TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIATEADUSKOND
Molekulaar- ja rakubioloogia instituut
Bioinformaatika õppetool

Aleksander Sudakov

Genoomi signatuuri meetodi rakendamine horisontaalse geeniülekande detekteerimiseks

Bakalaureusetöö

Juhendaja Tõnu Margus

TARTU
2006

Sisukord

Lühendid ja mõisted

16S rRNA	ribosoomi väikese subühiku rRNA sedimentatsiooni koefitsiendiga 16 Svedbergi.
5S rRNA	ribosoomi suure subühiku rRNA sedimentatsiooni koefitsiendiga 5 Svedbergi
BLAST	programm sarnasuse otsimiseks järjestuse andmebaasides (*Basic Local Alignment Search Tool*)
bp	aluspaar (*Base Pair*)
Bootstrap	fülogeneesipuu usaldusväärsuse hinne
CAI	koodoni adaptatsiooniindeks (*Codon Adaptation Index*)
ClustalW	järjestuste globaalse joondamise programm
Ef-Tu	elongatsiooni faktor Tu, translatsiooni mehhanismi valk (*elongation factor Tu*)
GC-sisaldus	genoomi parameeter, mis näitab GC aluspaaride osakaalu DNAs
GS	genoomi signatuur (*Genomic Signature*)
HGT	horisontaalne (lateraalne) geeniülekanne (*Horisontal Gene Transfer*)
kbp	10^3 aluspaari
nt	nukleotiid
ORF	avatud lugemisraam (*Open Reading Frame*)
Ortoloogid	eri liikidest pärit järjestused, millel on üks ühine eellane
Paraloogid	samast liigist pärit järjestused, millel on üks ühine eellane (tekivad duplitseerumise teel)
PDS	fülogeneetiliselt kokkusobimatu järjestus (*Phylogenetically Discordant Sequence*)
Perl	tekstilise informatsiooni töötlemiseks sobiv programmeerimiskeel (*Practical Extraction and Report Language*)
RSCU	Suhteline sünonüümsete koodonite kasutuskoefitsient (*Relative Synonymous Codon Usage*)
SSU	ribosoomi väike subühik (*Small Subunit*)

Sissejuhatus

Horisontaalselt ülekandunud geenide olemasolu bakterite genoomides on nüüdseks üldtuntud ja aktsepteeritud fakt. Geneetilise info vahetus erinevate liikide vahel omab olulist rolli liikide evolutsioonis. Kuna genoomide vahel toimuvad ülekande protsessid, siis ei ole õige esitada liikide põlvnemist ainult evolutsioonipuu kujul. Olulised liigi omadusi määravad rakumehhanismid võivad olla pärit teiselt liigilt ja mitte eellaselt. Selle taustal on tekkinud mitmed küsimused, näiteks "tuumik" geenide olemasolu, geeni ja genoomi fülogeneesipuude erinevus ning fülogeneesipuu mittevastavus reaalsele evolutsioonile.

Fülogeneesipuude analüüs on horisontaalse ülekande detekteerimisel kõige usaldusväärsem meetod. Fülogeneesipuude vaheline erinevus osutab kõrvalekalletele evolutsiooni normaalsest kulgemisest. Samas on fülogeneetilistel meetoditel omad probleemid, mis tulenevad joondamise ja puu konstrueerimise vigadest, arvutusmahukusest ning analüüsi automatiseerimise raskusest.

Horisontaalse geeniülekande detekteerimiseks on lisaks välja töötatud ja testitud mitmeid meetodeid, kuid nende täpsus ja usaldusväärsus on puudulikud. Alternatiivsete lähenemisviiside seast eristub positiivselt genoomi signatuuri meetod. See ei vaja suurt arvutusvõimsust, on kergelt kasutatav vabalt valitud järjestuste analüüsil ning on teiste meetoditega võrreldes efektiivsem.

1. Kirjanduse ülevaade

1.1 Horisontaalne geeniülekanne

1.1.1 Horisontaalse geeniülekande mõiste

Geneetiline informatsioon pärandub valdavalt emarakult tütarrakkudele ehk vertikaalselt. Kuid on võimalik ka muul viisil geneetilise materjali liikumine, mida nimetatakse horisontaalseks või lateraalseks geeniülekandeks (*horizontal gene transfer*; HGT). Horisontaalne geeniülekanne on protsess, milles geneetiline informatsioon liigub teise raku, mis ei ole otsene järeltulija, genoomi. On mitmeid viise kuidas võõrast rakust pärit geneetiline materjal pääseb rakku (Twyman, 1998):

- transformatsioon – bakterirakk võtab vaba DNA väliskeskkonnast;
- konjugatsioon – DNA (tavaliselt plasmiidi) vahetu ülekanne ühest rakust teise;
- transduktsioon – DNA ülekanne viiruste (bakteriofaagide) abil.

HGT on fenomen, mille ulatus ja olemasolu on pikka aega olnud vaidlusalune küsimus. HGT on ka raske üheselt tõestada (Daubin *et al*., 2003). Tänu bakterite (ning vähemal määral ka eukarüootide) võimele integreerida võõrast DNA-d enda genoomi, on paljud liigid evolutsioneerunud kiiremini ja omandanud võime asustada uusi ökoloogilisi nišše (Boucher *et al*., 2003). Juba enne täisgenoomide sekveneerimist räägiti üksikutest HGT juhtudest, aga valitseva arvamuse järgi oli fenomen liiga harvaesinev selleks, et mõjutada evolutsiooni pilti. Ainus sündmus, kus tunnistati HGT tähtsust, oli geeniülekanne endosümbiontsetest organellidest: paljud geenid liikusid mitokondri ja kloroplasti genoomidest eukarüootide tuuma genoomi (Woese, 1987).

Täielikult sekveneeritud bakterigenoomide analüüs näitas, et tegemist on liiga sagedase nähtusega, et pidada selle mõju tähtsusetuks. Esimene kinnitus HGT esinemisele tuli *Escherichia coli* koodonisageduste analüüsist genoomi fragmentides. Umbes 15% geenidest näitasid ülejäänud genoomist oluliselt erinevat koodonsagedust (Koonin *et al*., 2001). Mõned neist geenidest olid selgelt suguluses bakteriofaagi geenidega või transponeeruvate

elementidega, millest järeldadati taoliste geenide genoomi välist päritolu.

1.1.2 Horisontaalse geeniülekande uuringute ajalugu

Esimesed horisontaalse geeniülekande juhtumid bakterite genoomide vahel on teada juba suhteliselt pikka aega.

1943. a. Rockefeller'i Ülikoolis läbiviidud eksperimendis, mis tõestas DNA rolli päriliku info kandjana, demonstreeriti mikroobide võimet integreerida enda genoomi keskkonnast pärit DNA-d (Avery *et al*., 1944). Edaspidi kirjeldati sarnast võimet mitmel bakteriliigil, lisaks ka ülekannet bakteriofaagide ning plasmiidide abil.

Antibiootikumi resistentsuse ning patogeensuse tunnuste levikut plasmiidide jm mobiilsete elementide abil tunti juba aastakümneid tagasi, kuid tol ajal ei arvatud, et horisontaalse geeniülekande fenomen on laialdaselt levinud.

Esimesed molekulaarse fülogeneesi uurimised on tehtud Zukerandl ja Pauligi poolt. Nad näitasid, et liikide fülogeneesi rekonstrueerimiseks saab kasutada molekulaarseid järjestusi kui traditsioonilisi fenotüübi karakteristikuid ning saada evolutsiooniliselt õigemaid tulemusi. Nad tutvustasid järjestuste neutraalse evolutsiooni teooriat ja eristasid ortolooge ja paralooge (Zukerandl, Paulig, 1965).

Algselt kasutati molekulaarse fülogeneesi leidmiseks valgulisi järjestusi, eriti bakteri ferredoksiine ja tsütokroome. Valkude analüüsi abil seoti kloroplastide päritolu tsüanobakteritega ja mitokondrid – proteobakteritega. Eukarüootse raku tuumas on tsütokroomide ja ferredoksiinide homoloogid pärit organellide genoomist ning ei sisalda informatsiooni eukarüoodi evolutsioonist. Schwartz ja Dayhoff uurisid 5S rRNA molekule, et leida eukarüootse tuuma päritolu. Selle järjestused saadi mitmest organismist. Kahjuks on 5S rRNA liiga lühike, et anda piisavat fülogeneetilist informatsiooni: antud artiklis oli tulemuseks eukarüootse tuuma genoomi päritolu gram-positiivsetest bakteritest (Schwartz, Dayhoff, 1978).

Tsütokroomi fülogeneesi uurimisel selgus veel, et eri molekulidel põhinevatel puudel on erinev topoloogia, mis erines ka üldtunnustatud klassifikatsioonist. Ambler ja kaasautorid pakkusid 1979. a. mitut võimalikku põhjust (Ambler *et al*., 1979) ja järeldasid, et antud valkude jõudmine mitmesse organismi on võimalik vaid liikidevahelise geeniülekande vahendusel. Siiski arvati järgmisel aastakümnel (suurel määral Carl Woese tööde mõjul), et

valgu ebakorrapärase esinemise liikide lõikes tingib uuritava valgu geeni liigispetsiifiline deleteerumine genoomist ja mitte horisontaalne geeniülekanne (Doolittle *et al*., 2003).

Carl Woese pooldas 1970-ndatel ribosoomi väikese subühiku (SSU) rRNA järjestuse valimist üldiseks molekulaarseks kellaks (Woese, 1987). Sellel molekulil on homoloogid kõikides prokarüootide, mitokondrite, kloroplastide ja eukarüootide tuuma genoomides. Ribosomaalses RNA's on väga konserveerunud piirkondi, mis sobivad kaugete liikide sugulussidemete kirjeldamiseks, ning samas on seal ka varieeruvaid piirkondi, mis kannavad piisavalt signaali lähedaste liikide eristamiseks. Lisaks interakteerub rRNA paljude teiste molekulidega ning on väga konserveerunud.

Kasutades SSU rRNA põhjal arvutatud fülogeneesipuud on lihtne paigutada bakteriliigid taksonoomilistesse rühmadesse, mis üldjoontes vastavad Bergey's klassifikatsioonile. 16S rRNA abil avastati mõned uued rühmad, samas leiti, et mitmed eri taksonitest pärit liigid on tegelikult lähedased sugulased. Lisaks sai paigutada mitmed kahtluse all olevad organismid arhea taksonisse. Üldine fülogeneesipuu SSU rRNA põhjal on kolmeharulise tähe kujuline (*unresolved polychotomy*). See võib olla aga SSU rRNA vähese eraldusvõime tõendiks (Doolittle *et al*., 2003).

SSU rRNA andmete abil oli tõestatud ka korduva endosümbioosi hüpotees (*Serial Endosymbios Hypotyhesis*), ehk mitokondri ja kloroplasti pärinemine endosümbiontsetest bakteritest (Doolittle *et al*., 2003).

Yan Boucher kaasautoritega (2003) tegi ülevaate horisontaalse ülekande rollist metabolismisüsteemide levikus. Eelnevad uuringud näitavad muuhulgas, et HGT vastutab fotosüsteemide tüüp 1 ja 2 ühinemise eest tsüanobakterites varajases evolutsioonijärgus. Mitmel juhul tõestati fülogeneetiliste puude abil HGT rolli fotosünteesi, aeroobse hingamise, lämmastiku fikseerimise, sulfaadi redutseerimise, isoprenoidide biosünteesi, signaali ülekande, gaasivesiikuli abil pinnal hõljumise ja muude mehhanismide levikus eubakterite ja arheate lõikes (Boucher *et al*., 2003).

1.1.3 Horisontaalse geeniülekande uurimise hetkeseis ja ulatus

Tänu viimasel ajal saadud tõendustele horisontaalse geeniülekande levikust, on huvi HGT fenomeni uurimise vastu oluliselt kasvanud. Lisaks eraldiseisvatele uurimistele on loodud ka terveid bakterigenoome hõlmavad andmebaasid.

Garcia-Vallvé poolt tehtud ennustused, mis põhinevad mitme parameetri hindamisel nagu G+C sisaldus, koodon- ja aminohapete kasutus on toodud andmebaasis Horizontal Gene Transfer Database (http://www.fut.es/~debb/HGT/). See sisaldab ka informatsiooni geenide kohta, mis erinevad nendes parameetrites genoomi keskmisest – tõenäolise horisontaalse geeniülekande tagajärg; kõrgelt ekspresseeruvad geenid on HGT hinnangust välja jäetud. Andmebaas sisaldab otsingumootorit ja organismide nimekirja ning võimalust pääseda ligi iga geeni statistilistele parameetritele 94 eubakteri ja arhea genoomi kohta (Garcia-Vallvé, 2003). Anomaalsete geenide detekteerimise protseduur põhineb Garcia-Vallvé *et al*. (2000) artiklil.

A. Tsirigos ja I. Rigoutsos (2005) andmebaas sisaldab 123 prokarüootse liigi genoomi horisontaalselt ülekandunud geenide andmeid (http://cbcsrv.watson.ibm.com/HGT/). Kasutatud on genoomi signatuuril põhinevat parameetrilist meetodit.

Viimaste aastate kirjanduses on palju diskuteeritud nn. *core* ehk tuumik geenide üle (Doolittle *et al*., 2003). Just nende seast otsitakse geene, milles ei ole toimunud HGT. Arvatakse, et infotöötlemise mehhanismide geenid (replikatsioon, transkriptsioon ja translatsioon) on vähem vastuvõtlikud horisontaalsele ülekandele; metabolismi ja struktuurgeenid – rohkem (Nelson *et al*., 1999; Brochier *et al*., 2002; Boucher *et al*., 2003). Osade tööde tulemused näitavad aga, et selline tuum leidub ainult lähedastel liikidel (Nesbø *et al*., 2001). *Super-tree* lähenemine seob mitme geeni või valgu fülogeneesipuud liikide fülogeneesipuu saamise eesmärgil. See jätab välja teistest erineva fülogeneesiga geenid ning arvestab vaid sama topoloogiat toetavate geenidega (Calteau *et al*., 2004, Cicarelli *et al*., 2006, Daubin *et al*., 2002). Sellised geenid moodustavad teatud määral tuumiku, kuigi osa neist ei levi kõigis organismides.

