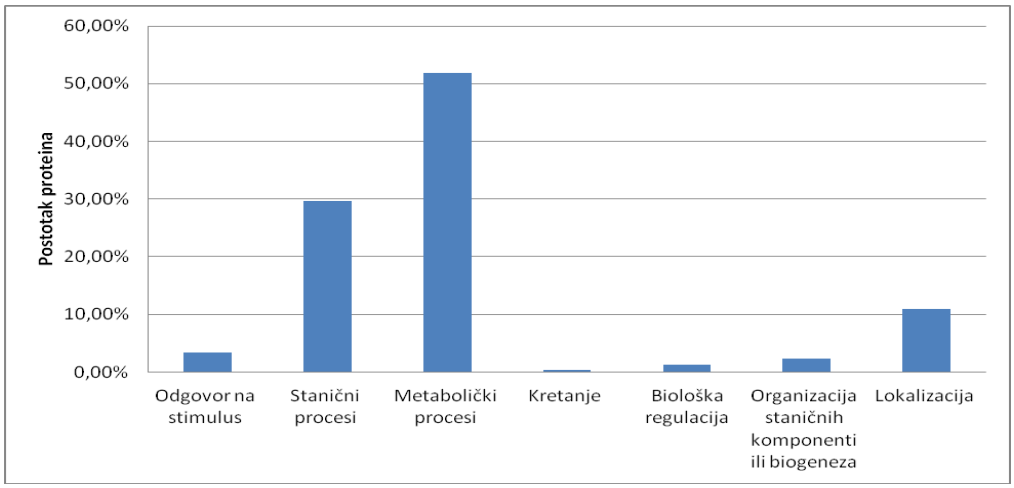
Proteinskim sljedovima iz prvog filostratuma koji su kategorizirani kao najstariji te proteinskim sljedovima iz 9.,10.,11. i 12. koji su kategorizirani kao mlađi analizirana je uloga u biološkim procesima, molekularna funkcija i stanična lokalizacija prema anotaciji rječnika Gene Ontology. Upotrijebljeni su rezultati dobiveni korištenjem sveobuhvatne, reducirane i nasumične baze te rezultati s najvišom graničnom vrijednošću soja K-12 budući da je njegov proteom dio baze Gene Ontolgy. Kako je većina mladih proteina kategorizirana kao neklasificirani s obzirom na biološke procese, molekularnu funkciju i staničnu lokalizaciju to je potvrđeno alatom Enrichment analysis. Navedeni alat pronalazi prisutnost određenih GO kategorija u zadanom setu proteina s obzirom na čitav proteom. Kao rezultati te analize navedene su tablice koje sadrže 7 stupaca. Prvi stupac sadrži pronađene GO kategorije, drugi broj proteina iz te kategorije u čitavom proteomu, treći broj proteina iz kategorije u analiziranom setu proteina, četvrti očekivani broj proteina te kategorije u analiziranom setu s obzirom na cjelokupni proteom, peti povećanu ili smanjenu prisutnost proteina iz te kategorije, šesti izračunato povećanje ili smanjenje proteina iz pojedine kategorije s obzirom na pojavnost u čitavom proteomu, a sedmi stupac izračunatu p-vrijednosti koja nam govori kolika je vjerojatnost da će se toliki broj proteina neke kategorije u analiziranom setu pojaviti s obzirom na broj proteina te kategorije u čitavom proteomu.

4.7.1. Analiza proteina iz pojedinih filostratuma mape *E.coli* K-12 dobivene korištenjem sveobuhvatne baze

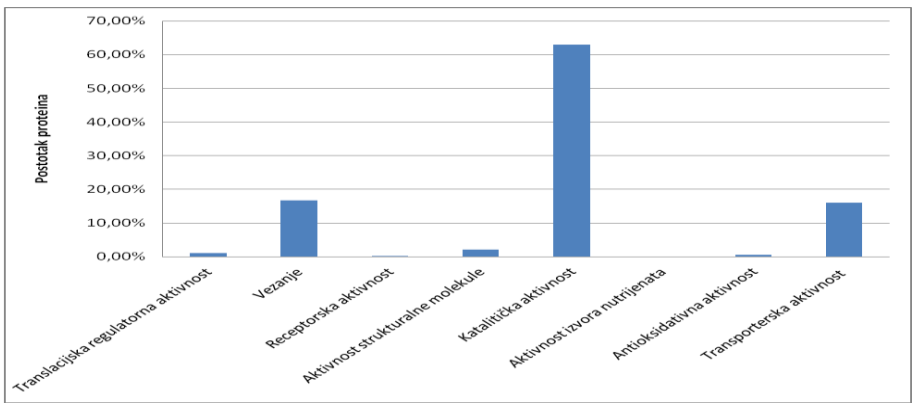
Od svih proteinskih sljedova u proteomu *E.coli* K-12 njih 2606 ima poznatu ulogu u biološkim procesima, 2217 poznatu molekularnu funkciju i 939 poznatu staničnu lokaciju. Od 3683 proteinskih sljedova iz prvog filostratuma 2584 ima poznatu ulogu u biološkim procesima, 2203 molekularnu funkciju i 931 poznatu staničnu lokazilaciju. Proteini iz prvog filostratuma najzastupljeniju ulogu imaju u metaboličkim i staničnim procesima te lokalizaciji (Slika 11.). Najčešća molekularna funkcija tih proteina je katalitička aktivnost, vezanje i transportna aktivnost (Slika 12.).

U stanici se najviše nalaze u njenoj unutrašnjosti, membrani i makromolekulskim kompleksima (Slika 13.).

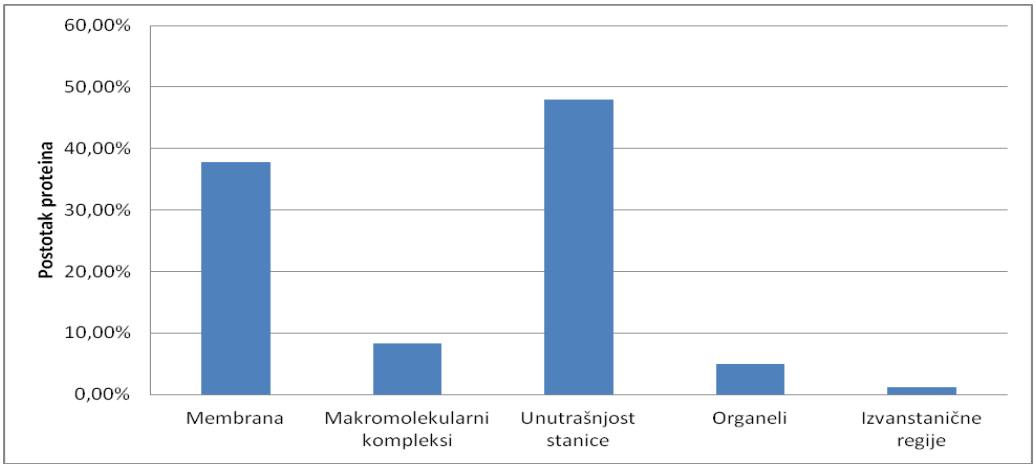
Proteini iz mlađih filostratuma su s obzirom na ulogu u biološkim procesima, molekularnu funkciju i staničnu lokaciju neklasificirani jer su sve tri kategorije nepoznate budući da se radi većinom o nekarakteriziranim proteinima. To je potvrđeno pomoću alata Enrichment Analysis koji u određenom setu proteina određuje koje GO anotacijske kategorije su prezastupljene u odnosu na skupinu iz koje je set uzet što je u ovom slučaju proteom *E.coli* K-12. Jedini prezastupljeni pojam koji se pojavljuje u sve tri promatrane kategorije je pojam neklasificirano koji je potvrđen niskom P-vrijednošću (Tablica 29.,30. i 31.).



Slika 11. Prikaz zastupljenosti pojedinih bioloških procesa prema Gene Ontology anotaciji u kojima sudjeluju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem sveobuhvatne baze.



Slika 12. Prikaz zastupljenosti pojedinih molekularnih funkcija prema Gene Ontology anotaciji koje imaju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem sveobuhvatne baze.



Slika 13. Prikaz zastupljenosti pojedinih staničnih lokacija prema Gene Ontology anotaciji u kojima su locirani proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem sveobuhvatne baze.

Tablica 29. Rezultati Enrichment Analysis alata za proteine iz mlađih filostratuma filostratigrafske mape *E.coli* K-12 na sveobuhvatnoj bazi s obzirom na ulogu proteina u biološkim procesima.

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1649	151	73,9	+	2,04	0,00E+00
Biološki procesi	2613	40	117,1	-	0,34	3,96E-27
Stanični procesi	1889	26	84,65	-	0,31	5,51E-17
Stanični procesi jednostaničnog organizma	1024	10	45,89	-	0,22	2,01E-09
Biosintetski procesi jednostaničnog organizma	417	4	18,69	-	0,21	1,95E-02

Tablica 30. Rezultati Enrichment Analysis alata za proteine iz mlađih filostratuma filostratigrafske mape *E.coli* K-12 na sveobuhvatnoj bazi s obzirom na molekularnu funkciju proteina.

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1712	173	76,72	+	2,25	0,00E+00
Vezanje proteina	926	13	41,5	-	0,31	1,07E-05
Vezanje	1474	15	66,06	-	0,23	2,26E-15

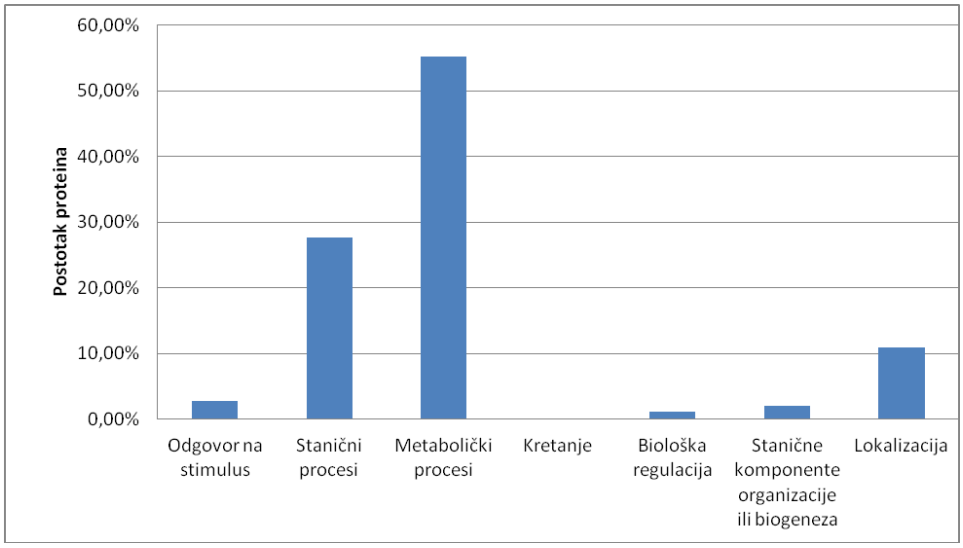
Tablica 31. Rezultati Enrichment Analysis alata za proteine iz mlađih filostratuma filostratigrafske mape *E.coli* K-12 na sveobuhvatnoj bazi s obzirom na staničnu lokalizaciju proteina.

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1922	161	86,13	+	1,87	0,00E+00
Stanična periferija	1013	19	45,4	-	0,42	8,29E-05
Membrana	1131	20	50,69	-	0,39	3,45E-06
Plazma membrana	926	16	41,5	-	0,39	7,05E-05
Stanična unutrašnjost	2286	30	102,45	-	0,29	1,54E-25
Stanica	2287	30	102,49	-	0,29	1,44E-25
Stanične komponente	2340	30	104,87	-	0,29	3,56E-27
Unutrašnje komponente membrane	487	5	21,82	-	0,23	7,93E-04
Integralni dijelovi plazma membrane	319	3	14,3	-	0,21	2,69E-02
Dijelovi membrane	532	5	23,84	-	0,21	1,30E-04

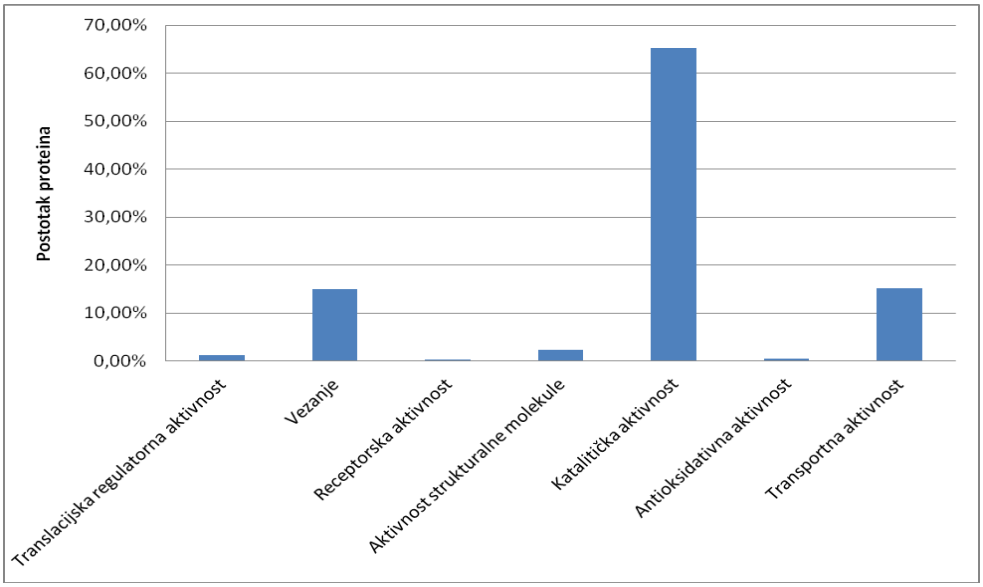
4.7.2. Analiza proteina iz pojedinih filostratuma mape *E.coli* K-12 dobivene korištenjem reducirane baze

Od 2091 proteinskog slijeda iz prvog filostratuma mape dobivene korištenjem reducirane baze 1951 protein ima poznatu ulogu u biološkim procesima, 1767 molekularnu funkciju i 657 poznatu staničnu lokalizaciju. Proteini iz prvog filostratuma najzastupljeniju ulogu imaju u metaboličkim i staničnim procesima te lokalizaciji (Slika 14.) . Najčešća molekularna funkcija tih proteina je katalitička aktivnost, vezanje i transportna aktivnost (Slika 15.). U stanici se najviše nalaze u njenoj unutrašnjosti, membrani i makromolekulskim kompleksima (Slika 16.). Unatoč tome što je ovdje puno manje proteinskih sljedova nego u prvom filostratumu mape dobivenoj na sveobuhvatnoj bazi uloga proteina, molekularna funkcija i stanična lokalizacija je gotovo pa ista.

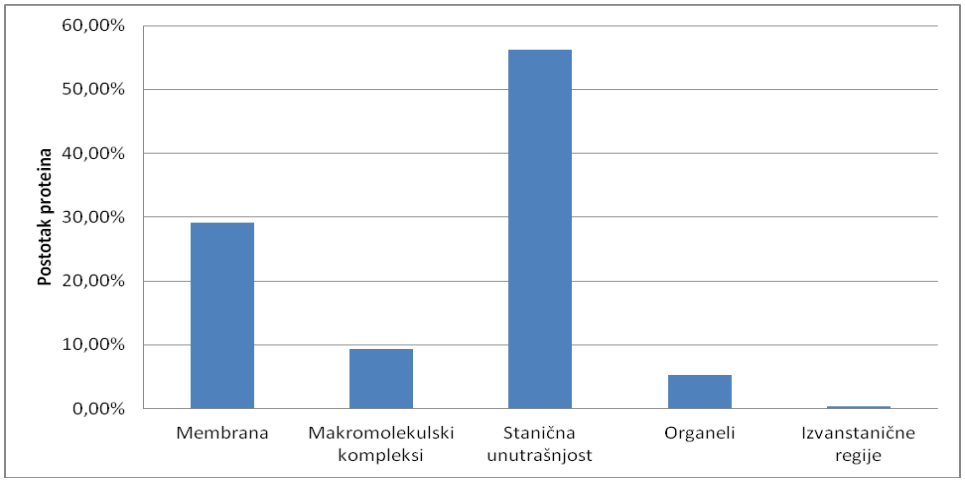
U mlađim (9.,10.,11. i 12) filostratumima s obzirom na ulogu u biološkim procesima prezastupljeni su proteini koji imaju ulogu u formiranju biofilma, biološkoj i staničnoj adheziji te neklasificirani proteini (Tablica 32.). Prezastupljeni proteini s obzirom na molekularnu funkciju su oni neklasificirani (Tablica 34.), a prezastupljeni proteini s obzirom na lokaciju u stanici su oni koji se nalaze na pilusu i staničnim projekcijama te opet neklasificirani proteini (Tablica 33.)



Slika 14. Prikaz zastupljenosti pojedinih bioloških procesa prema Gene Ontology anotaciji u kojima sudjeluju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem reducirane baze.



Slika 15. Prikaz zastupljenosti pojedinih molekularnih funkcija prema Gene Ontology anotaciji koje obavljaju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem reducirane baze.



Slika 16. Prikaz zastupljenosti pojedinih staničnih lokacija prema Gene Ontology anotaciji u kojima su locirani proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem reducirane baze.