1.2 Horisontaalse geeniülekande detekteerimise meetodid

1.2.1 Meetodite klassifikatsioon

HGT detekteerimise meetodid võib jagada mitmeks klassiks:

- **Parameetrilised meetodid** põhinevad eraldi võetud liigi genoomi järjestuse omaduste analüüsil. Meetodid identifitseerivad geene, mis on atüüpilised oma järjestuse omadustelt ülejäänud genoomi suhtes. Parameetrite hulka kuulub näiteks GC-osakaal,

nukleotiidi kasutus igas koodoni positsioonis, aminohappe valik, geeni ehitus. Eeldatakse, et geenid, mis on evolutsioneerunud ühe genoomi piires, omavad ühtlast väärtust antud parameetrites (Sandberg *et al*., 2003). Atüüpilised geenid on tõenäoliselt evolutsioneerunud teise liigi genoomis ning on tekkinud HGT tulemusena. Meetodite eelis on võimalus analüüsida suurt hulka genoom nõudmata suurt arvutusvõimsust. Sellistel meetoditel on suur valepositiivsete ja valenegatiivsete tulemuste protsent (Koski *et al*., 2001), leitud geenid tuleb enne horisontaalse ülekande järeldamist analüüsida fülogeneetiliste meetodite abil.

- **Fülogeneetilised meetodid** põhinevad fülogeneetiliste puude ehitamisel ning nende võrdlemisel. Erinevate geenide ortoloogidel põhinevate evolutsioonipuude võrdlemine on peamine viis tõestada horisontaalset ülekannet. On mitmeid puude ehitamise meetodeid, mis erinevad omavahel saadud tulemuste täpsuselt ja arvutusmahu poolest. Tänapäeval on rRNA evolutsioonipuu võetud etaloniks, millega võrrelda teisi organismi geene. Fülogeneesipuude võrdlemine lubab leida ka varajases evolutsioonis toimunud horisontaalülekande, aga ka duplikatsiooni ja deletsiooni sündmuseid. Meetodite puuduseks on arvutustöö mahukus, käsitsi võrdlemise vajadus ning kaugete liikide vaheliste puude ebatäpsus.

- **Füleetiline kokkusobimatus**. Andmebaasi otsingu põhjal identifitseeritakse nn. füleetiliselt kokkusobimatud järjestused, millel on erinev leviku muster liikide lõikes. Füleetilise jaotuse erinevused võivad olla tingitud horisontaalselt ülekandunud geenist, aga ka liigispetsiifilisest geeni deletsioonist ja duplikatsioonist. Seda meetodit kasutatakse fülogeneesipuude täpsuse suurendamiseks eemaldatates kokkusobimatud järjestused fülogeneesipuu analüüsist.

- **Paradoksaalne parim sarnasus (*paradoxical best hit*)**. Meetod põhineb samuti andmebaasi otsingul. Horisontaalset ülekannet võib oletada, kui geeni kõige sarnasem ortoloog on pärit evolutsiooniliselt kaugest liigist. Ka selle meetodi puhul võib valepositiivse tulemuse anda liigispetsiifiline deletsioon.

- **Geenide sisalduse ja järjekorra analüüsil põhinevad meetodid. E**ristatakse meetodeid, mis analüüsivad geenikoostist ning geenide järjekorda. Peale väheste erandite on nende rakendus horisontaalse geeniülekande detekteerimisel väike. Näiteks võib kolme või enama geeni samas järjekorras esinemine kaugetes liikides suure tõenäosusega

tähendada horisontaalselt ülekandunud operoni (Koonin *et al*., 2001; Omelchenko *et al*., 2003). Geenikomplekti võrdlus ning geenide järjekorra konserveerumise võrdlus on kasutust leidnud lähedaste liikide puhul (Koski *et al*., 2001).

1.2.2 Parameetrilised meetodid

Bakteri genoomi alad võivad ülejäänud genoomist erineda rea tunnuste poolest. Nendeks võib olla aluseline koostis, puriinide ja pürimidiinide suhe, oligonukleotiidide sagedus või koodonkasutus (Ragan, 2001a). On leitud, et horisontaalselt ülekandunud aladel on sageli hinnatavate parameetrite osas genoomi keskmisest statistiliselt usaldusväärne erinevus. Samas võivad ka kõrge ekspressiooni tasemega geenid ja ebatavalise funktsiooniga geenid ületada kõrvalekalde piiri hinnatava parameetriosas, ning klassifitseeruda ekslikult horisontaalselt ülekandunuks.

Nn. surrogaatsed HGT detekteerimise meetodid ei vaja fülogeneetiliste puude ehitamist. Antud lähenemine vähendab arvutusliku töö mahtu ja lubab vaadelda geene, millel ei ole leitud ortolooge. Oma töös võrdles Mark A. Ragan (2001b) nelja sellist meetodit HGT kandidaatide detekteerimiseks ORFide leidmiseks E. coli genoomist. Lisaks on Ragan'i sõnul veel vähemalt kolm taolist meetodit, aga nad on rakendatavad ainult väikestele regioonidele (nt 50 kb lõikudele):

• anomaalsused nukleotiidide tasakaalus või koodonkasutuses (*compositional difference*);

• Markovi mudel atüüpiliste geenide leidmiseks;

• fülogeneetiliselt kokkusobimatud järjestused, mille homoloogide fülogeneetiline kuuluvus erineb ülejäänud genoomi järjestuse homoloogide omast;

• geenide ebatavaline jaotumine eri organismides (*distribution profile*).

Eri meetodite abil leitud ORFide arv erineb oluliselt (5%-10% genoomist), sõltuvalt kasutatud künnise (*threshold*) väärtustest. Autor järeldab, et selliste meetodite tulemuste ühisosa ei peegelda horisontaalselt ülekandunud geene (Ragan, 2001b).

Garcia-Vallvé kaasautoritega (2000) analüüsisid 24 täisgenoomi mitme parameetrilise meetodiga. Igale genoomile arvutati keskmine väärtus ning igale geenile standardhälbe (σ) väärtus järgmistes parameetrites: koodonkasutus, suhteline sünonüümsete koodonite kasutus (*relative synonymous codon usage*, RSCU), üldine G+C sisaldus, G+C sisaldus igas

koodonipositsioonis (G+C[1], G+C[2], G+C[3]), aminohappeline koostis. Atüüpiliseks nimetati sellised geenid, mille parameetrid erinesid genoomi keskmisest väärtusest rohkem kui 1,5σ võrra. Autorid said tulemuseks 24 genoomis HGT osakaaluks 1,56% *Borrelia burgdorferi* ning kuni 14,47% *Bacillus subtilis* ja *Mycoplasma genitalium* puhul. Tulemused paigutati andmebaasi aadressil http://www.fut.es/~debb/HGT/ ning täiendati hiljem uute genoomidega sama metoodi järgi (Garcia-Vallvé, 2003).

1.2.3 Fülogeneetilised meetodid

Fülogeneetilised meetodid kasutavad fülogeneesipuude võrdlemist ja ennustavad nende erinevuste põhjal võimalikku horisontaalset ülekannet. Selle tulemused on suuresti sõltuvuses õigete fülogeneesipuude ehitamisest.

Evolutsiooniliste puude ehitamiseks on kolm lähenemist: *maximum parsimony*, *maximum likelihood* ning *distance* meetodid (Nei, 1996).

- *Distance* puhul arvutatakse evolutsiooniline kaugus kõikide järjestuste paaridele ning evolutsioonipuu ehitatakse paarikaupa kaugustest, kasutades *least-squares*, *minimum evolution*, *four-cluster analysis* või *neighbor-joining* meetodit. Kaugus väljendatakse tavaliselt nukleotiidide või aminohapete asenduste arvuga saidi kohta.
- *Maximum likelihood* meetod on täpsem, kuid aeglasem. Hinnatakse nukleotiidide või aminohapete asenduse tõenäosust. Üle kümne järjestuse analüüsil muutub arvutamine väga aeglaseks (arvutusmahukus kasvab logaritmiliselt).
- *Maximum parsimony* arvutab hüpoteetilised eellasjärjestused ning valib sellise puu, milles on toimunud minimaalne nukleotiidide või aminohapete asenduste arv igas harus. See meetod kasutab ka insertsioonide ja deletsioonide andmeid ning annab reeglina õige parema puu topoloogia. Tulemust võib negatiivselt mõjutada korduvate asenduste toimumine samas positsioonis, eriti nukleotiididse järjestuse puhul.

Fülogeneetiliste meetodite abil on tehtud mitmeid HGT uuringuid ka enne täisgenoomide sekveneerimist. Tänapäeval on saadaolevad arvutusvõimsused kasvanud ning see võimaldab läbi viia suuremahulist fülogeneetilist analüüsi mitme liigi ja geeni lõikes.

Carl Woese ja kaasautorid (2000) analüüsisid aminoatsüül-tRNA süntetaaside fülogeneesi ning järeldasid paljudel neist horisontaalse geeniülekande toimumist. *Asp*, *Glu*, *Phe*, *Leu*, *Tyr* ja *Trp* aminohapete tRNA süntetaasid toetavad üldist evolutsioonipuud; *Ile*,

His, *Pro* ja *Met* süntetaaside puhul on osadel eubakteritel kasutuses arhealt pärit ensüümi versioon; *Val*, *Arg*, *Thr*, *Ala* ja *Asn* on tugevasti vastuolus kanoonilise evolutsioonipuuga ning eukarüootide ensüümid on selgelt bakteriaalse päritoluga. Ülejäänud 6 ensüümi: *Cys*, *Ser*, *Lys*, *Gly*, *Asn* ja *Gln* ei vasta kanoonilisele puule üldse (Woese *et al*., 2000). Leitud HGT sündmused toimusid kaugete liikide vahel varajases evolutsiooni staadiumis.

V. Daubin ja kaasautorid (2003) otsisid seost DNA omandamise ja fülogeneetilise puude sobimatuse vahel. Kasutatakse korraga nelja suguluses oleva liigi genoomi järjestust. Kõigi liikide jaoks on HGT toetavate geenide arv väike (tüüpiliselt 0-3%). Suhteliselt suur on see arv E. coli (6,7%) ja Chlamydophila pneumoniae (8,2%) liikide siseselt. Selline tulemus näitab ulatusliku homoloogse rekombinatsiooni esinemist nende liikide lähisugulastega. Lisaks järeldatakse, et fülogeneesipuude ehitamisel sageli kasutatav likelihood-mapping meetod ei anna õiget hinnangut HGT kohta kaugete liikide vahel (Daubin *et al*., 2003).

Novichkov kaasautoritega (2004) pakuvad uut fülogeneetilist meetodit, mis ei vaja fülogeneesipuude ehitamist või käsitsi võrdlemist. Viiesajast analüüsitud ortoloogsete geenide komplektist umbes 70% olid kooskõlas null-hüpoteesiga, mis ütleb, et geenid on muutunud sama kiirusega ning päranduvad vertikaalselt. Autorid viisid läbi ka fülogeneetilise analüüsi geenidele, mis näitasid suurimat variatsiooni. Molekulaarse kella evolutsioonist kõrvalekaldega analüüsitud geenidest üle poole olid tõenäoliselt horisontaalse ülekande ühe vormi – ksenoloogse geeniasenduse (*xenologous gene displacement*) – tulemus. Ülejäänud geenide erinevus null-hüpoteesist võib olla tingitud evolutsiooni kiiruse liigispetsiifilise muutusega.

Ühendades *Escherichia coli* MG1655 parameetrilise meetodi põhjal saadud HGT tulemused nelja lähedase liigi fülogeneetilise analüüsiga, märkisid J.G. Lawrence ja H. Ochman (2002) tulemuste ulatuslikku kattumist. 755-st atüüpilisest geenist olid 627 (83%) ka fülogeneetilise meetodiga detekteeritud kui HGT. Tähtis on ka, et mõlemal viisil leitud ORF-id on genoomis enamasti järjestikku ühendatud, ehk eri meetodid leiavad sama horisontaalse ülekande sündmuse erinevaid alasid. Lõpphinnang on, et *E. coli* MG1655 genoomis on vähemalt 221 sündmuse käigus omandatud geene 24,5% (Lawrence, Ochman, 2002).

1.2.4 Füleetilise jaotuse meetod

Fülogeneetiliselt kokkusobimatud järjestused (*phylogenetically discordant sequence*,

PDS), mis on Clarke ja kaasautorite (2002) artikli põhiteema, on genoomi need ORF-id, millel on statistiliselt erinev sarnasuse liikide muster (*similarity relationship pattern*). Andmebaasi otsing tagastab lähimate homoloogsete liikide nimekirja. Enamikul genoomi geenidest on see nimekiri sarnane, kuid keskmiselt 10,8% ORF-idest oli see erinev.

Andmebaasi otsingu abil identifitseeriti füleetiliselt kokkusobimatud järjestused ning eemaldati nad fülogeneesipuu analüüsist. Analüüsitud ORF-ide põhjal koostatud evolutsioonipuud lahutasid edukalt suuremad bakterirühmad, mis olid kooskõlas ka rRNA puuga. PDS järjestuste analüüsist eemaldamisel puude konfiguratsioon peaaegu ei muutunud, kuid *bootstrap* toetus puude harudel tõusis. Võib järeldada, et kokkusobimatud ORF-id ei toeta kindlat alternatiivset puu topoloogiat ning suure tõenäosusega on need horisontaalselt ülekandunud (Clarke *et al*., 2002).

Mark Ragani võrdlevas töös (2001b) on PDS meetodi tulemused korrelatsioonis GC-osakaalu ning Markovi mudelit kasutavate parameetriliste meetoditega.

1.2.5 Paradoksaalne sugulus

Paradoksaalse parima sarnasuse (*paradoxical best hit*) meetod põhineb valkude või DNA andmebaasi otsingul (Koonin *et al*., 2001). Võrreldakse kõiki genoomi järjestusi kõigi andmebaasis olevate järjestustega, et leida homolooge. Klassifitseerides leitud homolooge sarnasuse (*alignment score* või *expected value*) ning evolutsioonilise kauguse järgi teadaolevas taksonoomilises jaotuses, leitakse sellised geenid, mille lähim sugulane (*best hit*) on evolutsiooniliselt kaugest liigist. Kuigi ka liigispetsiifiline deletsioon võib põhjustada sellist efekti, on geeni esinemise ja puudumise kindel muster oluline HGT näitaja (Koonin *et al*., 2001).

Thermotoga maritima oli üks esimesi baktereid, milles leiti ja tõestati ulatuslik horisontaalne geeniülekanne. 16S rRNA analüüsi järgi on T. maritima üks vanimaid harusid eubakterite riigis. 1877 geenist 24% olid sarnasemad arhea geenidele. Aquifex aeolicus genoomis oli see number 16% ning B. subtilis vaid 7%. Enamik infotöötlemise geene on sarnasemad eubakteri homoloogidega, kuid 49% transportvalkudest, 60% elektronahela valkudest ning 42% tundmatu funktsiooniga valkudest on sarnasemad arheast pärit homoloogiga. Genoomi järjestuse $\chi 2$ analüüs leidis 51 regiooni, mis on oluliselt erineva koostisega, kusjuures 42 neist sisaldasid geene, mille kõige sarnasem homoloog oli pärit

arheast või termofiilse A. aeolicus genoomist (Nelson *et al*., 1999).