Tablica 32. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 na reduciranoj bazi s obzirom na ulogu u staničnim procesima

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Mogućnost stvaranja biofilma	17	11	2,39	+	4,61	2,85E-02
Stanična adhezija	31	18	4,35	+	4,14	5,16E-04
Biološka adhezija	34	18	4,77	+	3,77	1,85E-03
Neklasificirano	1649	431	231,37	+	1,86	0,00E+00
Odgovor na stimulus	688	56	96,53	-	0,58	7,72E-04
Biološki procesi	2613	167	366,63	-	0,46	1,38E-58
Stanični procesi	1889	104	265,05	-	0,39	1,77E-41
Stanični makromolekularni biosintetski procesi	304	15	42,65	-	0,35	3,63E-04
Makromolekularni biosintetski procesi	306	15	42,93	-	0,35	2,97E-04
RNA metabolički procesi	277	13	38,87	-	0,33	5,89E-04
Metabolizam ugljikohidrata jenostaničnih organizama	196	9	27,5	-	0,33	2,25E-02
Metabolizam nukleinskih kiselina	377	17	52,9	-	0,32	1,99E-06
Metabolizam ugljikohidrata	207	9	29,04	-	0,31	7,28E-03
Makromolekularni metabolički procesi	729	31	102,29	-	0,3	7,14E-16

Tablica 33. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 na reduciranoj bazi s obzirom na staničnu lokalizaciju proteina.

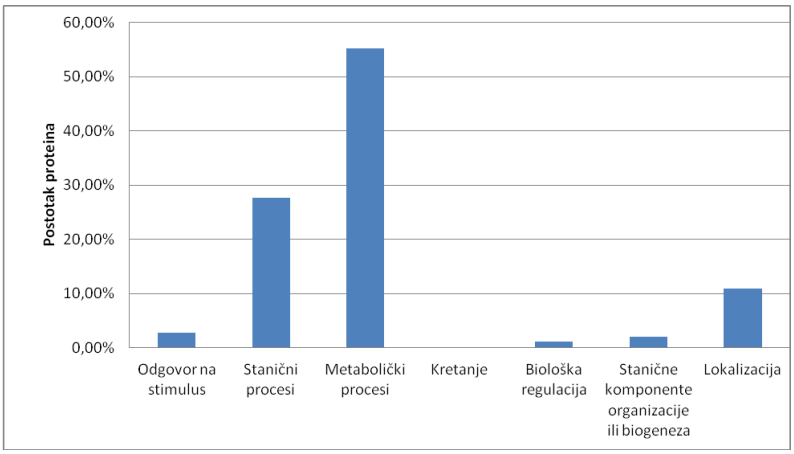
Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Pilus	24	18	3,37	+	5,35	1,76E-06
Stanične projekcije	34	19	4,77	+	3,98	6,23E-05
Neklasificirano	1922	460	269,68	+	1,71	0,00E+00
Stanična periferija	1013	85	142,13	-	0,6	4,73E-07
Membrana	1131	88	158,69	-	0,55	2,75E-10
Plazma membrana	926	70	129,93	-	0,54	1,48E-08
Stanična unutrašnjost	2286	136	320,75	-	0,42	8,91E-52
Stanica	2287	136	320,89	-	0,42	7,48E-52
Stanične komponente	2340	138	328,32	-	0,42	9,72E-55
Dijelovi membrane	532	19	74,64	-	0,25	9,41E-14
Unutrašnje komponente membrane	487	17	68,33	-	0,25	1,10E-12
Integralni dijelovi plazma membrane	319	11	44,76	-	0,25	6,44E-08
Unutrašnji dijelovi plazma membrane	335	11	47	-	0,23	1,00E-08
Integralni dijelovi membrane	458	15	64,26	-	0,23	1,80E-12
Dijelovi plazma membrane	409	12	57,39	-	0,21	6,39E-12

Tablica 34. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 na reduciranoj bazi s obzirom na molekularnu funkciju proteina.

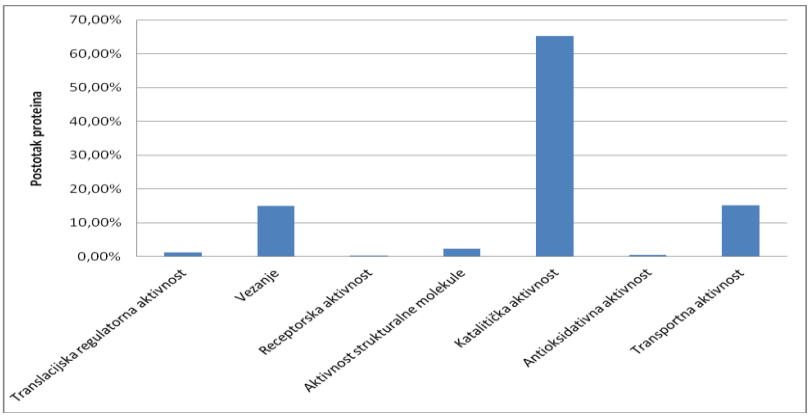
Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1712	514	240,21	+	2,14	0,00E+00
Vezanje proteina	926	48	129,93	-	0,37	8,96E-17
Vezanje	1474	60	206,82	-	0,29	2,53E-41
Vezanje DNA	205	8	28,76	-	0,28	2,11E-03
Molekularna funkcija	2550	84	357,79	-	0,23	3,61E-116
Vezanje nukleinskih kiselina	275	9	38,59	-	0,23	3,29E-06

4.7.3. Analiza proteina iz pojedinih filostratuma mape *E.coli* K-12 dobivene korištenjem nasumične baze

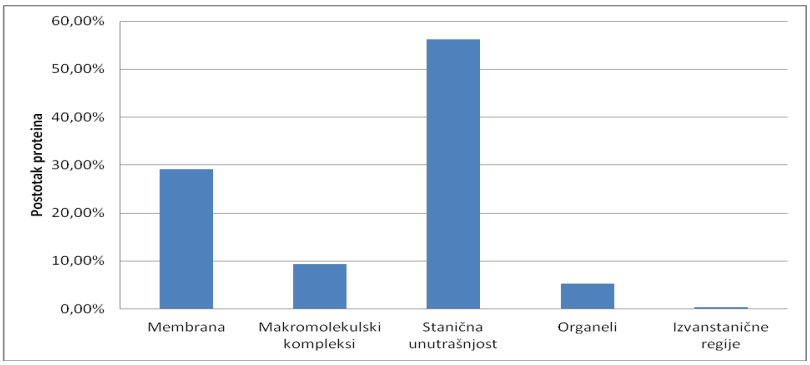
Od 1903 proteinska slijeda iz prvog filostratuma mape dobivene korištenjem nasumične baze 1790 proteina ima poznatu ulogu u biološkim procesima, 1651 molekularnu funkciju i 603 poznatu staničnu lokazilaciju. Proteini iz prvog filostratuma najzastupljeniju ulogu imaju u metaboličkim i staničnim procesima te lokalizaciji (Slika 17.) . Najčešća molekularna funkcija tih proteina je katalitička aktivnost, vezanje i transportna aktivnost (Slika 18.). U stanici se najviše nalaze u njenoj unutrašnjosti i mebrani (Slika 19.). Kao i u rezultatima dobivenim sveobuhvatnom i reduciranom bazom uloge u biološkim procesima, molekularna funkcija i stanična lokalizacija su isti. U mlađim filostratumima s obzirom na ulogu u biološkim procesima zastupljeni su proteini koji sudjeluju adheziji i raznim procesima formiranja biofilma dok s obzirom na staničnu lokaciju zastupljeni su proteini koji su dio staničnih projekcija i pilusa, ali s obzirom na oba kriterija pojavljuje se i prezastupljenost neklasificiranih proteina (Tablica 35. i 36.). S obzirom na molekularnu funkciju prezastupljeni su samo neklasificirani proteini (Tablica 37.).



Slika 17. Prikaz zastupljenosti pojedinih bioloških procesa prema Gene Ontology anotaciji u kojima sudjeluju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem nasumične baze.



Slika 18. Prikaz zastupljenosti pojedinih molekularnih funkcija prema Gene Ontology anotaciji koje obavljaju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem nasumične baze.



Slika 19. Prikaz zastupljenosti pojedinih staničnih lokacija prema Gene Ontology anotaciji u kojima su locirani proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem nasumične baze.

Tablica 35. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 na nasumičnoj bazi s obzirom na ulogu u biološkim procesima

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Biološka adhezija	34	22	4,91	+	4,48	7,94E-06
Mogućnost stvaranja potopljenog biofilma jedne vrste	17	11	2,46	+	4,48	3,72E-02
Stanična adhezija	31	19	4,48	+	4,24	1,78E-04
Mogućnost stvaranje potopljenog biofilma	23	13	3,32	+	3,91	3,15E-02
Stvaranje biofilma jedne vrste	59	23	8,53	+	2,7	1,89E-02
Stvaranje biofilma	65	25	9,39	+	2,66	1,06E-02
Stanični procesi koji uključuju više organizama	93	31	13,44	+	2,31	1,66E-02
Procesi koji uključuju više organizama	94	31	13,59	+	2,28	2,02E-02
Neklasificirano	1649	440	238,34	+	1,85	0,00E+00
Odgovor na stimulus	688	58	99,44	-	0,58	6,54E-04
Biološki procesi	2613	176	377,66	-	0,47	5,63E-58
Stanični procesi	1889	109	273,02	-	0,4	1,09E-41
Stanični makromolekulski biosintetski procesi	304	14	43,94	-	0,32	4,66E-05
Makromolekulski biosintetski procesi	306	14	44,23	-	0,32	3,76E-05

Tablica 36. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 na reduciranoj bazi s obzirom na staničnu lokalizaciju proteina.

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Pilus	24	17	3,47	+	4,9	1,47E-05
Stanične projekcije	34	18	4,91	+	3,66	3,87E-04
Neklasificirano	1922	463	277,79	+	1,67	0,00E+00
Stanična periferija	1013	96	146,41	-	0,66	4,03E-05
Membrana	1131	100	163,47	-	0,61	8,12E-08
Plazma membrana	926	79	133,84	-	0,59	9,54E-07
Stanična unutrašnjost	2286	150	330,4	-	0,45	1,21E-47
Stanica	2287	150	330,55	-	0,45	1,02E-47
Stanični dijelovi	2340	153	338,21	-	0,45	5,08E-50
Dijelovi membrane	532	26	76,89	-	0,34	1,59E-10
Unutrašnji dijelovi membrane	487	23	70,39	-	0,33	6,53E-10
Integralni dijelovi plazma membrane	319	15	46,11	-	0,33	4,23E-06
Unutrašnji dijelovi plazma mebrane	335	15	48,42	-	0,31	7,50E-07
Plazma membrana	409	18	59,11	-	0,3	8,47E-09

Tablica 37. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 na nasumičnoj bazi s obzirom na molekularnu funkciju proteina.

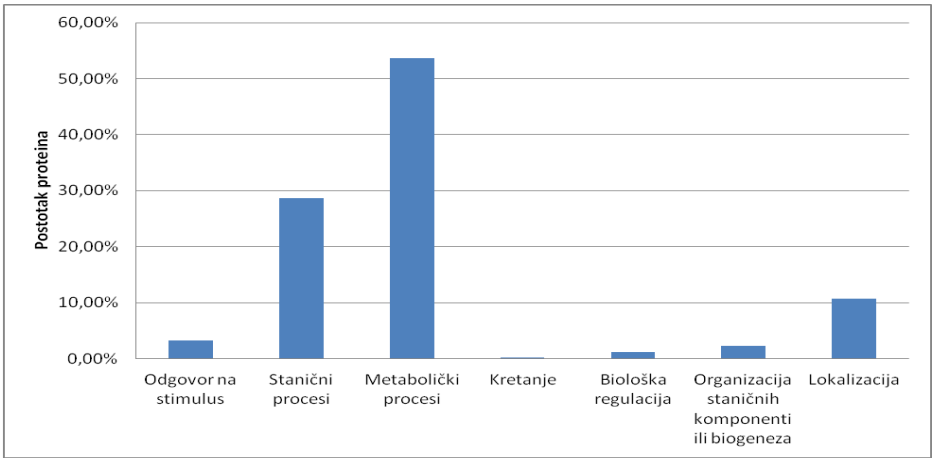
Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1712	521	247,44	+	2,11	0,00E+00
Vezanje proteina	926	54	133,84	-	0,4	4,04E-15
Vezanje DNA	205	10	29,63	-	0,34	1,25E-02
Vezanje	1474	70	213,04	-	0,33	2,55E-37
Vezanje nukleinskih kiselina	275	11	39,75	-	0,28	2,17E-05
Molekularna funkcija	2550	95	368,56	-	0,26	8,84E-112
Hidrolazna aktivnost koja djeluje na esterske veze	171	6	24,72	-	0,24	3,42E-03
Vezanje identičnih proteina	272	9	39,31	-	0,23	1,84E-06

4.7.4. Analiza proteina iz pojedinih filostratuma mape *E.coli* K-12 dobivene korištenjem granične vrijednosti

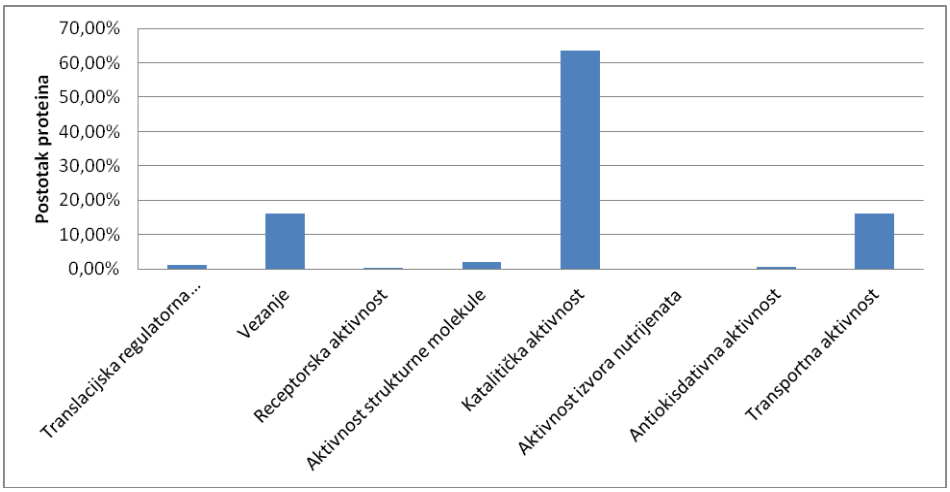
Od dobivenih filostratigrafskih mapa s graničnim vrijednostim odabrana je mapa s najvišom graničnom vrijednošću budući da se ta mapa u rezultatima najviše razlikuje od mape dobivene bez korištenja granične vrijednosti.

Od 2609 proteinskih sljedova iz prvog filostratuma mape dobivene korištenjem granične vrijednosti 15, 2280 ima poznatu ulogu u biološkim procesima, 2025 molekularnu funkciju i 793 poznatu staničnu lokalizaciju. Proteini iz prvog filostratuma najzastupljeniju ulogu opet imaju u metaboličkim i staničnim procesima te lokalizaciji (Slika 20.) . Najčešća molekularna funkcija tih proteina je katalitička aktivnost, vezanje i transportna aktivnost (Slika 21.). U stanici se najviše nalaze u unutrašnjosti stanice i mebrani (Slika 22.).

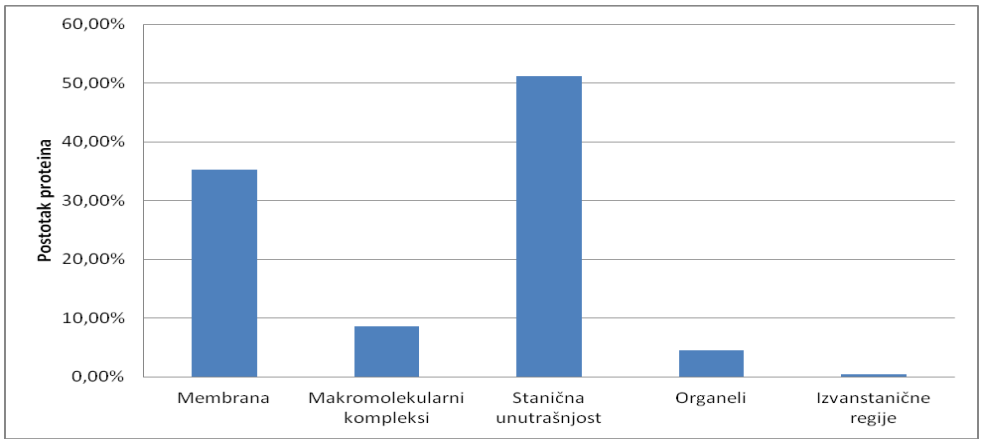
Mlađi proteini su s obzirom na sva tri promatrana kriterija prezastupljeni u GO kategoriji neklasificirano dok u svimn ostalim kategorijama imaju zastupljenost ispod očekivane s obzirom na čitav proteom (Tablica 38.,39. i 40.)



Slika 20. Prikaz zastupljenosti pojedinih bioloških procesa prema Gene Ontology anotaciji u kojima sudjeluju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem granične vrijednosti 15.



Slika 21. Prikaz zastupljenosti pojedinih molekularnih funkcija prema Gene Ontology anotaciji koje obavljaju proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem granične vrijednosti 15.



Slika 22. Prikaz zastupljenosti pojedinih staničnih lokacija prema Gene Ontology anotaciji u kojima su locirani proteinski sljedovi koji se nalaze u prvom filostratumu filostratigrafske mape *E. coli* K-12 dobivene korištenjem granične vrijednosti 15.

Tablica 38. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 dobivenu korištenjem granične vrijednosti 15 s obzirom na molekularnu funkciju proteina.