1.3 Genoomi signatuur

1.3.1 Järjestuste liigispetsiifilisus

Genoomi signatuur (GS) on oligonukleotiidsete sõnede esinemise sageduste hulk järjestuses. Geneetilise info signatuuri sõne pikkus on tavapäraselt 2 kuni 10 nukleotiidi.

Igal bakteriliigil on genoomi lõikes talle omane signatuur. Ühe genoomi erinevate geenide vahel püsib signatuur suhteliselt muutumatuna (Dufraigne *et al*. 2005).

Horisontaalselt ülekandunud geenil on doonorliigiga sarnased parameetrid (näiteks aluspaari koostis, koodonkasutus) ning ka signatuur (Garcia-Vallvé *et al*, 2000). Teatud aja jooksul peale ülekande toimumist hakkavad uuele geenile mõjuma retsipiendi genoomile omased reeglid. Geeni parameetrid muutuvad genoomi keskmiste parameetrite sarnaseks. Ülekandunud geen retsipiendi genoomist on võimalik leida signatuuri erinevuse järgi, kuid evolutsioonis ammu toimunud ülekannete detekteerimine signatuuri ja muude parameetrite järgi on raskendatud.

Peab märkima, et igas genoomis esineb hulk geene, mis omavad alati ülejäänud genoomist erinevat signatuuri (Tsirigos, Rigoutsos, 2005). Nende hulka kuuluvad kõrgelt ekspresseeruvad ning ribosomaalsete järjestuste geenid.

1.3.2 Järjestuste klassifikatsioon oligomeeride abil

Libiseva akna abil leitakse oligonukleotiidide sagedused geenile, genoomile, või vaid mõnele regioonile. Sageduste väärtused moodustavad n-mõõtmelise vektori, kus $n = 4^m$; m = oligonukleotiidse sõne pikkus. Kui võtta sõne pikkuseks 4 nukleotiidi, on signatuuri mõõduks 64, 6 nukleotiidi puhul – 4096.

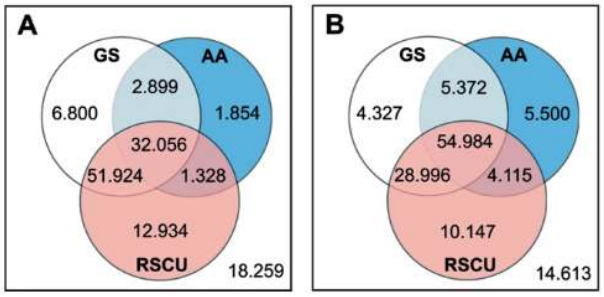
Sageduste vektorite võrdlemiseks ning sarnasuse hindamiseks võib kasutada kovariatsioon- ning korrelatsioonanalüüsi, arvutada Mahalanobis'e või Eucleides'e kaugus või teha χ^2 test (Tsirigos, Rigoutsos, 2005; Fertil *et al*., 2005).

1.3.3 Genoomi signatuuri rakendamine

Genoomi signatuuri meetod (GS) on olemuselt parameetriline. Võrreldes teiste parameetriliste meetoditega omab GS oluliselt suuremat lahutusvõimet ning täpsust.

Erinevate meetodite võrdlust on läbi viinud A. Tsirigos ja I. Rigoutsos (2005). Nad tõestasid originaalsete testimiskriteeriumite abil, et GS meetod leiab HGT suurema täpsusega kui muud parameetrilised meetodid: G+C sisaldus, koodonkasutus või koodoni adaptatsiooni indeks (CAI).

GS ja RSCU (sünonüümsete koodonite kasutus) meetodite järjestuse identifitseerimise täpsus 3-nukleotiidsete koodonite esinemise sageduse mõõtmisel on võrreldav (Sandberg *et al*., 2003). GS meetod saavutab parima tulemuse, kui kasutab pikemaid (kuni 9 nt) sõnesid (Tsirigos, Rigoutsos, 2005).



Joonis 1. Parameetriliste meetodite tulemuste kattuvus. A – sõne pikkus 3 nt, B – sõne pikkus 9 nt. GS – genoomi signatuur; AA – aminohappe kasutus; RSCU – sünonüümsete koodonite kasutus (Sandberg et al., 2003).

Signatuuri poolt antav informatsioon peegeldab suurel määral teiste parameetriliste meetodite informatsiooni. Signatuuri meetodi abil leitud geenihulk sisaldas endas kuni 95% võrra teiste meetodite abil leitud geene (Joonis 1) (Sandberg *et al*., 2003).

Lisaks sellele andis GS häid tulemusi katsetes, kus tuli leida varem kirjeldatud horisontaalse ülekande juhtumid (Dufraigne *et al*., 2005; Tsirigos, Rigoutsos, 2005; Sandberg *et al*., 2001). Horisontaalse ülekande puhul lubab GS leida ka organismi, millest pärineb ülekandunud järjestus (Sandberg *et al*., 2001; Dufraigne *et al*., 2005). Seejuures ei ole vajalik doonororganismi täisgenoomi järjestuse olemasolu.

Eelpooltoodust võib järeldada, et genoomi signatuuril on horisontaalse geeniülekande detekteerimisel suurim rakendamise efektiivsus.

2. Töö eesmärgid

Käesoleva uurimistöö üheks eesmärgiks oli genoomi signatuuri leidmise programmi koostamine ning rakendamine genoomifailide töötlemiseks.

Teiseks, 16S ribosomaalse RNA ja Ef-Tu valgu järjestustel põhinevate evolutsioonipuude konstrueerimine, võrdlemine ja vastuolude kontroll, kasutades maksimaalset arvu prokarüootide täisgenoome.

Kolmandaks, genoomi signatuuri ja fülogeneesipuude andmete integreerimine ning võimalike horisontaalse geeniülekande juhtumite analüüs genoomi signatuuri meetodi abil.

3. Materjal ja metoodika

3.1 Analüüsitavate liikide valim

Töös kasutatud liikide nimekirja valik põhines seisuga 10.01.06 täielikult sekveneeritud prokarüootide genoomide nimistul. 25 arhea ja 269 bakteriliigi nimetused on toodud Lisas 1.

Võimalikult suurt liikide arvu taotleti kolmel põhjusel. Esiteks, liikide mitmekesisuse tagamiseks. Teiseks, lähedaste liikide vahel toimuvatest protsessidest ülevaate saamiseks. Kolmandaks, genoomi signatuuri meetodi sobivuse testimiseks suurte andmemahtude analüüsimisel.

Ef-Tu valk uurimisobjektina valiti järgnevatel kaalutlustel:

- On levinud kõikides prokarüootsetes liikides
- Omab olulist funktsiooni translatsioonimehhanismis (aminoatsüül-tRNA kinnitamine ribosoomile) ning on seetõttu konserveerunud järjestusega.
- Sisaldab informatsiooni kaugete liikide omavahelise suguluse kohta.
- Annab rohkem võimalusi HGT toimumiseks, kuna võib esineda genoomis kahe koopiana.
- On põhjalikult uuritud ning annoteeritud.

Ef-Tu valgu järjestust on muuhulgas kasutatud 16S rRNA andmete täiendamiseks elusloodu evolutsioonipuu juure määramisel (Woese, 2000).

Ef-Tu geen paikneb genoomis koos teiste translatsiooni valkude geenidega (Lathe, Bork, 2001), mis on sageli erineva nukleotiidse koostisega. Sellest tulenevalt on raskendatud tüüpilisuse hindamine parameetriliste meetoditega, kuna ei ole teada, kas konkreetsel juhul on tegemist HGT või geeni loomuliku seisundiga (Nakamura *et al*., 2004). Siin muutub oluliseks genoomi signatuuri meetodi võime näidata geeni ja sellega külgneva ala tüüpilisuse väärtust.

3.2 Järjestuste andmebaasid

Ribosomal Database Project andmebaas (Cole *et al*., 2005) sisaldab ribosomaalse RNA järjestusi joondatud ja annoteeritud kujul. Esindatud on üle 100 000 erineva rRNA geeni järjestuse. Antud töö jaoks kasutati joonduse versiooni 37 FastaA tekstiformaadis (http://rdp.cme.msu.edu/download/release9_37_aligned.fasta.bz2).

Andmebaasis sisalduvatel järjestustel on kasutatud erilist liiginime formaati ning seetõttu teostati otsing modifitseeritud liikide nimekirja alusel.

Liikide täisgenoomide nukleotiidsed ja aminohappelised järjestused ning annotatsioonid saadi GenBank andmebaasist (Benson *et al*., 2006). Genoomi signatuuri analüüsi programm kasutab FastaA formaadis „fna“ laiendiga nukleotiidse järjestuse faile ning „ptt“ ja „rnt“ annotatsiooni faile. Ef-Tu valgujärjestuste eraldamiseks FastA formaadis „faa“ laiendiga failidest teostati otsing järjestuste nime järgi.

Puuduvatele rRNA ja Ef-Tu järjestustele tehti otsing NCBI (National Center for Biotechnology Information) Entrez otsingusüsteemis (http://www.ncbi.nlm.nih.gov/entrez/).

3.3 Programmid ja parameetrid

DNA ja valgujärjestused joondati programmis ClustalW (Chenna *et al*., 2003). Kasutati käsurealt parameetrite sisestamist. Valgujärjestuse puhul märgiti võrdlusmaatriksi tüübiks BLOSUM. Nukleotiidsete järjestuste lisamine olemasolevale joondusele toimus järjestuse profiili joondamise teel (*profile alignment*).

Joondatud järjestusel eemaldati üle 90% lünklikud veerud programmis Belvu (autor prof. Erik Sonnhammer). Joonduse konverteerimine PHYLIP formaati toimus programmis GeneDoc (Nicholas *et al*., 1997).

Fülogeneesipuude konstrueerimine ja edasine töötlus viidi läbi paketi PHYLIP versioon 3.6a ja 3.6b programmide *dnaml*, *proml*, *seqboot*, *retree, consense* abil (Felsentstein, 1993). *Dnaml* ja *proml* on programmid fülogeneesipuude ehitamiseks, vastavalt DNA või valgu järjestustele, suurima tõepära (*maximum likelihood*) meetodil. See on üks usaldusväärsemaid, kuid väga arvutusmahukas fülogeneesipuu ehitamise meetod (Nei, 1996). Kasutati vaikimisi parameetreid, välja arvatud juhuslikult valitud järjestuste sisendjärjekord (*randomize input order of sequences*, parameeter J), mis vähendab algmaterjali struktuuri mõju tulemustele

(Tuimala, 2005). *Bootstrap* andmehulkade moodustamiseks kasutati programmi *seqboot*. *Bootstrap* andmehulgad sisestati mitmekordse andmehulgana (*multiple data set*). 100 *bootstrap* puud ühendati programmis *consense* ning ühendatud puu muudeti juureta (*unrooted*) fülogeneesipuuks programmi *retree* abil.

Fülogeneesipuude visualiseerimiseks kasutati programmi *MEGA* versioon 3.1 (Kumar *et al*., 2004).

Mitmesuguste faili sisuga teostatud operatsioonide jaoks, nagu liikide nimede asendamine, failide konverteerimine, otsing jm, kasutati autori poolt kirjutatud programme, kasutades Perl (Practical Extraction and Reporting Language, http://www.perl.org) programmeerimiskeelt (Tisdall, 2003).

Genoomi signatuuri analüüs oli teostatud autori poolt kirjutatud programmi Sig.pl abil, mida vaadeldakse lõigus 4.3 (Lisa 4).

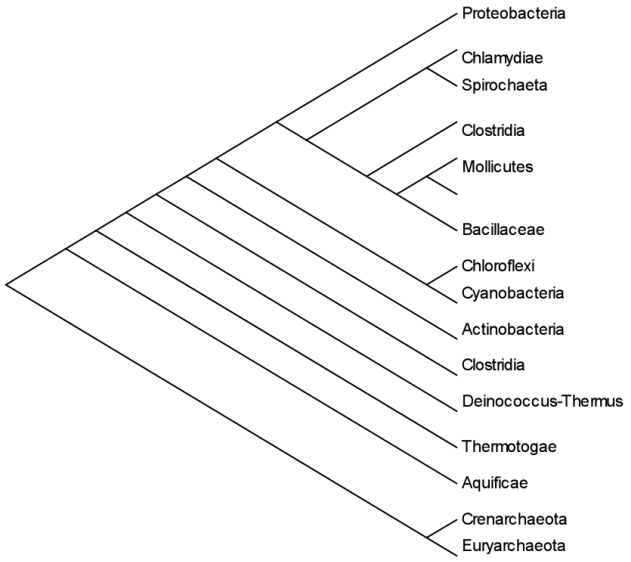
Genoomi signatuuri leidmise programm Sig.pl kasutab Jettero Helleri statistika analüüsi moodulit, mille kasutamine toimub GPL litsentsi alusel.

4. Tulemused

4.1 16S rRNA fülogeneesipuu koostamine

4.1.1 Algandmete töötlemine

Kõigi uurimistöös kasutatud liikide jaoks oli vajalik konstrueerida 16S rRNA fülogeneesipuu (Joonis 2). Seda kasutatakse evolutsiooni kõrvalekallete identifitseerimiseks.



Joonis 2. 16S rRNA järjestuste põhjal konstrueeritud puu skemaatiline konfiguratsioon

Ribosomal Database Project andmebaasis sisalduvad eelnevalt joondatud RNA järjestused paljudele liikidele ning nende alamliikidele ja tüvedele. Selle joonduse kasutamine on tagab korrektse evolutsioonipuu, sest joonduse koostamisel on arvestatud 16S rRNA omaduste ning struktuuridega. Andmebaasist liiginimedele vastavate järjestuste eraldamine toimus Perl programmi abil, mis võrdles 294 liigi nimekirja ja järjestuste nimesid. RDP andmebaasis leidus täpne vaste 284 liigile.

10 liigi või tüve puhul ei leitud joondatud RNA järjestust või oli see puudulik. NCBI Entrez andmebaasis tehti otsing puuduvatele järjestustele ning leiti 9 vastet. *Streptococcus pneumoniae TIGR4* tüvele vastavat 16S rRNA järjestust leida ei õnnestunud. RDP 284 liigi järjestuste joondusele lisati 9 puuduvat järjestust. Kasutati CLUSTALW *profile alignment* võimalust.

4.1.2 *Maximum likelihood* puu ja *bootstrap* kontroll

16S ribosomaalse RNA järjestuse põhjal konstrueeriti fülogeneesipuu (vt Lisa 2) programmi *dnaml* abil. See kasutab suurima tõepära (*maximum likelihood*) meetodit. Käesolevas analüüsis langes 16S rRNA puu konfiguratsioon kokku üldaktsepteeritud ribosomaalse klassifikatsiooniga (Guerrero, 2001).