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1712	372	171,92	+	2,16	0,00E+00
Vezanje proteina	926	37	92,99	-	0,4	2,29E-10
Vezanje	1474	44	148,02	-	0,3	1,52E-28
Vezanje nukleinskih kiselina	275	8	27,62	-	0,29	4,40E-03
Molekularna funkcija	2550	56	256,08	-	0,22	7,19E-87

Tablica 39. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 dobivene korištenjem granične vrijednosti 15 s obzirom na ulogu u biološkim procesima

Naziv GO antocaijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1649	317	165,6	+	1,91	0,00E+00
Odgovor na stimulus	688	39	69,09	-	0,56	1,22E-02
Biološki procesi	2613	111	262,4	-	0,42	5,00E-47
Stanični procesi	1889	68	189,7	-	0,36	3,18E-33
Stanični makromolekulski biosintetski procesi	304	9	30,53	-	0,29	2,02E-03
Makromolekulski biosintetski procesi	306	9	30,73	-	0,29	1,73E-03
Metabolizam ugljikohidrata	207	5	20,79	-	0,24	2,14E-02
Stanični procesi koji uključuju jedan organizam	1024	23	102,83	-	0,22	1,69E-22
Biosintetski procesi koji uključuju jedan organizam	417	9	41,88	-	0,21	1,69E-07
Procesi koji uključuju jedan organizam	1414	30	142	-	0,21	9,00E-36
Makromolekulski stanični procesi	729	15	73,21	-	0,2	1,12E-15

Tablica 40. Rezultati Enrichment Analysis alata za proteine iz mlađih (9.,10.,11. i 12.) filostratuma filostratigrafske mape *E.coli* K-12 dobivene korištenjem granične vrijednosti 15 s obzirom na staničnu lokalizaciju proteina.

Naziv GO anotacijske kategorije	Broj proteina u proteomu	Stvarna prisutnost	Očekivana prisutnost	Stanje prisutnosti	Povećanje	P-vrijednost
Neklasificirano	1922	332	193,01	+	1,72	0,00E+00
Stanična periferija	1013	63	101,73	-	0,62	2,61E-04
Membrana	1131	66	113,58	-	0,58	2,77E-06
Plazma membrana	926	52	92,99	-	0,56	2,24E-05
Stanična unutrašnjost	2286	95	229,57	-	0,41	2,17E-38
Stanica	2287	95	229,67	-	0,41	1,91E-38
Stanične komponente	2340	96	234,99	-	0,41	8,18E-41
Unutrašnji dijelovi membrane	487	14	48,91	-	0,29	8,74E-08
Integralni dijelovi plazma membrane	319	9	32,03	-	0,28	8,67E-05
Membranski dijelovi	532	15	53,42	-	0,28	8,21E-09
Unutrašnji dijelovi plazma membrane	335	9	33,64	-	0,27	2,41E-05
Integralni dijelovi membrane	458	12	45,99	-	0,26	6,21E-08
Dijelovi plazma membrane	409	10	41,07	-	0,24	2,16E-07

5.RASPRAVA

5.1. Filostratigrafija bakterije *Escherchia coli* sojeva K-12 i ISC11 na sveobuhvatnoj bazi

Sve dosadašnje filostratigrafske mapa izrađene su na eukariotima dok jedina mapa izrađena na prokariotskom organizmu je ona bakterije *Bacillus subtilis subsp. subtilis* str. 168 (Lenuzzi, 2014.). Jedino je moguće usporediti dobivene rezultate s rezultatima bakterije *Bacillus subtilis subsp. subtilis* str. 168 u kontekstu bakterija.

Prema filostratigrafskim mapama sojeva K-12 i ISC11 dobivenih na sveobuhvatnoj proteinskoj bazi eukariota, bakterija i arheja velika većina proteina je smještena u prvi filostratum, tj. određeno je da su nastali vrlo rano u evolucijskoj prošlosti točnije do nastanka prvog staničnog organizma. Bakterija *Bacillus subtilis subsp. subtilis* str. 168 ima 64,6% sveukupnih proteina u prvom filostratumu dok sojevi K-12 i ISC11 87,01% odnosno 82,31% sveukupnih proteina u prvom filostratumu. Iako je taj evolucijski period trajao vrlo dugo te i eukarioti imaju najviše proteina nastalih u tom periodu postotci dobiveni za sojeve *E.coli* su iznad očekivanih i dosada nije zabilježen toliko visok postotak proteina u najstarijem filostratumu. Razlog tako visokom postotku može biti veliki broj proteinskih sljedova različitih eukariotskih i arhejskih vrsta u odnosu na filostratume koji imaju manji broj proteinskih sljedova te time manju šansu za pronalazak homologa. Također ako je homolog pronađen u više filostratuma onda se smješta u onaj koji je najstariji od svih njih što pogoduje prvom filostratumu budući da je on uvijek najstariji. Toliko visok postotak proteinskih sljedova u najstarijem filostratumu ne slaže s činjenicom da je dvije trećine gena iz pangenoma *E.coli* podrijetlom iz drugih vrsta i da su se pojavili procesom horizontalnog prijenosa gena (Zhaxybayeva, Olga et al., 2011). Geni koji se prenose horizontalnim prijenosom između bakterija su oni koji omogućuju bakteriji preživljavanje u promjenjivom okolišu te su to evolucijski mlađi geni (Heuer and Smalla, 2007.) no u dobivenim filostratigrafskim mapama korištenjem sveobuhvatne baze broj proteina u mlađim filostratumima je puno manji nego što se očekivalo s obzirom na horizontalni prijenos. Jedini manji pikovi prisutni su 2.,5 i 9. filostratumu no oni se s obzirom na malu količinu signala ne mogu odrediti kao evolucijski važni događaji.

5.2. Provjera robusnosti filostratigrafske mape bakterije *Escherchia coli* sojeva K-12 i ISC11 na reduciranoj i nasumičnoj bazi

Rezultati stratifkacije ovisni su o proteomu pojedinog organizma, filogenetskom stablu tog organizma i veličini baze proteoma s kojima se odabrani proteom uspoređuje. Zbog dobivenih neočekivanih rezultata za oba korištena soja *E.coli* na sveobuhvatnoj bazi napravljena je provjera robusnosti korištenjem reducirane i nasumične baze.

Obje baze sadrže jednak broj predstavnika svakog filostratuma, a razlikuju se po tome što se u reduciranoj bazi nalaze predstavnici odabrani po kriteriju filogenetske pokrivenosti, a u nasumičnoj bazi po kriteriju slučajnosti. Iz dobivenih rezultata može se vidjeti da se postotak proteinskih sljedova u prvom filostratumu drastično smanjio korištenjem obje baze za oba soja. Veliki broj proteinskih sljedova iz prvog filostratuma svebuhvatne baze je očito utjecao na dobivanje tako visokog postotka proteinskih sljedova u tom filostratumu. Smanjenjem broja proteina u prvom filostratumu pojavljuju se jasni pikovi u 6.,7. i 9. filostratumu u oba soja dok se u soju *E.coli* ISC11 pojavljuje još dodatni pik u posljednjem 12. filostratumu. Smatram da ovakva provjera robusnosti je korisna ako želimo saznati u kojim filostratumima možemo očekivati veći broj proteinskih sljedova, a u kojima manji. Međutim, ne bi se trebala koristiti za interpretaciju evolucijskih obrazaca na razini genoma jer ovisi o kriteriju odabira proteoma i o odabranom smanjenom broju te je moguće unošenje pristranosti svojim odabirom u rezultate. Korištenjem reducirane i nasumične baze dobiveni su vrlo slični rezultati no pitanje je bili bi se dobili slični rezultati da se koristila reducirana baza koja je sagrađena na temelju nekog drugog kriterija. Da bi se izbjegla pristranost i neovisnost o kriteriju odabira baze mogla bi se napraviti genomska filostratigrafija korištenjem velikog broja nasumičnih baza različitih veličina i obradom rezultata napraviti distribuciju koja najbolje opisuje sve rezultate. Takva distribucija bi se mogla koristiti u analizi evolucijskih obrazaca genoma. Još jedna dodatna prednost te metodem bila bi u tome što bi se izbjegle velike sveobuhvatne baze koje svojom veličinom otežavaju rad i analize. U ovom istraživanju to nije bilo moguće izvesti zbog tehničkih ograničenja poput nedostatka mjesta na serveru i klasteru no i nije bilo potrebe budući da je robusnost filostratigrafskih mapa provjerena na drugi način.

5.3. Provjera robusnosti filostratigrafske mape bakterije *Escherchia coli* sojeva K-12 i ISC11 promjenom broja hitova po filostratumu

Robusnost filostratigrafskih mapa sojeva K-12 i ISC11 provjerena je još uvođenjem uvjeta da broj hitova u određenom filostratumu mora biti iznad zadane granične vrijednosti ili u protivnom taj hit se uklanja iz rezultata dobivenih programom BLAST. Tako na primjer ako neki protein ima 2 hita s proteinskim sljedovima iz prvog filostratuma i tri hita s proteinima iz nekog mlađeg filostratuma, a zadana granična vrijednost je 2 onda će taj protein biti smješten u mlađi filostratum jer rezultat iz prvog filostratuma je uklonjen zbog toga što nije iznad zadane granične vrijednosti.