Puu hargnemise usaldusväärsuse kontroll on teostatud *bootstrap* meetodil (Efron *et al*., 1996). 100 *bootstrap* koopia loomiseks kasutati PHYLIP paketi programmi *seqboot*. *Bootstrap* puude ühendamiseks kasutati programmi *consense „Majority rule extended"* reeglit. See võimaldas lahendada korrektselt ka väikese *bootstrap* väärtusega fülogeneesipuu harud (näiteks puu juure ligidal). Puu visualiseerimiseks paigutati fülogeneesipuu juurepunkt eubakterite ja arheate lahknemise punkti.

4.2 Ef-Tu fülogeeneesipuu koostamine

4.2.1 Algandmete töötlemine

Korrektse valgu fülogeneesi pildi saamiseks on vajalik analüüsida võimalikult täielikku valkude järjestuste hulka (Hoef-Emden, 2004).

Ef-Tu geeni järjestused on võrdlemisi hästi annoteeritud (Lathe, Bork, 2001). Tuleb arvestada, et elongatsiooni faktor Tu (Ef-Tu) valk võib genoomis esineda kuni kahes koopias (Ke *et al*., 2000). Seda arvestades on valgujärjestuste eraldamiseks tehtud otsing GenBank aminohappe järjestustes iga genoomi kohta. Puuduvatele järjestustele tehti lisaks otsing, kasutades NCBI Entrez otsingusüsteemi. Kokku analüüsiti 398 Ef-Tu valgu järjestust. Kahe liigi puhul, Magnetospirillum magneticum AMB-1 ning Wolinella succinogenes, ei leidunud annoteeritud Ef-Tu järjestust.

Eristatud valgujärjestused joondati programmi ClustalW abil, kasutades BLOSUM maatriksit.

4.2.2 *Maximum likelihood* puu ja *bootstrap* kontroll

Ef-Tu valgu järjestuse põhjal konstrueeriti fülogeneesipuu (vt Lisa 3) programmi *proml* abil. See kasutab suurima tõepära (*maximum likelihood*) meetodit.

Puu hargnemise usaldusväärsuse kontroll on teostatud *bootstrap* meetodil (Efron *et al*.,

1996). 100 *bootstrap* koopia loomiseks kasutati PHYLIP paketi programmi *seqboot*. *Bootstrap* puude ühendamiseks kasutati programmi *consense „Majority rule extended"* reeglit. See võimaldas lahendada korrektselt ka väikese *bootstrap* väärtusega fülogeneesipuu harud (näiteks puu juure ligidal). Puu visualiseerimiseks paigutati fülogeneesipuu juurepunkt eubakterite ja arheate lahknemise punkti. Ef-Tu valgu fülogeneesipuu andmed on kooskõlas 16S rRNA evolutsioonipuuga, välja arvatud järgmises lõigus nimetatud liikide puhul.

4.2.3 Ef-Tu ja 16S rRNA puude võrdlusanalüüs

Kahe puu analüüsil leitud erinevusi võib seletada, esiteks, järjestuste joondamise ning puu ehitamise vigadega; teiseks, evolutsiooni kõrvalekallete esinemisega. Kõige tõenäolisem evolutsiooni pildi kõrvalekallete põhjustaja on horisontaalne geeniülekanne.

Enterobacteriaceae rühmas esines väikese *bootstrap* väärtusega või täielikult lahendamata harusid. Selle rühma bakterid on evolutsiooniliselt väga lähedal ning Ef-Tu geen ei kanna piisavat informatsiooni liikide korrektseks lahutamiseks.

Yersinia nelja liigi puhul paigutusid Ef-Tu geeni koopiad kahte rühma. Ülejäänud liikide puhul on kaks geeni koopiat alati koos. Selline kahe koopia järjestuse sarnasus on tõenäoliselt põhjustatud geenikonversiooni mehhanismi poolt (Lathe, Bork, 2001).

Puude võrdlemisel on arvesse võetud Ef-Tu ebapiisav lahutusvõime puu juure lähedaste bakterirühmade eristamisel. Seetõttu vale paigutusega puul, kuid madala *bootstrap* väärtusega liigirühmi ei vaadeldud.

Edasine GS meetodiga HGT analüüs viidi läbi järgmiste fülogeneesipuudel erinevalt paigutunud liikidega:

- Bdellovibrio bacteriovorus HD100
- Borrelia burgdorferi B31
- Borrelia garinii PBi
- Campylobacter jejuni RM1221
- Campylobacter jejuni subsp. jejuni NCTC 11168
- Dehalococcoides ethenogenes 195
- Dehalococcoides sp. CBDB1
- Hahella chejuensis KCTC 2396

- Helicobacter hepaticus ATCC 51449
- Helicobacter pylori 26695
- Helicobacter pylori J99
- Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
- Leptospira interrogans serovar Lai str. 56601
- Rhodobacter sphaeroides 2.4.1
- Symbiobacterium thermophilum IAM 14863
- Thiomicrospira denitrificans ATCC 33889
- Treponema denticola ATCC 35405
- Treponema pallidum subsp. pallidum str. Nichols

4.3 Genoomi signatuuri analüüsi programm Sig.pl

4.3.1 Programmi eesmärk

Genoomi signatuuri ja fülogeneesipuude andmete integreerimise ülesande jaoks ei sobi olemasolev genoomi signatuuri analüüsi programm GENSTYLE (Fertil et al., 2005). Käesoleva töö käigus oli koostatud programm GS meetodi implementeerimiseks ning võimaluste uurimiseks (Lisa 4). Üks programmi koostamise eesmärkidest oli selle meetodi integreereerimine olemasolevate kontrollitud fülogeneesimeetoditega.

Selleks, et näha, kas ülekanne toimus kogu geeni ulatuses, hõlmas vaid osa sellest või haaras ka külgnevat ala, oli programmi koostamise eesmärgiks arvutada genoomi signatuuri väärtus arvestamata geeni piiridega. Programm pidi andma informatsiooni, millised geenid asuvad igas segmendis. Tüüpilisuse väärtused arvutati samaaegselt kovariatsiooni ja korrelatsiooni indeksitena, et võrrelda nende näitajate poolt antava informatsiooni kvaliteeti.

Programm pidi olema kasutatav vabalt valitud geenide jaoks ning lihtsalt täiustatav püstitatud ülesande jaoks optimaalse algoritmi leidmiseks.

4.3.2 Sisendfailid ja parameetrid

Programm kasutab sisendfailina tabelit, mis sisaldab liigi nime, järjestuse faili nime ning kataloogi. Käsurealt saab programm tabeli reanumbreid tähistavad kaks arvu, mis

määravad töölõigu alguse ja lõpu.

Toetudes Tsirigos ja Rigoutsos (2005) töö tulemustele, oli sõne pikkuseks valitud 6 nt. See tagab informatsiooni piisava usaldusväärsuse optimaalse arvutusajaga.

Signatuur saavutab maksimaalse täpsuse alates 1000 nt segmendi laiusest (Sandberg *et al*., 2001). Samal ajal segmendi laius 1000 nt on väiksem, kui Ef-Tu geeni keskmine pikkus. See võimaldab saada geeni õige signatuuri, kuna enamasti jääb üks segment antud geeni piiridesse.

Sammu pikkuseks järjestusel on 500 nt ning iga järgmine analüüsitav segment kattub eelmisega pooles ulatuses.

Segmendi sees loetakse iga oligonukleotiidi esinemise sagedused 1 nt sammuga. Samaaegselt arvestatakse nii pärisuunalise kui vastusuunalise ahelaga.

Programm võrdleb igat segmenti kogu genoomi signatuuriga ja tagastab nii kovariatsiooni kui korrelatsiooni väärtused.

4.3.3 Programmi töökäik

Programmi käivitamisel loetakse sisendfail, ning käivitatakse järgmine alamprogramm iga sisendfailis sisalduva liigi või replikoni kohta.

Alamprogramm teostab järgmised toimingud:

1. Loetakse sisse kogu genoomi järjestus
2. Arvutatakse kogu genoomi signatuur ning salvestatakse see mällu
3. Käivitatakse tsükkel, mis:

a) loeb segmendi;

b) arvutab selle signatuuri;

c) võrdleb segmendi signatuuri ja genoomi signatuuri (kahel viisil);

d) kirjutab tulemuse tabelisse;

e) kontrollib, millised geenid asuvad segmendi alal ning kirjutab nende nimed tabelisse;

f) liigub järjestusel edasi.

4. Programm paneb tabeli nimeks analüüsitud faili nime, millele lisatakse laiend „csv“.

Sig.pl programm tagastab ühe tabelifaili analüüsitud genoomi kohta. Tabelis on segmendi kohta üks rida, mis sisaldab järjekorra numbrit, kovariatsiooni ja korrelatsiooni

indekseid ning annotatsioonifailidest võetud geeninimesid. RNA järjestuste eristamiseks märgistatakse need eraldi veerus.

4.3.4 Kovariatsioon ja korrelatsioon

Kovariatsioon ja korrelatsioon on statistilised meetodid kahe muutuja omavahelise sõltuvuse määramiseks. Antud programmis kasutatakse nende väärtuseid kahe signatuuri omavahelise sarnasuse hindamiseks.

Kui korrelatsiooni väärtus muutub vahemikus 1 kuni -1, siis kovariatsiooni väärtuse ulatus sõltub hinnatavate arvude suurusest. Nullilähedane kovariatsiooni väärtus tähendab seose ja sellega ka sarnasuse puudumist analüüsitava segmendi ja genoomi signatuuride vahel. Negatiivne kovariatsiooni väärtus tähendab, et analüüsitava segmendi signatuur on pöördvõrdeline ülejäänud genoomi suhtes. Sel juhul on signatuuride vaheline erinevus suurim.

GS analüüsil andis kovariatsiooni indeks usaldusväärsemaid tulemusi kui korrelatsiooni indeks.

4.3.5 Genoomi signatuuri meetodi jõudlus

GS meetodi oluliseks eeliseks on võimalus võrrelda ning klassifitseerida järjestusi väiksema arvutusmahuga, kui järjestuste joondamise abil. Arvutusaeg kasvab lineaarses sõltuvuses analüüsitavate andmete mahuga. Keskmise suurusega genoomi järjestuse töötlemine 6 nt pikkuse sõne puhul kestab Athlon 64 2.4 Ghz arvutil 4 minutit. Analüüsitava sõne pikkuse suurendamine ühe nt võrra tõstab arvutusaega 4 korda.

Enamuse arvutusajast võtab kovariatsiooni ja korrelatsiooni väärtuste määramine kahe signatuuri vahel. Kasutades lihtsamat, näiteks Eucleides'e kauguse arvutamist võib programmi tööaega vähendada ligi kolm korda.

Programmi tööajal oli mälukasutus alla 100MB.

Kasutatud programmeerimiskeel Perl sobib tekstilise informatsiooni töötlemiseks ning võimaldab lihtsalt teostada muudatusi programmis. See on efektiivne bioloogiliste järjestuste töötlemise ja tõlgendamise ülesannete jaoks. Matemaatiline analüüs ning suuremahuline arvutamine on siiski C või C++ programmeerimiskeeltes kümneid kordi kiirem.

4.3.6 Meetodi perspektiivikus

Antud programmis on GS meetodit kasutatud iga ala tüüpilisuse võrdlemiseks genoomi keskmisega. Lihtsalt ja väheste muutustega võib sellele meetodile püstitada teisi ülesandeid. Näiteks, mingi ala võrdlus teiste liikide genoomidega, HGT doonori otsimiseks. Aga samuti üksteisele järgnevate alade võrdlemine signatuuri muutumise visualiseerimiseks. Teoreetiliselt võib viimane lähenemisviis olla abiks väga muutuvate, heterogeensete genoomide järjestuse analüüsil.

4.4 Geeniülekande hüpoteesi kontroll genoomi signatuuri meetodi abil

4.4.1 Horisontaalse geeniülekande detekteerimise kriteeriumid

Otsuse tegemine geeni evolutsioonilise ajaloo ning HGT esinemise kohta põhines järgmistel kriteeriumitel:

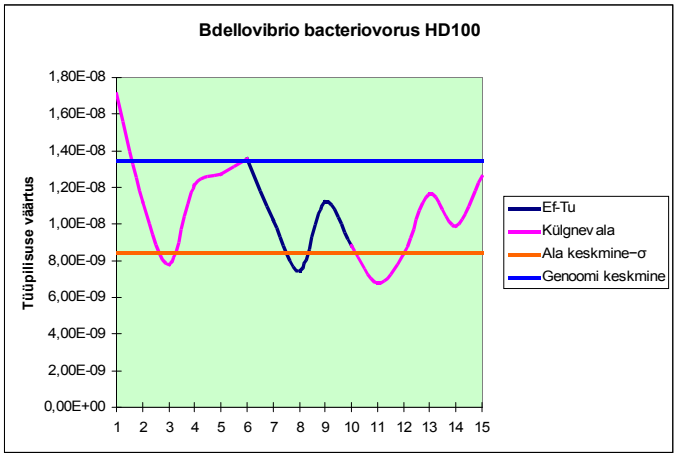
- Järjestamine tüüpilisuse väärtuse järgi ning grupeerimine „anomaalsesse“ rühma graafikul. Ef-Tu geeni puhul võib see anda valepositiivse tulemuse (Nakamura *et al*., 2004);
- Erinevus külgneva ala (100kb lai aken) keskmise väärtusega üle 1 standardhälbe (σ);
- Oluliselt madalam tüüpilisuse väärtus, võrreldes külgneva alaga.
- Sugulasliikides leitud HGT puhul sarnase tüüpilisuse mustri esinemine, mis eraldi ei viita HGT toimumisele.
- Kahtluse puhul võrdlemine tulemustega, mis on saadud kasutades 8 nt sõne pikkust.

Eelnevates töödes on kasutatud teatud lävendeid HGT esinemise automaatseks detekteerimiseks (Tsirigos, Rigoutsos, 2005; Nakamura *et al*., 2004). Selline lähenemisviis sobib HGT ulatuse määramiseks genoomide lõikes. Ef-Tu geeni ebatüüpilist koostist silmas pidades on siiski vajalik iga üksikjuhu eraldi vaatlemine.

4.4.2 Bdellovibrio bacteriovorus

Bdellovibrio bacteriovorus HD100 omab ühte Ef-Tu geeni (gi|42524390). See bakter kuulub δ-proteobakterite hulka, mida kinnitasid 16S rRNA puu andmed.

Ef-Tu valgu fülogeneesipuul positsioneerus see liik α-proteobakterite rühma keskele. *Bootstrap* väärtus 16% näitab sellise konfiguratsiooni vähest usaldusväärsust.



Joonis 3. Bdellovibrio bacteriovorus HD100 Ef-Tu ala tüüpilisuse väärtused (Lisa 5).

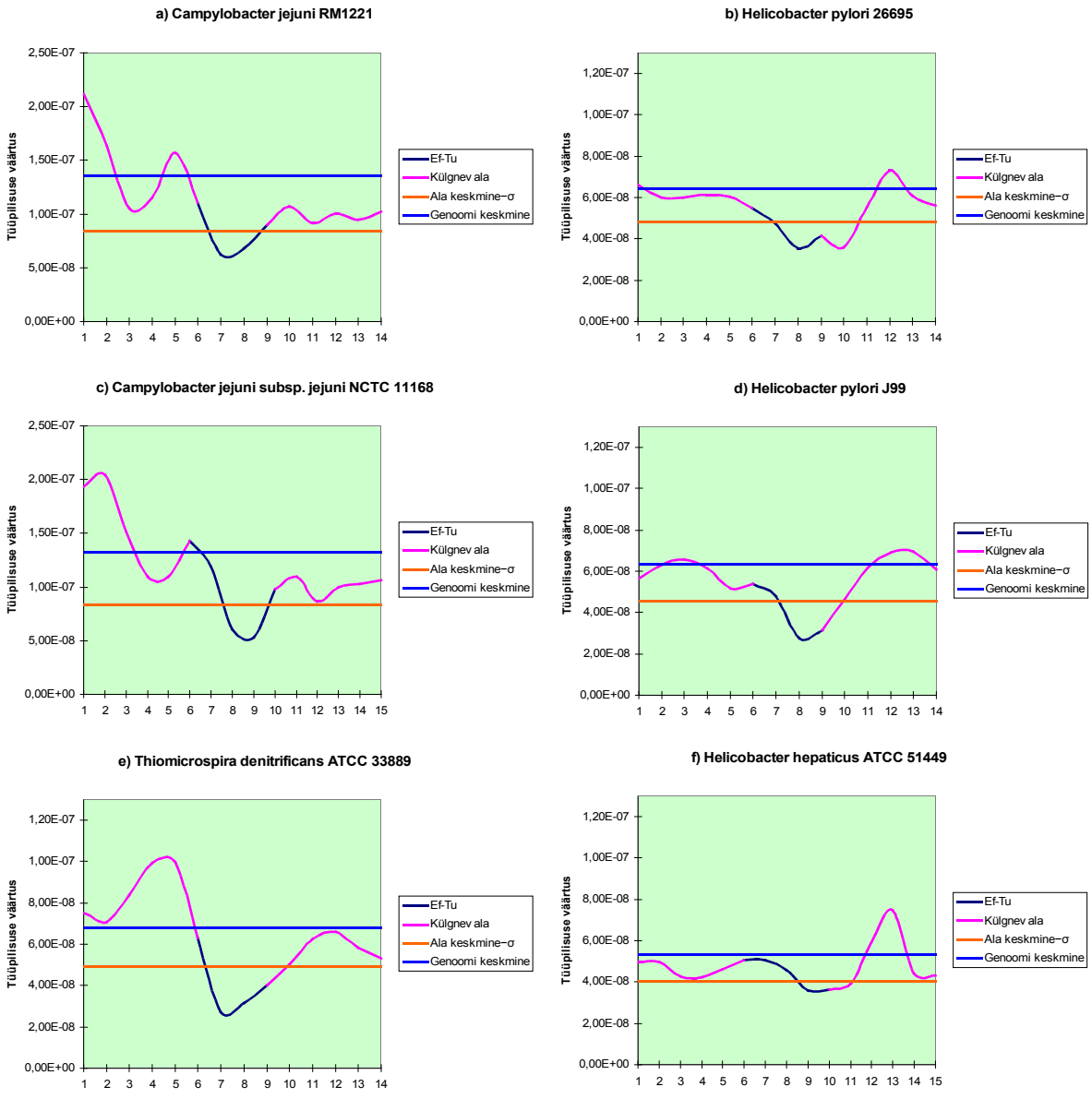
GS analüüs Ef-Tu valgu piirkonnas näitas, et antud geeni ala on muutuva tüüpilisuse väärtusega ning selle keskosa segmendil on tüüpilisuse väärtus alla standardhälbe piirist(Joonis 3). Kuid arvestades tüüpilisuse väärtuse suurt kõikumist geeniga vahetult külgneval alal tuleb järeldada, et antud juhul ei ole toimunud HGT.

4.4.3 Epsilon-proteobakterid

Campylobacter jejuni RM1221, *C. jejuni subsp. jejuni* NCTC 11168, *Helicobacter pylori* J99, *H. pylori* 26695, *H. hepaticus* ATCC 51449 ning *Thiomicrospira denitrificans* ATCC 33889 kuuluvad ε-proteobakterite taksonisse. 16S rRNA puul paiknesid nimetatud liigid *Bacteroides*/*Chlorobi* rühma kõrval. Bergey's klassifikatsioonis on *Bacteroides*/*Chlorobi* monofüleetiline hõimkond ning ε-proteobakterid on proteobakterite hõimkonna alamjaotus. Samas *bootstrap* väärtus 36% on rRNA puu puhul suhteliselt madal.

Ef-Tu fülogeneesipuul paigutusid kõik analüüsis osalevad 6 ε-proteobakterit ühtse rühmana tsüanobakterite kõrvalharusse. Sellise hargnemise *bootstrap* väärtus on kõigest 6%, kuid ε-proteobakterite ja ülejäänud proteobakterite suguluse puudumine annab märku anomaalia esinemisest. Fakt, et kuus liiki grupeerusid ühtse rühmana, kuid valesse harusse viitab Ef-Tu geeni HGT toimumise võimalusele nende liikide eelkäijal.

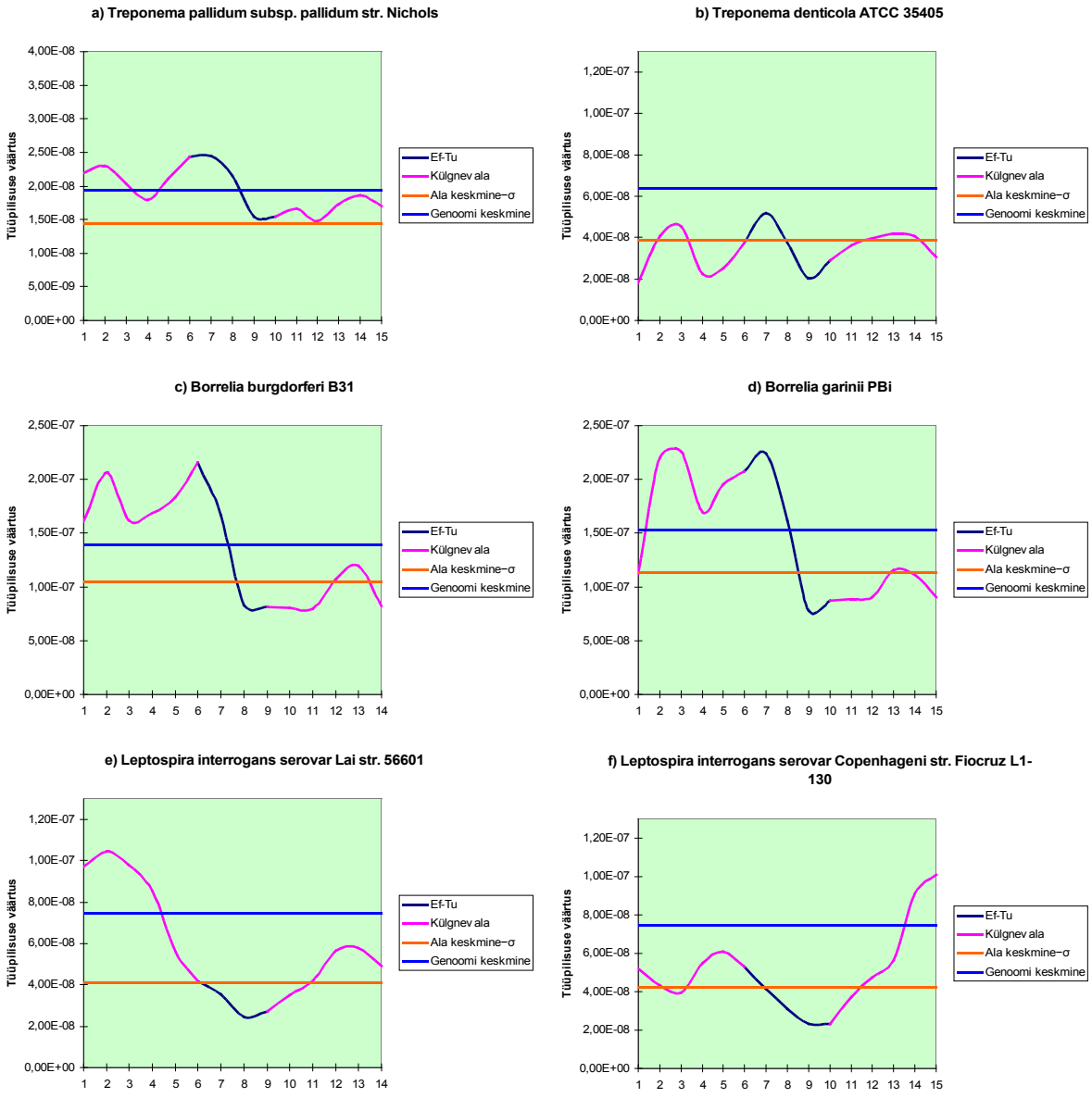
Genoomi signatuuri analüüs näitas Ef-Tu järjestuse tüüpilisuse väärtuse erinevust külgnevate alade suhtes ning ka genoomi keskmisest (Joonis 4). Külgnevad geenid ning nende signatuuri muster on sarnased kõigil kuuel liigil. Võib järeldada, et ε-proteobakterite ühise eelkäia Ef-Tu geen on üle kandunud teisest liigist.



Joonis 4. Epsilon-proteobakterite Ef-Tu geeni ala tüüpilisuse väärtuste kõverad. Kõigil liikidel on Ef-Tu geeni tüüpilisuse väärtus madalam kui külgnevatel aladel ning märgatavalt madalam kui genoomi keskmine. Geeni äärte kõrgem tüüpilisuse väärtus võib olla tingitud sellest, et mõõtmine toimus segmendis, mis hõlmas osaliselt ka külgnevat ala (Lisad 6-11).

**Helicobacter liikidel on Ef-Tu valk kodeeritud pärisuunaliselt ning ülejäänud liikidel vastassuunaliselt.*

4.4.4 Spiroheedid



Joonis 5. Spiroheetide Ef-Tu geeni ala tüüpilisuse väärtuste kõverad. Kõigil liikidel välja arvatud Leptospira interrogans serovar Copenhageni on Ef-Tu geeni lõpuosa tüüpilisuse väärtus madalam kui alguse osa, mis on tõenäoliselt tingitud ribosomaalsete geenide mõjust (Lisad 12-17).*

**Leptospira interrogans serovar Copenhageni Ef-Tu valk on kodeeritud vastassuunaliselt, ülejäänud liikidel pärisuunaliselt*

Borrelia burgdorferi B31, *B. garinii* PBi, *Leptospira interrogans serovar Lai str.* 56601, *L. interrogans serovar Copenhageni str. Fiocruz* L1-130, *Treponema pallidum subsp. pallidum str. Nichols* ja *T. denticola* ATCC 35405 kuuluvad Bergey's klassifikatsiooni järgi

monofüleetilisse *Spirochaeta* hõimkonda. 16s rRNA puu andmetel grupeeruvad spiroheedid monofüleetiliselt klass *Chlamydiae* bakterirühma kõrvalharusse (sellise topoloogia *bootstrap* väärtus 26% võib näidata, et 16S rRNA fülogeneesipuu juure lähedased harud ei ole korrektsed).

Ef-Tu fülogeneesipuul asetusid *Spirochaeta* liigid α-proteobakterite kõrvalharusse ning Leptospira liigid eraldusid ülejäänud spiroheetidest. Selline ebatavaline paiknemine pakkus huvi edasiseks analüüsiks.

GS tüüpilisuse väärtuste kõverad (Joonis 5) on keerulise kujuga, kuid geeni keskosa tüüpilisuse väärtus on sarnane külgnevate alade ja genoomi keskmiste väärtustega (välja arvatud *Leptospira* liikidel). Tõenäoliselt on taoline kõvera kuju tingitud Ef-Tu geeni järel olevate ribosomaalsete valkude järjestuste mõjust (Lisad 12-17). Toodud andmete põhjal ei saa väita spiroheetide Ef-Tu geeni kohta HGT esinemist või puudumist. Antud juhul on vajalik täiendav fülogeneetiline ja ka signatuuri uuring.

4.4.5 Rhodobacter sphaeroides

Ef-Tu fülogeneesipuu koostamise aluseks võeti *Rhodobacter sphaeroides 2.4.1* valkude nimistust kaks Ef-Tu geeni (gi|77462257 ja gi|77464017). Ef-Tu puul paigutub geen gi| 77464017 arheate rühma kõrvalharusse. *Bootstrap* väärtus on 28%. Teise geeni paigutus langes kokku 16S rRNA puu andmetega ning asub α-proteobakterite rühma sees. Valgu järjestusel põhinev otsing RefSeq andmebaasist (http://www.ncbi.nlm.nih.gov/RefSeq/) näitas, et antud geenil on homoloogia *TypA* ja *LepA* GTP-siduvate valkudega.