U ovom radu je korišteno sedam različitih graničnih vrijednosti te se pratio utjecaj povećanja granične vrijednosti na distribuciju proteinskih sljedova po filostratumima. Korištenjem najmanje granične vrijednosti primjećen je najveći pad postotka proteinskih sljedova u prvom filostratumu. Daljnjim porastom granične vrijednosti broj proteinskih sljedova u prvom filostratumu je linearno padao dok je u mlađim filostratumima zabilježen porast proteinskih sljedova. Najveći porast postotka proteinskih sljedova je zabilježen u 5.,6.,7., i 9., filostratumu te su u njima sada vidljivi jasni signali. Ti isti filostratumi su imali jasne pikove prilikom provjere robusnosti reduciranom i nasumičnom bazom što potvrđuje hipotezu da takva provjera robusnoti je korisna prilikom određivanja filostratuma u kojima se očekuje veći signal.

Korištenjem granične vrijednosti 9 kod soja K-12 došlo se do postotka proteinskih sljedova u prvom filostratumu jednakom onome koju bakterija *Bacillus subtilis subsp. subtilis* str. 168 posjeduje dok je za soj ISC11već granična vrijednost 5 dovela do pojave tog postotka proteinskih sljedova što je logično s obzirom da je soj ISC11 već u početku je imao manje proteinskih sljedova u prvom filostratuma. U oba soja povećanjem kriterija granične vrijednosti pojačavaju se signali mlađim filostratumima posebice se vide jasni pikovi u 5. i 9. filostratumu koji su bili vrlo slabi signali u prvotnim filostratigrafskim mapama ovih sojeva. Peti filostratum predstavlja vrijeme nakon odvajanja koljena Synergistetes i nastanak brojnih bakterijskih koljena iz Grupe C. Deveti filostratum predstavlja period nakon odvajanja ostalih članova razreda Gammaproteobacteria i nastanak reda Enterobacteriaceae. Signal u devetom filostratumu je zanimljiv jer postoje podaci da od tog perioda u genom *E.coli* K-12 je unešeno 755 protein-kodirajućih gena putem horizontalnog prijenosa gena

(Lawrence and Ochman, 1998) međutim ovdje je utvrđena pojava tek 422 protein-kodirajuća gena korištenjem najstrože granične vrijednosti. Od tih 422 protein-kodirajuća gena ne može se tvrditi da su svi ušli u genom horzontalnim prijenosom gena budući da geni mogu nastati mutacijama i duplikacijama. Oba soja u svim dobivenim filostratigrafskim mapama pokazuju vrlo sličnu distribuciju proteinskih sljedova što ne mora biti slučaj budući da sojevi *E.coli* mogu se razlikovati u 80% genoma. Sojevi korišteni u ovom radu razlikuju se i veličinom proteoma te ulogom tako da je K-12 najčešće korišteni laboratorijski soj, a soj ISC11 je klinički izolat nedavno izoliran i sekvenciran no unatoč tome imaju sličnu distribuciju proteinskih sljedova po filostratumima.

5.4. Analiza proteina iz pojedinih filostratuma prema rječniku Gene Ontology

Proteinski sljedovi identificirani kao najstariji i najmlađi u proteomu *E.coli* soja K-12 na sveobuhvatnoj, reduciranoj i nasumičnoj bazi te korištenjem granične vrijednosti 15 su analizirani s obzirom na ulogu u biološkim procesima, molekularnu funkciju i staničnu lokalizaciju.

Svi stariji proteini bez obzira na način identifikacije pokazaju jednaku ulogu u biološkim procesima, molekularnoj funkciji i staničnoj lokalizaciji. Tako svi stariji proteini imaju najveću ulogu u metaboličkim i staničnim procesima te lokalizaciji. Molekularna funkcija im je vezana uz katalitičku aktivnost, vezanje i transport. U stanici se najviše nalaze u njenoj unutrašnjosti i membrani. Važno je napomenuti da svi starij geni su bili dobro anotirani i za veliku većinu njih postoje dostupni podaci u bazam.

Analizom mlađih proteina otkriveno je da većina njih je u biološkim baza kategorizirana kao nekarakterizirani proteini i nisu postojali nikakvi podaci o njihovoj ulozi u biološkim procesima, molekularnoj funkciji i staničnoj lokalizaciji. Korištenjem Enrichment analysis alata utvrđeno je da mlađi proteini identificrani na sveobuhvatnoj bazi i korištenjem najviše granične vrijednosti su prezastupljeni u kategoriji neklasificiranih proteina s obzirom na ulogu u biološkim procesima, molekularnu funkciju i staničnu lokalizaciju. Mlađi proteini identificirani na reduciranoj i nasumičnoj

bazi uz prezastupljnost u kategoriji neklasificiranosti s obzirom na sva tri kriterija pokazuju prezastupljenost u ulozi stanične adhezije i stvaranja biofilma te lokalizacije na pilusu i staničnim projekcijama. S obzirom da su ti rezultati dobiveni na bazama koje su puno manje od sveobuhvatne i nisu identificirani u rezultatima dobivenim korištenjem najveće granične vrijednosti koja je najviše povećala broja mladih proteina te samim time i vjerojatnost pojave tih proteina smatram da prezastupljenost tih proteina treba uzeti s rezervom. Jedina prezastupljena kategorija koja se pojavila u svim mladim proteinima bez obzira na način njihove identifkacije je bila neklasificiranost te se time potvrdilo da su evolucijski mlađi proteini *E.coli* K-12 neistraženi i nekarakterizirani s obzirom na ulogu u biološkim procesima, molekularnu funkciju i staničnu lokaciju.

6.ZAKLJUČAK

U sklopu ovog rada napravljene su filostratigrafske mape bakterije *Escherichia coli* sojeva K-12 i ISC11 na sveobuhvatnoj, reduciranoj i nasumičnoj proteinskoj bazi. Robusnost filostratigrafskih mapa ispitana je uvođenjem uvjeta o graničnom broju hitova unutar filostratuma. Analizirana je molekularna funkcija, uloga u biološkim procesima i stanična lokalizacija proteina koji su određeni kao evolucijski stari i onih koji su određeni kao evolucijski mlađi. Nakon provedenih analiza, dobivenih rezultata i provedene rasprave mogu se donijeti sljedeći zaključci:

1. Nakon provedene filostratigrafske analize bakterije *Escherichia coli* sojeva K-12 i ISC11 na sveobuhvatnoj bazi u slučaju neočekivanih rezultata nužno je provjeriti robusnost dobivene mape

2. Provjera robusnosti filostratigrafske mape na reduciranoj i nasumičnoj mapi pouzdan je način određivanja onih filostratuma u kojima možemo očekivati veći broj proteinskih sljedova

3. Provjera robusnosti uvođenjem uvjeta o graničnom broju hitova u filostratumu uzrokuje najveće smanjenje proteinskih sljedova u prvom filostratumu te povećanje signala u mlađim filostratumima bakterije *E.coli* sojeva K-12 i ISC11

4. U oba korištena soja bakterije *Escherichia coli* mogu se utvrditi tri jasna makroevolucijska događaja u kojima su nastali većina proteina

5. Analizom uloge u biološkim procesima, molekularne funkcije i stanične lokalizacije može se sa sigurnošću utvrditi da evolucijski stariji proteini bakterije *E.coli* K-12 imaju dobro proučenu i određenu ulogu u biološkim procesima, molekularnu funkciju i staničnu lokalizaciju. Evolucijski mladi proteini bakterije *E.coli* K-12 su u bazama podataka opisani kao nekarakterizirani i neklasificirani proteini te je njihova uloga u biološkim procesima, molekularna funkcija i stanična lokalizacija nepoznata.

7. ZAHVALE

Zahvaljujem se svojim mentorima, izv. prof. dr. sc. Tomislavu Domazetu-Loši i izv. prof. dr. sc. Damjanu Franjeviću na pruženoj prilici izrade ovog rada te na stručnoj pomoći i vodstvu.

Zahvaljujem se dr.sc. Martinu Sebastijanu Šestaku na rješavanju praktičnih problema, korisnim savjetima i nesebičnom dijeljenju iznimno bogatog znanja.

Zahvaljuem se svojoj obitelji i prijateljima na pruženoj podršci i ljubavi.

8. LITERATURA

Acinas, S. G. et al. "Divergence And Redundancy Of 16S Rrna Sequences In Genomes With Multiple Rrn Operons". *Journal of Bacteriology* 186.9 (2004): 2629-2635.

Achtman, M. & Wagner, M. "Microbial diversity and the genetic nature of microbial species". *Nature Rev. Microbiol.* 6(2008): 431–440.

Bachmann, B. J. "Pedigrees of some mutant strains of Escherichia coli K-12". *Bacteriological reviews* 36.4(1972): 525–557.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF: "A kingdom level phylogeny of eukaryotes based on combined protein data". *Science* 290(2000):9 72-977.

Barion, Sacha et al. "The First Lines Of Divergence In The Bacteria Domain Were The Hyperthermophilic Organisms, The Thermotogales And The Aquificales, And Not The Mesophilic Planctomycetales". *Biosystems* 87.1 (2007): 13–19.