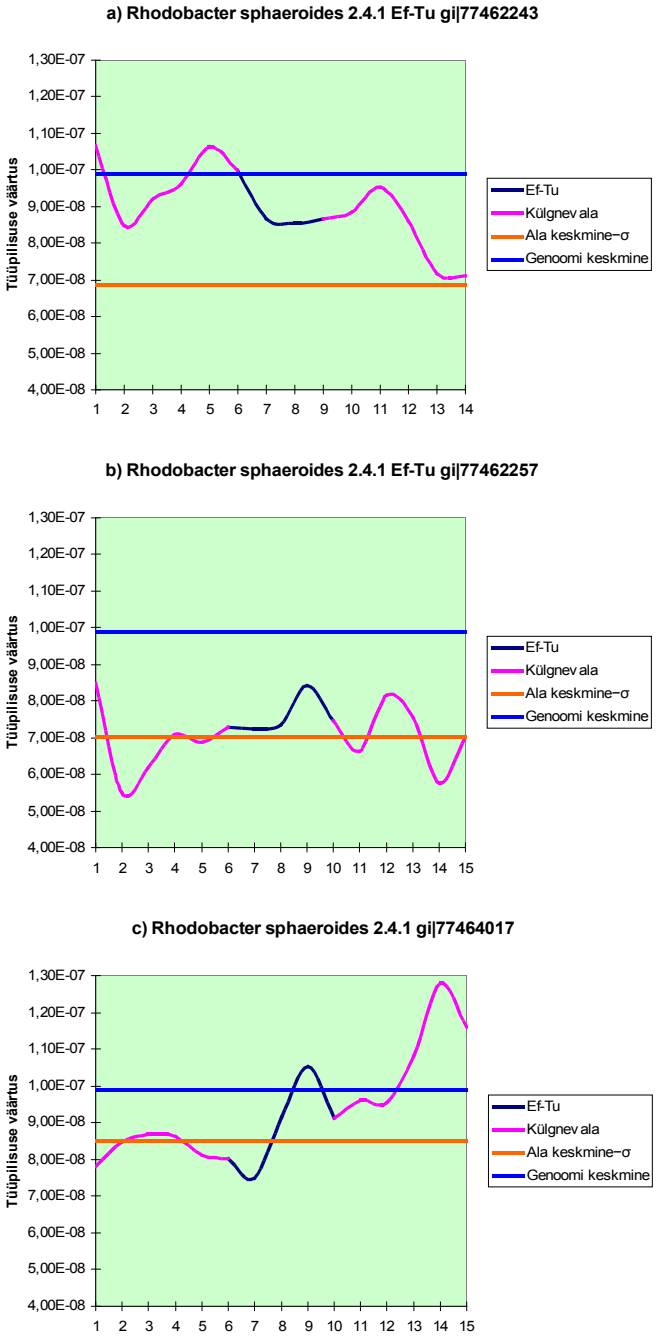
Täiendava otsingu tulemusena valkude nimistust selgus, et *Rhodobacter sphaeroides 2.4.1* GenBank valgujärjestuses esineb kolm Ef-Tu nimelist valku (lisaks eelpool nimetatutele ka gi|77462243). Antud bakteri sugulasliikidel esineb kaks Ef-Tu geenikoopiat.

Genoomi signatuuri väärtus ei näidanud ebatüüpilisust Ef-Tu geenide puhul (Joonis 6). Sarnast tüüpilisuse väärtuse kõverate pilti näitavad paljud gram-negatiivsed bakteriliigid, kelle genoomis esineb kaks Ef-Tu geeni koopiat. Geeni gi|77464017 tüüpilisus kõigub suures vahemikus.

Toodud faktidest võib järeldada, et geeni gi|77464017 nimetus GenBank andmebaasis on vale ning seetõttu sattus ta antud uuringus ekslikult Ef-Tu valkude hulka.

4.4.6 Hahella chejuensis

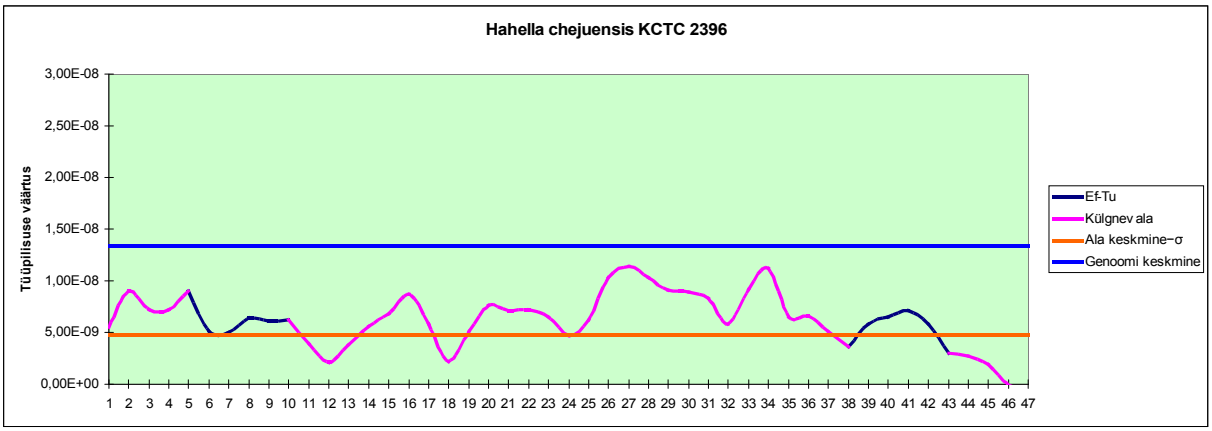
Hahella chejuensis KCTC 2396 genoomis on leitud kaks Ef-Tu geeni. Ef-Tu ja 16S rRNA fülogeneesipuudel paigutus antud liik γ-proteobakterite erinevatesse alamrühmadesse. 16S rRNA puul on *Hahella* lähimad sugulasliigid määratud korrektselt. Ef-Tu puu näitab valkude järjestuse sarnasust *Hahella* ja *Pseudomonas* liikide vahel.



Joonis 6. Rhodobacter sphaeroides Ef-Tu geenide tüüpilisuse näitajad. a) ja b) õiged Ef-Tu geenid, c) valesti identifitseeritud Ef-Tu geen (Lisa 18, 19, 20)

Genoomi signatuuri analüüs näitas, et mõlemad Ef-Tu geenid ja nendega külgnevad alad omavad sarnast tüüpilisuse väärtust (kovariatsiooni väärtuse kõikumine 5,05 x 10^{-9} kuni 7,08 x 10^{-9}). Genoomi järjestuses

paiknevad Ef-Tu geenid lähestikku, 50 kb pika madala tüüpilisuse väärtuse saarel (Ef-Tu saare tüüpilisuse keskmise väärtus 6,10 x 10^{-9}; genoomi keskmine 1,35 x 10^{-8}, maksimaalne 3,90 x 10^{-8}, minimaalne -5,16 x 10^{-9}).



Joonis 7. Hahella chejuensis Ef-Tu geenide tüüpilisuse väärtuste kõver. Kaks Ef-Tu geeni paiknevad genoomis lähestikku (Lisa 21)

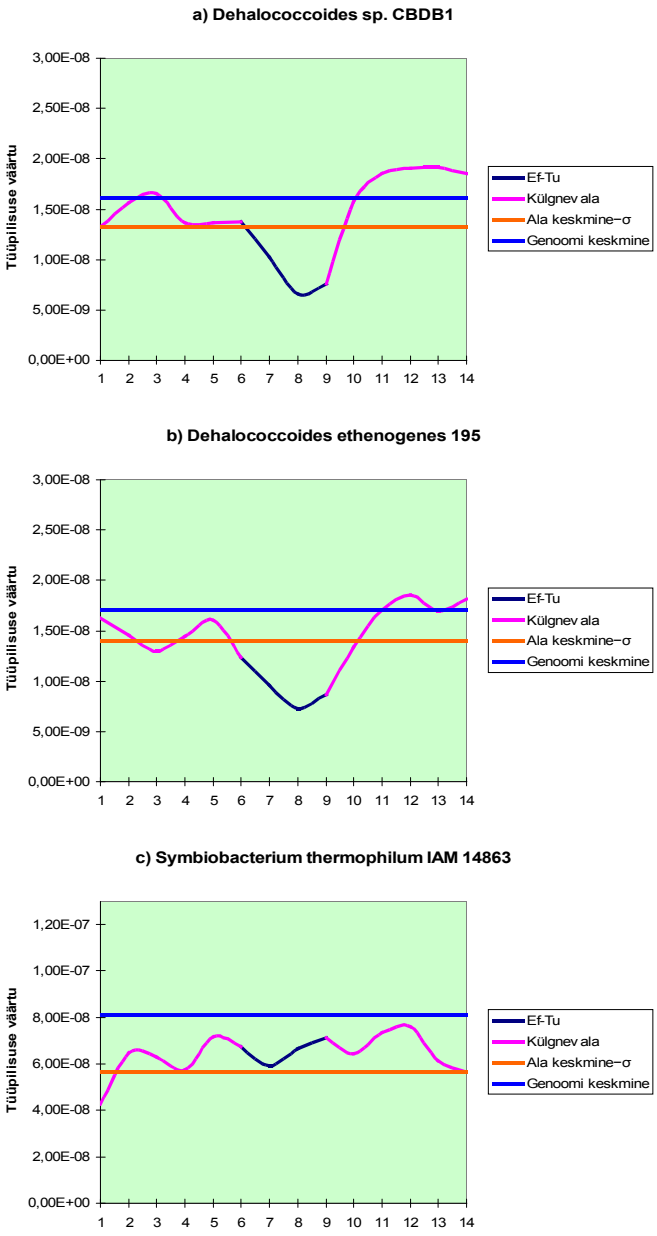
H. chejuensis genoom on väga suur (7,2 Mb) ja heterogeenne. HGT on mänginud olulist rolli selle bakteriliigi evolutsioonis. 23% genoomist on moodustatud 69 ebatüüpilise koostisega genoomi saarekese poolt (Jeong *et al*, 2005). Genoomi signatuuri andmed tõendavad, et tüüpilisuse väärtus on genoomi ulatuses väga varieeruv (Joonis 7). Seetõttu on signatuuri andmete tõlgendamine raskendatud. Lokaalsete signatuuride erinevuste võrdlusest võib järeldada, et Ef-Tu valkude puu erinevus 16S rRNA puust on tingitud muudest põhjustest kui HGT.

4.4.7 Symbiobacterium ja Dehalococcoides liigid

Dehalococcoides sp. CBDB1 ja *D. ethenogenes 195* kuuluvad fotosünteesivate väävlibakterite klassi (*Chloroflexi*). 16S rRNA puul paigutuvad nad *Cyanobacteria* rühma välisharusse. *Symbiobacterium thermophilum IAM 14863* asetub 16S rRNA puul Actinobacteria rühma välisharusse, mis on kooskõlas Bergey's klassifikatsiooniga (Garrity et al., 2004). Ef-Tu valgu fülogeneesipuul paigutuvad kolm nimetatud liiki *Clostridia* klassi bakterite rühma keskele.

Läbiviidud genoomi signatuuride analüüs näitas, et mõlema *Dehalococcoides* liigi puhul näitab Ef-Tu valgu geeni järjestuse signatuur olulist ebatüüpilisust. Geeniga külgnevatel aladel suureneb tüüpilisuse väärtus järsult (Lisa 22, 23). Need faktid viitavad Ef-Tu geeni ülekande toimumisele kahe liigi eelkäija genoomis (Joonis 8a, 8b).

S. thermophilum genoomi signatuuri analüüs ei näidanud kõrvalekaldeid Ef-Tu geenis ja selle ümbruses. Tüüpilisuse väärtused on genoomi keskmise lähedased ja muutuvad vähe (5,91 x 10^{-8} kuni 7,12 x 10^{-8}, keskmine 8,14 x 10^{-8}) (Joonis 8c, Lisa 24). Arvestades madala *bootstrap* väärtusega Ef-Tu puu *S. thermophilum* harus võib oletada, et tegemist on puu konstrueerimise veaga ning geeni ajalugu ei ole mõjutatud HGT poolt.



Joonis 8. a) ja b) Dehalococcoides liikide Ef-Tu geeni tüüpilisuse väärtuste kõverad, mis tõendavad HGT toimumist. c) Symbiobacterium liigi Ef-Tu tüüpilisuse väärtuste kõver (Lisa 22, 23, 24).

5. Arutelu ja järeldused

Horisontaalse geeniülekande (HGT) detekteerimine bakterite genoomides omab olulist väärtust bakterite evolutsiooni uurimisel. Olemasolevaid HGT detekteerimise meetodeid võib jagada kirjanduse ülevaate põhjal kolmeks grupiks: fülogeneesipuudel põhinevad, andmebaasiotsingul põhinevad ning parameetrilised meetodid.

Viimase aja horisontaalse geeniülekande detekteerimise alased uurimistööd on pühendatud uutele meetoditele, mis ei vaja töömahukat fülogeneesipuude uurimist. Laialdaselt kasutatakse otsingut andmebaasidest, mis annab seda paremaid tulemusi, mida rohkem genoome on sekveneeritud.

Genoomi signatuuri meetod on andnud horisontaalselt ülekandunud geenide detekteerimisel häid tulemusi (Dufraigne et al., 2005; Tsirigos, Rigoutsos, 2005), kuid selle laialdane kasutamine on pärsitud sobiva tarkvaralahenduse puudumisest. Käesoleva töö eesmärkideks oli genoomi signatuuri (GS) meetodit imlementeeriva programmi loomine ning GS ja fülogeneesipuu meetodite integreerimine horisontaalse ülekande detekteerimiseks.

Uurimise aluseks võeti Ef-Tu valk, mis on fülogeneesi mõttes sobiv – konserveerunud ning laialt levinud valk tagas informatiivse fülogeneesipuu. Samas osutus Ef-Tu valgu geen genoomi signatuuri seisukohalt keerukaks uurimisobjektiks. Paljude liikide puhul paistab silma Ef-Tu geeni ebatüüpiline koostis. Tsirigos ja Rigoutsos (2005) on analüüsinud GS abil täisgenoome ja paigutanud tulemused andmebaasi (http://cbcsrv.watson.ibm.com/HGT/). Ef-Tu geen on selles andmebaasis enamikus genoomides märgitud kui ülekandunud. Tõenäoliselt ei ole aga kasutatud kriteeriumid piisavad automaatse otsuse tegemiseks selle geeni puhul.

Matemaatilise modelleerimise alusel võib väita, et ülekanduv järjestus ei ole piiratud geeni täpsete mõõtmetega vaid võib hõlmata tervet geeni või osa sellest ning ka geeniga külgnevat ala (Novozhilov *et al*., 2005). Horisontaalse geeniülekande uurimistöödes kasutatakse aga terviklikke geene või avatud lugemisraame (Koonin *et al*., 2001; Ragan, 2001b; Tsirigos, Rigoutsos, 2005) ning seetõttu võivad ülekande detekteerimise tulemused

olla ebatäpsed.

Selleks, et kontrollida, kas geeni külgnevate aladega arvestamine lisab uut informatsiooni, arvutab käesolevas töös kasutatud algoritm signatuuri kogu genoomile sõltumata geeni piiridest. Genoomi signatuuri meetodit implementeeriv programm Sig.pl kasutab GenBank andmebaasi järjestusi. Ef-Tu geeni ajaloo kohta otsuse tegemiseks vaadeldi lokaalset ja ülegenoomilist signatuuri muutumist geeniga külgnevatel aladel. Horisontaalse ülekande kriteeriumiks oli Ef-Tu geeni oluline erinevus genoomi keskmisest signatuurist ning samaaegselt ka erinevus sellega külgnevatest aladest.

Ef-Tu geeni vaatlemisel ainult genoomi signatuuri abil ei saa teha lõplikku järeldust horisontaalse ülekande kohta. Külgnevate alade tüüpilisuse väljaselgitamine on oluline fülogeneesi puude informatsiooni kinnitamiseks või ümberlükkamiseks kuna see annab täielikuma pildi geeni ajaloost.

294 liigi 16S rRNA puu ei näidanud kõrvalekaldeid üldtunnustatud evolutsioonipuust ning seda kasutati evolutsiooni normaalse kulgemise etalonina. Võrdluses Ef-Tu geeniga näitas 16S rRNA järjestus paremat lahutusvõimet ka sügavate harude puhul. Ef-Tu geeni järjestuse fülogeneesipuu eristas suuremad bakterigrupid õigesti, kuigi *bootstrap* väärtus puu juure lähedastes harudes on madal.

Ef-Tu geeni põhjal konstrueeritud fülogeneesipuu erines vaid üksikutes harudes 16S rRNA puust. Kahe puu analüüsil leitud erinevusi võib seletada esiteks järjestuse joondamise ning puu ehitamise vigadega, teiseks evolutsiooni kõrvalekallete esinemisega. Selliseid võimalikke kõrvalekaldeid analüüsiti Sig.pl programmi abil GS meetodiga.

Käesoleva töö raames on üheksal liigil leitud Ef-Tu HGT sündmus, kolme liigi puhul saadud tulemuseks HGT puudumine. Ef-Tu geeni HGT leidis tõestust 6 liigil ε-proteobakterite rühmas ning kahel *Dehalococcoides* liigis. Analüüsi tulemusena oli HGT toimumine välistatud *Bdellovibrio bacteriovorus, Symbiobacterium thermophilum*, *Rhodobacter sphaeroides* ning *Hahella chejuensis* liikide Ef-Tu geenides.

Spiroheetide juhtumi analüüs GS abil ei andnud ühest vastust. Genoomi signatuuri andmed on spiroheetide Ef-Tu geeni puhul ebakindlad ning fülogeneesipuu vähese tõenäosusväärtusega. Seetõttu on HGT hüpoteesi kontrollimiseks vajalik täiendava uuringu läbiviimine.

Genoomi signatuuri programmi edasisel täiendamisel tuleb silmas pidada, et kõrgelt

ekspresseeruvate ja RNA geenide, milledel on alati atüüpiline koostis, genoomi keskmise signatuuri arvutamisest eemaldamine, muudab keskmise väärtuse sarnasemaks ülejäänud geenide suhtes ning see toob võimalikud ülekandunud geenid rohkem esile. Lisaks geeni ja genoomi keskmiste väärtuste võrdlusele tuleks lisameetodina rakendada geeni ja sellega külgnevate alade tüüpilisuse omavahelist võrdlust, mis annab täiendavat informatsiooni geeni ehitusest. Käesolevas töös oli signatuuri sõne pikkuseks valitud 6 nt eelkõige arvutusmahukuse vähendamiseks. Samas on Sandberg kaastöötajatega (2003) ning Tsirigos, Rigoutsos (2005) uurinud sõne pikkuse mõju ning leidnud, et sõne suurendamine 8 nukleotiidini annab parima tulemuse järjestuse omaduste määramiseks (Tsirigos, Rigoutsos, 2005).

Tänapäeva suuremahuliste analüüside tegemine bioloogiliste järjestuste ja muude bioinformaatika probleemide puhul vajab korduvat programmide käivitamist ning otsingu tegemist järjestustes. Selliste ülesannete automatiseerimine tõstab töö kiirust ja efektiivsust.

Tehtud töö alusel võib teha järgmised järeldused:

- 16S rRNA geeni ja Ef-Tu valgu järjestuste põhjal on koostatud 294 prokarüootse liigi fülogeneesi puud. 16S rRNA puu on kooskõlas üldtunnustatud evolutsioonipuuga. Kahe puu erinevused näitasid võimalikke HGT juhtumeid.
- Genoomi signatuuri meetod võimaldab matemaatiliselt hinnata järjestuse tüüpilisust antud genoomi lõikes. HGT detekteerimise põhiline kriteerium on järjestuse tüüpilisuse madal väärtus.
- Reeglina annab genoomi signatuuri meetod häid tulemusi, kuid selle laialdane kasutamine on pärsitud sobiva tarkvaralahenduse puudumisest.
- Käesoleva uurimistöö käigus koostati genoomi signatuuri leidmise programm ning teostati analüüs, mis hõlmas 18 liigi Ef-Tu geeni. Välja töötatud algoritm lubab hinnata ka geeniga külgnevate alade signatuuri, mis annab täiendavat informatsiooni geeni ajaloost ja struktuurist.
- Kaheksal liigil on leitud Ef-Tu HGT sündmus, nelja liigi puhul saadud tulemuseks HGT puudumine. Spiroheetide juhtumi analüüs GS abil ei andnud ühest vastust.
- Ef-Tu geeni HGT oli tõendatud ε-proteobakterite rühmas ning *Dehalococcoides*

liikides.

• Analüüsi tulemusena oli HGT välistatud *Bdellovibrio bacteriovorus, Symbiobacterium thermophilum*, *Rhodobacter sphaeroides* ning *Hahella chejuensis* liikide Ef-Tu geenides.

• Genoomi signatuuri meetod tõestas oma informatiivsust. Välja töötatud programmi võib lihtsalt adapteerida eri ülesannete täitmiseks.

6. Kokkuvõte

Käesoleva töö eesmärkideks oli genoomi signatuuri (GS) meetodit implementeeriva programmi loomine ning GS ja fülogeneesipuu meetodite integreerimine horisontaalse ülekande detekteerimiseks.

Reeglina annab genoomi signatuuri meetod häid tulemusi, kuid selle laialdane kasutamine on pärsitud sobiva tarkvaralahenduse puudumisest.

16S rRNA geeni ja Ef-Tu valgu järjestuste põhjal on konstrueeritud 294 prokarüootse liigi fülogeneesi puud, kusjuures 16S rRNA puu oli kooskõlas kanoonilise evolutsioonipuuga. On koostatud Sig.pl programm genoomi signatuuri arvutamiseks. 18 liigi Ef-Tu geeni ala on analüüsitud genoomi signatuuri meetodiga. Ef-Tu geeni ajaloo kohta otsuse langetamiseks on tehtud lokaalse ja ülegenoomilise signatuuri vaatlus.

Kaheksal liigil on leitud Ef-Tu geenis HGT sündmus, nelja liigi puhul on tõestatud HGT puudumine. Spiroheetide juhtumi analüüs GS abil ei andnud ühest vastust.

Ef-Tu geeni HGT on tõestatud ε-proteobakterite rühmas (*Campylobacter spp*., *Helicobacter spp*. ja *Thiomicrospira denitrificans*) ning kahes *Dehalococcoides* liikides. HGT oli välistatud *Bdellovibrio bacteriovorus, Symbiobacterium thermophilum*, *Rhodobacter sphaeroides* ning *Hahella chejuensis* liikide Ef-Tu geenides.

7. Summary

Implementation of genomic signature method for detection of horizontal gene transfer

Aleksander Sudakov

Detection of horizontal gene transfer (HGT) in bacterial genomes has a great value for evolutionary studies. Current methods for detection of HGT can be divided into three major groups: ones that are based on phylogenetic trees, ones that utilize database searches, and parametric methods. Phylogenetic trees are commonly used to picture true evolutionary history of a gene or a protein. HGT may be detected as incongruence between trees. Recent studies in the field of detection of HGT focus on evaluating the extent of HGT in genomes and on developing new methods that do not require manual analysis of phylogenetic trees.

Genomic signature measures occurrences of oligonucleotides in a sequence. Genomic signatures can be utilized to detect regions with atypical composition. This method shows better results than other parametric methods. The genomic signature is a flexible and powerful tool for analysis of genomic sequences.

In this work the genomic signature method was used to support or reject a horizontal gene transfer hypothesis in a number of species that show incongruence between phylogenetic trees. Two maximum likelihood phylogenetic trees were constructed for 294 prokaryotic species based on sequence of 16S rRNA gene and on sequence of elongation factor Tu protein (Ef-Tu). 18 species were further analysed using the program Sig.pl that implements genomic signature method. Typicality score for Ef-Tu gene and adjoining sequences were evaluated.

In 8 species genomic signature data supports HGT hypothesis for Ef-Tu gene: 6 ε-proteobacteria species (*Campylobacter spp*., *Helicobacter spp*. and *Thiomicrospira denitrificans*) and 2 *Dehalococcoides species*. In four species genomic signature found no atypical composition and HGT was rejected (*Bdellovibrio bacteriovorus, Symbiobacterium*

thermophilum, *Rhodobacter sphaeroides* and *Hahella chejuensis*). A group of 6 *Spirochaeta* species showed inconsistent results.

8. Kirjanduse loetelu

Ambler, R.A., Daniel, M., Hermoso, J., Meyer, T.E., Bartsch, R.G., Kamen, M.D. 1979. Cytochrome c2 sequence variation among recognized species of purple nonsulfur photosynthetic bacteria. Nature 278: 659-660.

Avery O.T., MacLeod C.M., McCarty M. 1944. Studies on chemical nature of the sunstance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. 79: 137-158.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J, Ostell, J., Wheeler, D.L. 2006. GenBank. Nucleic Acids Research 34: D16-D20.

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E.R., Nesbø, C.L., Case, R.J., Doolittle, W.F. 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu. Rev. Genet. 37: 283-328.

Brochier, C., Bapteste, E., Moreira, D., Philippe, H. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends in Genetics 18: 1-5.

Calteau, A., Daubin, V., Perrieère, G. 2004. Super-tree approach for studying the phylogeny of prokaryotes: new results on completely sequenced genomes. LNCS 3039: 700-708.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D. 2003. Multiple sequence alignment with Clustal series of programs. Nucleic Acids Research 34: 3497-3500.

Cicarelli, F.D., Doerks, T., Mering von, C., Creevey, C.J., Snel, B., Bork, P. 2006. Towards automatic reconstruction of a highly resolved tree of life. Science 311: 1283:1287.

Clarke, G.D.P., Beiko, R.G., Ragan, M.A., Charlebois, R.L. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. Journal of Bacteriology 184: 2072-2080.

Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam, S.A., McGarrell, D.M., Garrity, G.M., Tiedje, J.M. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Research 33: D294-D296.

Daubin, V., Gouy, M., Perrière, G. 2002. A phylogenetic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Research 12: 1080-1090.

Daubin, V., Moran, N.A., Ochman, H. 2003. Phylogenetics and the cohesion of bacterial genomes. Science 201: 829-832.

Doolittle, W.F., Boucher, Y., Nesbø, C.L., Douady, C. J., Andersson, J. O., Roger, A. J.

2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? The Royal Society 358: 39-58.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A., Deschavanne, P. 2005. Detection and characterization of horizontal gene transfer in prokaryotes using genomic signature. NAR 33: 1-12.

Efron, B., Halloran, E., Holmes, S. 1996. Bootstrap confidence levels for phylogenetic trees. PNAS 93: 7085-7090.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Fertil, B., Massin, M., Lespinats, S., Devic, C., Dumee, P., Giron, A. 2005. GENSTYLE: exploration and analysis of DNA sequences with genomic signature. NAR 33:W512-W515.

Garcia-Vallvé, S. 2003. Horizontal gene transfer database (HGT-DB) at URL: http://www.fut.es/~debb/HGT/.

Garcia-Vallvé, S., Romeu, A., Palau, J. 2000. Horizontal Gene transfer in Bacterial and Archaeal Complete Genomes. Genome Research 10: 1719-1725.

Garrity, G.M., Bell, J.A., Lilburn, T.G. 2004. Taxonomic outline of prokaryotes Bergey's manual of systematic bacteriology. Second edition. Springer 399 p.

Guerrero, R. 2001. Bergey's manuals and the classification of prokaryotes. Int. Microbiol. 4: 103-9.

Hoef-Emden, K. 2004. An introduction to molecular phylogeny. Köln University.

Jeong, H., Yim, J.H., Lee, C., Choi, S., Park, Y., Yoon, S.H., Hur, C., Kang, H., Kim, D., Lee, H.H., Park, K.H., Park, S., Park, H., Lee, H.K., Oh, T.K., Kim, J.F. 2005. Genomic blueprint of *Hahella chejuensis*, a marine microbe producing an algicidal agent. Nucleic Acids Research 33: 7066-7073.

Ke, D., Boissinot, M., Huletsky, A., Picard, F.J., Frenette, J., Ouellette, M., Roy, P.H., Bergeron, M.G. 2000. Evidence for horizontal gene transfer in evolution of elongation factor Tu in enterococci. Journal of Bacteriology 182: 6913-6920.

Koonin, E.V., Makarova, K.S., Aravind, L. 2001. Horisontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55: 709-42.

Koski, L.B., Morton, R.A., Golding, G.B. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Molecular Biology and Evolution 18: 404-412.

Kumar, S., Tamura, K., Nei, M. 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Briefings in Bioinformatics 5: 150-163.

Lathe, W.C., Bork, P. 2001. Evolution of *tuf* genes: ancient duplication, differential loss and gene conversion. Federation of European Biochemical Societies 502: 113-116.

Lawrence, J.G., Ochman, H. 2002. Reconciling the many faces of lateral gene transfer. Trends in Microbiology 10: 1-4.

Nakamura, Y., Itoh, T., Matsuda, H., Gojobori, T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nature Genetics 36: 760-766.

Nei, M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. 30: 371-403.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., Fraser, C.M. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature 399: 323-329.

Nesbø, C.L., Boucher, Y., Doolittle, W.F. 2001. Defining the Core of Nontransferable Prokaryotic Genes: The Euryarchaeal Core. J. Mol. Evol. 53: 340-350.

Nicholas, K.B., Nicholas H.B. Jr., Deerfield, D.W. II. 1997 GeneDoc: analysis and visualization of genetic variation, EMBNEW News 4: 14.

Novichkov, P.S., Omelchenko, M.V., Gelfland, M.S., Mironov, A.A., Wolf, Y.I., Koonin, E.V. 2004. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. J. of Bacteriol. 186: 6575-6578.

Novozhilov, A.S., Karev, G.P., Koonin, E.V. 2005. Mathematical modeling of evolution of horizontally transferred genes. Molecular Biology and Evolution 22: 1721-1732.

Omelchenko, M.V., Makarova, K.S., Wolf, Y.I., Rogozin, I.B., Koonin, E.V. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biology 4: R55.

Ragan, M.A. 2001a. Detection of lateral gene transfer among microbial genomes. Current Opinion in Genetics and Development 11: 620-626.

Ragan, M.A. 2001b. On surrogate methods for detecting lateral gene transfer. FEMS Microbiology Letters 201: 187-191.

Sandberg, R., Bränden, C-I., Ernberg, I., Cöster, J. 2003. Quantifying the species-specifity in genomic signatures, synonymous codon choise, amino acid usage and G + C content. Gene 311: 35-42.