Beiko, R. G., Doolittle, W. F. & Charlebois, R. L. "The impact of reticulate evolution on genome phylogeny". *Syst. Biol*. 57(2008): 844–856.

Blattner, F. R. et al. "The Complete Genome Sequence Of Escherichia Coli K-12". *Science* 277.5331 (1997): 1453-1462.

Boussau, Bastien et al. "Parallel Adaptations To High Temperatures In The Archaean Eon". *Nature* 456.7224 (2008): 942-945.

Brocchieri L. "Phylogenetic inferences from molecular sequences: review and critique". *Theor Popul Biol* 59 (2001): 27-40.

Brochier, Céline, and Hervé Philippe. "Phylogeny: A Non-Hyperthermophilic Ancestor For Bacteria". *Nature* 417.6886 (2002): 244-244.

Di Giulio, Massimo. "The Ancestor Of The Bacteria Domain Was A Hyperthermophile". *Journal of Theoretical Biology* 224.3 (2003): 277-283.

Domazet-Lošo, T., and D. Tautz. "An Evolutionary Analysis Of Orphan Genes In Drosophila". *Genome Research* 13.10 (2003): 2213-2219.

Domazet-Lošo, Tomislav, Josip Brajković, and Diethard Tautz. "A Phylostratigraphy Approach To Uncover The Genomic History Of Major Adaptations In Metazoan Lineages". *Trends in Genetics* 23.11 (2007): 533-539.

Domazet-Lošo, T., and D. Tautz. "An Ancient Evolutionary Origin Of Genes Associated With Human Genetic Diseases". *Molecular Biology and Evolution* 25.12 (2008): 2699-2707

Domazet-Lošo, Tomislav, and Diethard Tautz. "A Phylogenetically Based Transcriptome Age Index Mirrors Ontogenetic Divergence Patterns". *Nature* 468.7325 (2010a): 815-818.

Domazet-Lošo, Tomislav, and Diethard Tautz. "Phylostratigraphic Tracking Of Cancer Genes Suggests A Link To The Emergence Of Multicellularity In Metazoa". *BMC Biology* 8.1 (2010b): 66.

Doolittle, W. F. "Phylogenetic classification and the universal tree". *Science* 284 (1999),:2124–2128.

Eardly, B. D. et al. "Rhizobial 16S Rrna And Dnak Genes: Mosaicism And The Uncertain Phylogenetic Placement Of Rhizobium Galegae". *Applied and Environmental Microbiology* 71.3 (2005): 1328-1335.

Eisen, J. A. "Assessing evolutionary relationships among microbes from wholegenome analysis". *Curr. Opin. Microbiol*. 3(2000):475–480.

Foster PG, Hickey DA "Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions". *J Mol Evol* 48 (1999): 284-290.

Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., and Woese, C. R. "The phylogeny of prokaryotes", *Science* 209 (1980): 457–463.

Fuerst, J. A. "The Planctomycetes: Emerging Models For Microbial Ecology, Evolution And Cell Biology". *Microbiology* 141.7 (1995): 1493-1506.

Galtier, N., N. Tourasse, and M. Gouy. "A Nonhyperthermophilic Common Ancestor To Extant Life Forms". *Science* 283.5399 (1999): 220-221.

Gogarten-Boekels, M., Hilario, E., and Gogarten, J. P. "The effects of heavy meteorite bombardment on the early evolution–The emergence of the three domains of life". *Orig. Life Evol.Biosph*. 25 (1995): 251–264.

Gupta, R. S. "Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes " . *Microbiol. Mol. Biol. Rev.* 62 (1998): 1435–1491.

Gupta, R. S. "The phylogeny of Proteobacteria: relationships to other eubacterial phyla and eukaryotes". *FEMS Microbiol. Rev*. 24 (2000):367–402.

Gupta, Radhey S., and Emma Griffiths. "Critical Issues In Bacterial Phylogeny". *Theoretical Population Biology* 61.4 (2002): 423-434.

Hasegawa, Masami, and Tetsuo Hashimoto. "Ribosomal RNA Trees Misleading?". *Nature* 361.6407 (1993): 23-23.

Heure, H. and Smalla K. "Horizontal gene transfer between bacteria". *Environ. Biosafety Res.* 6 (2007): 3-13
Hong Nhung, Pham et al. "Phylogeny And Species Identification Of The Family Enterobacteriaceae Based On Dnaj Sequences". *Diagnostic Microbiology and Infectious Disease* 58.2 (2007): 153-161.

Hugenholtz, P. "Exploring prokaryotic diversity in the genomic era" . *Genome Biol.* 3, REVIEWS0003.1–REVIEWS0003.8 (2002).

Imai, E., Honda, H., Hatori, K., Brack, A., and Matsuno, K. "Elongation of oligopeptides in a simulated submarine hydrothermal system". *Science* 283 (1999): 831–833.

Jain, R., M. C. Rivera, and J. A. Lake. "Horizontal Gene Transfer Among Genomes: The Complexity Hypothesis". *Proceedings of the National Academy of Sciences* 96.7 (1999): 3801-3806.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H: "Phylogenomics: the beginning of incongruence? ". *Trends Genet* 22(2006): 225-231.

Kaminuma, Eli et al. "DDBJ Launches A New Archive Database With Analytical Tools For Next-Generation Sequence Data". *Nucleic Acids Research* 38.suppl_1 (2009): D33-D38.

Khalturin, Konstantin et al. "More Than Just Orphans: Are Taxonomically-Restricted Genes Important In Evolution?". *Trends in Genetics* 25.9 (2009): 404-413.

Kitahara, Kei, and Kentaro Miyazaki. "Revisiting Bacterial Phylogeny". *Mobile Genetic Elements* 3.1 (2013): e24210.

Lake, J. A. "The order of sequence alignment can bias the selection of tree topology ". *Mol. Biol. Evol.* 8 (1991): 378–385.

Land, Miriam et al. "Insights From 20 Years Of Bacterial Genome Sequencing". *Functional & Integrative Genomics* 15.2 (2015): 141-161.

Lawrence, J. G., and H. Ochman. "Molecular Archaeology Of The Escherichia Coli Genome". *Proceedings of the National Academy of Sciences* 95.16 (1998): 9413-9417.

Lederber, J. "E. coli K-12.". *Microbiology today* (2004)31:116.

Lenuzzi, M. "Filostratigrafska analiza bakterije *Bacillus subtilis*" Pohranjeno u repozitoriju Prirodoslovno-matematičkog fakulteta u Zagrebu(2014)

Levine, Norman D. "Buchanan, R. E. & Gibbons, N. E., Eds. 1974.Bergey's Manual Of Determinative Bacteriology.8Th Ed. Williams & Wilkins Co., Baltimore, Md. 21202. Xxvi + 1246 Pp. $45.00". *The Journal of Protozoology* 22.1 (1975): 7-7.

Lim, K., Y. Furuta, and I. Kobayashi. "Large Variations In Bacterial Ribosomal RNA Genes". *Molecular Biology and Evolution* 29.10 (2012): 2937-2948.

Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW. "Substitutional bias confounds inference of cyanelle origins from sequence data". *J Mol Evol* 34 (1992):153-162.

Loomis WF and Smith DW. "Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison". *Proc Natl Acad Sci USA* 87 (1990):9093-9097.

Ludwig, W., and Schleifer, K. H. "Phylogeny of Bacteria beyond the 16S rRNA standard". *ASM News* 65 (1999): 752–757.

Ludwig, W., and Klenk, H.-P. 2001. Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systematics. in ‘‘Bergey’s Manual of Systematic Bacteriology’’ (D. R. Boone, and R. W. Castenholz, Eds.), pp. 49–65, Springer-Verlag, Berlin.

Miller, S. R. et al. "Discovery Of A Free-Living Chlorophyll D-Producing Cyanobacterium With A Hybrid Proteobacterial/Cyanobacterial Small-Subunit Rrna Gene". *Proceedings of the National Academy of Sciences* 102.3 (2005): 850-855.

Mount D. W. (2004): *Bioinformatics: Sequence and Genome Anaylsis* (2nd Edition). Cold Spring Harbor Press.

Neme, Rafik, and Diethard Tautz. "Phylogenetic Patterns Of Emergence Of New Genes Support A Model Of Frequent De Novo Evolution". *BMC Genomics* 14.1 (2013): 117.

Nisbet, E. G., and Sleep, N. H. 2001. "The habitat and nature of early life", *Nature* 409 (2001): 1083–1091.

Olsen, G. J., Woese, C. R., and Overbeek, R. "The winds of (evolutionary) change: Breathing new life into microbiology", *J. Bacteriol*. 176 (1994), 1–6.

Pace, N. R. "Origin of life–facing up to the physical setting". Cell 65(1991): 531–533.

Paradis, S. et al. "Phylogeny Of The Enterobacteriaceae Based On Genes Encoding Elongation Factor Tu And F-Atpase -Subunit". *International Journal of Systematic and Evolutionary Microbiology* 55.5 (2005): 2013-2025.

Pardi, Fabio, and Nick Goldman. "Resource-Aware Taxon Selection For Maximizing Phylogenetic Diversity". *Systematic Biology* 56.3 (2007): 431-444.

Patwardhan A, Ray S, Roy A "Molecular Markers in Phylogenetic Studies – A Review". *J. Phylogen Evolution Biol* 2 (2014):131.

Rappé, Michael S., and Stephen J. Giovannoni. "The Uncultured Microbial Majority". *Annual Review of Microbiology* 57.1 (2003): 369-394.

Schouls, L. M., C. S. Schot, and J. A. Jacobs. "Horizontal Transfer Of Segments Of The 16S Rrna Genes Between Species Of The Streptococcus Anginosus Group". *Journal of Bacteriology* 185.24 (2003): 7241-7246.

Singleton P (1999). *Bacteria in Biology, Biotechnology and Medicine* (5th ed.). Wiley. pp. 444–454. ISBN 0-471-98880-4.

Stetter K. O. "Ciba Found. Symp". 202 (1996):1-10.

Šestak, Martin Sebastijan et al. "Phylostratigraphic Profiles Reveal A Deep Evolutionary History Of The Vertebrate Head Sensory Systems". *Frontiers in Zoology* 10.1 (2013): 18.

Šestak, Martin Sebastijan, and Tomislav Domazet-Lošo. "Phylostratigraphic Profiles In Zebrafish Uncover Chordate Origins Of The Vertebrate Brain". *Molecular Biology and Evolution* 32.2 (2014): 299-312.

Tatum E. L.; Lederberg J. "Gene recombination in the bacterium Escherichia coli". *J. Bacteriol*. 53 (1974): 673–684.

Tenaillon, Olivier et al. "The Population Genetics Of Commensal Escherichia Coli". *Nature Reviews Microbiology* 8.3 (2010): 207-217.

Wang, Yue, and Zhenshui Zhang. "Comparative Sequence Analyses Reveal Frequent Occurrence Of Short Segments Containing An Abnormally High Number Of Non-Random Base Variations In Bacterial Rrna Genes". *Microbiology* 146.11 (2000): 2845-2854.

Williams, K. P. et al. "Phylogeny Of Gammaproteobacteria". *Journal of Bacteriology* 192.9 (2010): 2305-2314.

Woese, C. R., Stackebrandt, E., Macke. R. J., and Fox, G. E. "A phylogenetic definition of the major eubacterial taxa", *Syst. Appl. Microbiol*. 6 (1985): 143–151.

Woese, C. R. "Bacterial evolution ". *Microbiol. Rev*. 51 (1987): 221–271.

Wu, M. & Eisen, J. A. "A simple, fast, and accurate method of phylogenomic inference". *Genome Biol*. 9 (2008): R151

Wu, D. et al. Complete genome sequence of the aerobic CO-oxidizing thermophile Thermomicrobium roseum. PLoS One 4, e4207 (2009).

Wu, Dongying et al. "A Phylogeny-Driven Genomic Encyclopaedia Of Bacteria And Archaea". *Nature* 462.7276 (2009): 1056-1060.

Zhaxybayeva, Olga et al. "Lateral Gene Transfer ". Current Biology 21.7(2011): 242 - 246

Zuckerkandl, Emile, and Linus Pauling. "Molecules As Documents Of Evolutionary History". *Journal of Theoretical Biology* 8.2 (1965): 357-366.

9.SAŽETAK

Hrvoje Mišetić

Filostratigrafska analiza bakterije *Escherichia coli*

Bakterijska filogenija je predmet rasprava između znanstvenika još od njihova otkrića te i dalje ne postoji robusno filogenetsko stablo koje prikazuje evolucijske odnose bakterija. Uzrok tome su brojni problemi poput horizontalnog prijenosa gena, brza stopa mutacija bakterijskog genoma i niz pretpostavki te parametara koji utječu na grananje filogenetskog stabla. Genomska filostratigrafija je analitička metoda koja proučava makroevolucijske trendove u evoluciji određene vrste. Filostratigrafske mape su dosada uglavnom rađene samo na eukariotima. U ovom radu je izrađena filostratigrafska mapa bakterije *Escherichia coli* sojeva K-12 i ISC11. Robusnost mape je testirana na dva način: korištenjem reducirane i nasumične proteinske baze te uvođenjem uvjeta o graničnom broju hitova po filostratumu. Prvim načinom se distribucija proteinskih sljedova po filostratumima za oba soja promijenila te su dobiveni jasni signali u pojedinim filostratumima. Varijacijom granične vrijednosti o broju hitova po filostratumu za rezultate dobivene na sveobuhvatnoj bazi uočeno je da porastom granične vrijednosti se smanjuje broj proteinskih sljedova u prvom filostratumu, a povećava broj proteina u mlađim filostratumima posebno u onim koji su imali jasan signal pri testiranju robusnosti na reduciranoj i nasumičnoj proteinskoj bazi. Testiranjem robusnosti filostratigrafske mape oba soja *E.coli* utvrđena su tri ključna makroevolucijska događaja pri kojem su nastali većina proteinskih sljedova. Nadalje, analizirana je uloga u biološkim procesima, molekularna funkcija i stanična lokalizacija evolucijski starih i evolucijski mladih proteina koji su identificirani u filostratigrafskim mapa *E.coli* K-12 dobivenim korištenjem sveobuhvatne, reducirane i nasumične baze te najveće granične vrijednosti. Ustanovljeno je da većina starih proteina ima ulogu u staničnim i metaboličkim procesima, s obzirom na molekularnu funkciju najviše pokazuju katalitičku i transportnu aktivnost te ulogu u vezivanju, a lokalizirani su ponajviše u staničnoj unutrašnjosti i na membrani. Evolucijski mladi proteini su nekarakterizirani i neklasificirani te je njihova uloga u biološkim procesima, molekularna funkcija te stanična lokalizacija nepoznata. Ovo je jedna od prvih

filostratigrafskih analiza napravljenih na prokariotskom organizmu i to na onome koji je najkorišteniji prokariotski model u molekluarnoj biologiji te prva analiza koja uključuje testiranje uvođenjem uvjeta o grančnom broju hitova po filostratumu i analizu uloge u biološkim procesima, molekularne funkcije i stanične lokalizacije identificiranih evolucijski starih i mladih proteina iz proteoma bakterije *Escherichia coli* K-12.

Ključne riječi: filostratigrafija, filogenija,biološke baze podataka, *Escherichia coli*

10.SUMMARY

Hrvoje Mišetić

Phylostratigraphic analyses of *Escherichia coli*

Bacterial phylogeny has been the subject of numerous discussions in the scientific community since the time of their discovery, but there is still no robust phylogenetic tree describing the evolutionary relationships among all bacterial phyla. There are many reasons why bacterial phylogeny is a complex issue, such as lateral gene transfer, high mutation rate in bacterial genomes, and numerous variables that have a different impact on the branching order of bacteria. Genomic phylostratigraphy is an analytical method used for discovering macroevolutionary trends in the evolution of a species. Thus far, it has mainly been applied to eukaryotes . In this paper, the phylostratigraphic map of bacterium *Escherichia coli* has been created for strains K-12 and ISC11. The quality of the maps was tested using two different approaches. The first approach was to use a reduced and a random protein database in order to construct the phylostratigraphic map. This caused a different distribution of proteins among phylostratums for both strains and also showed significant signals in certain phylostratums. The second approach was to introduce a condition according to which there has to be a certain number of hits per phylostratum before a protein can be assigned it. By applying multiple sets of cut-off values for hits per phylostratum while utilizing the complete protein database, it was detected that an increase of the cut-off value causes a decrease in the number of proteins in the first phylostratum and an increases in the number of proteins in later phylostratums, especially in the ones that contained strong signals while testing the results on the reduced and random protein databases. Based on these tests of the

phylostratigraphic maps for both strains, three main macroevolutionary events were identified as moments when the majority of proteins appeared in the proteomes of the tested strains. Furthermore, Gene Ontology was used to analyze the molecular function, biological processes and cellular localization of the proteins that were identified as evolutionarily old or new in the proteome of *E.coli* K-12 strain. This was conducted based on results obtained using the complete, reduced and random protein databases and also on the results obtained using the highest cut-off score. It was concluded that older proteins mostly have roles in metabolic processes, while their molecular function relates to catalytic and transport activity along with a role in binding partner molecules. They are mostly localized in the cytoplasm and on the plasma membrane. Evolutionarily new proteins are mostly uncharacterized and unclassified so their roles in biological processes, as well as their molecular functions and cellular localization still remain unknown. This is one of the first phylostratigraphic analyses on the prokaryotic organism and the first analysis that includes testing based on the cut-off number of hits per phylostratum and also determinating the molecular function, role in the biological processes and cell localization of the evolutionary old and new proteins from the proteome of *Escherichia coli* K-12.

Keywords: phylostratigraphy, phylogeny, bioinformatic databases, *Escherichia coli*