Sandberg, R., Winberg, G., Bränden, C-I., Kaske, A., Ernberg, I., Göster, J. 2001. Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier. Genome Research 11: 1404-1409.

Schopf, J.W. 1992. The oldest fossils and what they mean. Major Events in history of life. p. 29-64.

Schwartz, R.M., Dayhoff, M.O. 1978. Origins of procaryotes, mitochondria and chloroplasts. Science 199: 395-403.

Tisdall, J.D. 2003. Mastering Perl for bioinformatics. O'Reilly 377 p.

Tsirigos, A., Rigoutsos, I. 2005. A new computational method for the detection of horizontal gene transfer events. Nucleic Acids Research 33: 922-933.

http://cbcsrv.watson.ibm.com/HGT/.

Tuimala, J. 2005. A primer to phylogenetic analysis using the PHYLIP package. CSC Scientific Computing. 55 p.

Twyman, R.M. 1998. Advanced molecular biology. A concise reference. BIOS Scientific publishers. Oxford. p. 201-221.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. 51: 221-271.

Woese, C.R. 2000. Interpreting the universal tree. PNAS 97: 8392-8396.

Woese, C.R., Olsen, G.J., Ibba, M., Söll, D. 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol. Mol. Biol. Rev. 64: 202-236.

Zukerandl, E., Paulig, L., 1965. Evolutionary divergence and convergence in proteins. Evolving genes and proteins. New York: Academic Press. p. 97-166.

9. Lisad

Lisa 1. Kasutatud bakteriliikide nimekiri

Arhead

1. Aeropyrum pernix K1
2. Archaeoglobus fulgidus DSM 4304
3. Haloarcula marismortui ATCC 43049
4. Halobacterium sp. NRC-1
5. Methanothermobacter thermautotrophicus str. Delta H
6. Methanocaldococcus jannaschii DSM 2661
7. Methanococcus maripaludis S2
8. Methanopyrus kandleri AV19
9. Methanosarcina acetivorans C2A
10. Methanosarcina barkeri str. fusaro
11. Methanosarcina mazei Go1
12. Methanosphaera stadtmanae DSM 3091
13. Nanoarchaeum equitans Kin4-M
14. Natronomonas pharaonis DSM 2160
15. Picrophilus torridus DSM 9790
16. Pyrobaculum aerophilum str. IM2
17. Pyrococcus abyssi GE5
18. Pyrococcus furiosus DSM 3638
19. Pyrococcus horikoshii OT3
20. Sulfolobus acidocaldarius DSM 639
21. Sulfolobus solfataricus P2
22. Sulfolobus tokodaii str. 7
23. Thermococcus kodakarensis KOD1
24. Thermoplasma acidophilum DSM 1728
25. Thermoplasma volcanium GSS1

Bakterid

26. Acinetobacter sp. ADP1
27. Agrobacterium tumefaciens str. C58 Cereon
28. Agrobacterium tumefaciens str. C58 UWashington
29. Anabaena variabilis ATCC 29413
30. Anaplasma marginale str. St. Maries
31. Aquifex aeolicus VF5
32. Azoarcus sp. EbN1
33. Bacillus anthracis str. Ames
34. Bacillus anthracis str. 'Ames Ancestor'
35. Bacillus anthracis str. Sterne
36. Bacillus cereus ATCC 10987
37. Bacillus cereus ATCC 14579
38. Bacillus cereus E33L
39. Bacillus clausii KSM-K16
40. Bacillus halodurans C-125
41. Bacillus licheniformis ATCC 14580
42. Bacillus licheniformis ATCC 14580 DSM 13
43. Bacillus subtilis subsp. subtilis str. 168
44. Bacillus thuringiensis serovar konkukian str. 97-27
45. Bacteroides fragilis NCTC 9343
46. Bacteroides fragilis YCH46
47. Bacteroides thetaiotaomicron VPI-5482
48. Bartonella henselae str. Houston-1
49. Bartonella quintana str. Toulouse
50. Bdellovibrio bacteriovorus HD100
51. Bifidobacterium longum NCC2705
52. Bordetella bronchiseptica RB50
53. Bordetella parapertussis 12822
54. Bordetella pertussis Tohama I
55. Borrelia burgdorferi B31
56. Borrelia garinii PBi
57. Bradyrhizobium japonicum USDA 110
58. Brucella abortus biovar 1 str. 9-941
59. Brucella melitensis 16M
60. Brucella melitensis biovar Abortus 2308
61. Brucella suis 1330
62. Buchnera aphidicola str. APS (Acyrthosiphon pisum)
63. Buchnera aphidicola str. Bp (Baizongia pistaciae)
64. Buchnera aphidicola str. Sg (Schizaphis graminum)
65. Burkholderia mallei ATCC 23344
66. Burkholderia pseudomallei 1710b
67. Burkholderia pseudomallei K96243
68. Burkholderia sp. 383
69. Burkholderia thailandensis E264
70. Campylobacter jejuni RM1221
71. Campylobacter jejuni subsp. jejuni NCTC 11168
72. Candidatus Blochmannia floridanus
73. Candidatus Blochmannia pennsylvanicus str. BPEN
74. Candidatus Pelagibacter ubique HTCC1062
75. Carboxydothermus hydrogenoformans Z-2901
76. Caulobacter crescentus CB15
77. Chlamydia muridarum Nigg
78. Chlamydia trachomatis A/HAR-13
79. Chlamydia trachomatis D/UW-3/CX
80. Chlamydophila abortus S26/3
81. Chlamydophila caviae GPIC

82. Chlamydophila pneumoniae AR39
83. Chlamydophila pneumoniae CWL029
84. Chlamydophila pneumoniae J138
85. Chlamydophila pneumoniae TW-183
86. Chlorobium chlorochromatii CaD3
87. Chlorobium tepidum TLS
88. Chromobacterium violaceum ATCC 12472
89. Clostridium acetobutylicum ATCC 824
90. Clostridium perfringens str. 13
91. Clostridium tetani E88
92. Colwellia psychrerythraea 34H
93. Corynebacterium diphtheriae NCTC 13129
94. Corynebacterium efficiens YS-314
95. Corynebacterium glutamicum ATCC 13032 Bielefeld
96. Corynebacterium glutamicum ATCC 13032 Kitasato
97. Corynebacterium jeikeium K411
98. Coxiella burnetii RSA 493
99. Dechloromonas aromatica RCB
100. Dehalococcoides ethenogenes 195
101. Dehalococcoides sp. CBDB1
102. Deinococcus radiodurans R1
103. Desulfotalea psychrophila LSv54
104. Desulfovibrio desulfuricans G20
105. Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
106. Ehrlichia canis str. Jake
107. Ehrlichia ruminantium str. Gardel
108. Ehrlichia ruminantium Welgevonden
109. Ehrlichia ruminantium str. Welgevonden
110. Enterococcus faecalis V583
111. Erwinia carotovora subsp. atroseptica SCRI1043
112. Escherichia coli CFT073
113. Escherichia coli K12
114. Escherichia coli O157:H7
115. Escherichia coli O157:H7 EDL933
116. Francisella tularensis subsp. tularensis Schu 4
117. Fusobacterium nucleatum subsp. nucleatum ATCC 25586
118. Geobacillus kaustophilus HTA426
119. Geobacter metallireducens GS-15
120. Geobacter sulfurreducens PCA
121. Gloeobacter violaceus PCC 7421
122. Gluconobacter oxydans 621H
123. Haemophilus ducreyi 35000HP
124. Haemophilus influenzae 86-028NP
125. Haemophilus influenzae Rd KW20
126. Hahella chejuensis KCTC 2396
127. Helicobacter hepaticus ATCC 51449
128. Helicobacter pylori 26695
129. Helicobacter pylori J99
130. Idiomarina loihiensis L2TR
131. Lactobacillus acidophilus NCFM
132. Lactobacillus johnsonii NCC 533
133. Lactobacillus plantarum WCFS1
134. Lactobacillus sakei subsp. sakei 23K
135. Lactococcus lactis subsp. lactis Il1403
136. Legionella pneumophila str. Lens
137. Legionella pneumophila str. Paris
138. Legionella pneumophila subsp. pneumophila str. Philadelphia 1
139. Leifsonia xyli subsp. xyli str. CTCB07
140. Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
141. Leptospira interrogans serovar Lai str. 56601
142. Listeria innocua Clip11262
143. Listeria monocytogenes EGD-e
144. Listeria monocytogenes str. 4b F2365
145. Magnetospirillum magneticum AMB-1
146. Mannheimia succiniciproducens MBEL55E
147. Mesoplasma florum L1
148. Mesorhizobium loti MAFF303099
149. Methylococcus capsulatus str. Bath
150. Moorella thermoacetica ATCC 39073
151. Mycobacterium avium subsp. paratuberculosis K-10
152. Mycobacterium bovis AF2122/97
153. Mycobacterium leprae TN
154. Mycobacterium tuberculosis CDC1551
155. Mycobacterium tuberculosis H37Rv
156. Mycoplasma capricolum subsp. capricolum ATCC 27343
157. Mycoplasma gallisepticum R
158. Mycoplasma genitalium G-37
159. Mycoplasma hyopneumoniae 232
160. Mycoplasma hyopneumoniae 7448
161. Mycoplasma hyopneumoniae J
162. Mycoplasma mobile 163K
163. Mycoplasma mycoides subsp. mycoides SC str. PG1
164. Mycoplasma penetrans HF-2
165. Mycoplasma pneumoniae M129
166. Mycoplasma pulmonis UAB CTIP
167. Mycoplasma synoviae 53
168. Neisseria gonorrhoeae FA 1090
169. Neisseria meningitidis MC58
170. Neisseria meningitidis Z2491
171. Nitrobacter winogradskyi Nb-255
172. Nitrosococcus oceani ATCC 19707
173. Nitrosomonas europaea ATCC 19718
174. Nitrosospira multiformis ATCC 25196
175. Nocardia farcinica IFM 10152
176. Nostoc sp. PCC 7120
177. Oceanobacillus iheyensis HTE831
178. Onion yellows phytoplasma OY-M
179. Parachlamydia sp. UWE25
180. Pasteurella multocida subsp. multocida str. Pm70
181. Pelobacter carbinolicus DSM 2380
182. Pelodictyon luteolum DSM 273

183. Photobacterium profundum SS9
184. Photorhabdus luminescens subsp. laumondii TTO1
185. Pirellula sp. SH1
186. Porphyromonas gingivalis W83
187. Prochlorococcus marinus str. MIT 9312
188. Prochlorococcus marinus str. MIT 9313
189. Prochlorococcus marinus str. NATL2A
190. Prochlorococcus marinus subsp. marinus str. CCMP1375
191. Prochlorococcus marinus subsp. pastoris str. CCMP1986
192. Propionibacterium acnes KPA171202
193. Pseudoalteromonas haloplanktis TAC125
194. Pseudomonas aeruginosa PAO1
195. Pseudomonas fluorescens Pf-5
196. Pseudomonas fluorescens PfO-1
197. Pseudomonas putida KT2440
198. Pseudomonas syringae pv. phaseolicola 1448A
199. Pseudomonas syringae pv. syringae B728a
200. Pseudomonas syringae pv. tomato str. DC3000
201. Psychrobacter arcticus 273-4
202. Ralstonia eutropha JMP134
203. Ralstonia solanacearum GMI1000
204. Rhodobacter sphaeroides 2.4.1
205. Rhodopseudomonas palustris CGA009
206. Rhodospirillum rubrum ATCC 11170
207. Rickettsia conorii str. Malish 7
208. Rickettsia felis URRWXCal2
209. Rickettsia prowazekii str. Madrid E
210. Rickettsia typhi str. Wilmington
211. Salinibacter ruber DSM 13855
212. Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
213. Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150
214. Salmonella enterica subsp. enterica serovar Typhi str. CT18
215. Salmonella enterica subsp. enterica serovar Typhi Ty2
216. Salmonella typhimurium LT2
217. Shewanella oneidensis MR-1
218. Shigella boydii Sb227
219. Shigella dysenteriae Sd197
220. Shigella flexneri 2a str. 2457T
221. Shigella flexneri 2a str. 301
222. Shigella sonnei Ss046
223. Silicibacter pomeroyi DSS-3
224. Sinorhizobium meliloti 1021
225. Staphylococcus aureus RF122
226. Staphylococcus aureus subsp. aureus COL
227. Staphylococcus aureus subsp. aureus MRSA252
228. Staphylococcus aureus subsp. aureus MSSA476
229. Staphylococcus aureus subsp. aureus Mu50
230. Staphylococcus aureus subsp. aureus MW2
231. Staphylococcus aureus subsp. aureus N315
232. Staphylococcus epidermidis ATCC 12228
233. Staphylococcus epidermidis RP62A
234. Staphylococcus haemolyticus JCSC1435
235. Staphylococcus saprophyticus subsp. saprophyticus
236. Streptococcus agalactiae 2603V/R
237. Streptococcus agalactiae A909
238. Streptococcus agalactiae NEM316
239. Streptococcus mutans UA159
240. Streptococcus pneumoniae R6
241. Streptococcus pneumoniae TIGR4
242. Streptococcus pyogenes M1 GAS
243. Streptococcus pyogenes MGAS10394
244. Streptococcus pyogenes MGAS315
245. Streptococcus pyogenes MGAS5005
246. Streptococcus pyogenes MGAS6180
247. Streptococcus pyogenes MGAS8232
248. Streptococcus pyogenes SSI-1
249. Streptococcus thermophilus CNRZ1066
250. Streptococcus thermophilus LMG 18311
251. Streptomyces avermitilis MA-4680
252. Streptomyces coelicolor A3(2)
253. Symbiobacterium thermophilum IAM 14863
254. Synechococcus elongatus PCC 6301
255. Synechococcus elongatus PCC 7942
256. Synechococcus sp. CC9605
257. Synechococcus sp. CC9902
258. Synechococcus sp. WH 8102
259. Synechocystis sp. PCC 6803
260. Zymomonas mobilis subsp. mobilis ZM4
261. Thermoanaerobacter tengcongensis MB4
262. Thermobifida fusca YX
263. Thermosynechococcus elongatus BP-1
264. Thermotoga maritima MSB8
265. Thermus thermophilus HB27
266. Thermus thermophilus HB8
267. Thiobacillus denitrificans ATCC 25259
268. Thiomicrospira crunogena XCL-2
269. Thiomicrospira denitrificans ATCC 33889
270. Treponema denticola ATCC 35405
271. Treponema pallidum subsp. pallidum str. Nichols
272. Tropheryma whipplei str. Twist
273. Tropheryma whipplei TW08/27
274. Ureaplasma parvum serovar 3 str. ATCC 700970
275. Vibrio cholerae O1 biovar eltor str. N16961
276. Vibrio fischeri ES114
277. Vibrio parahaemolyticus RIMD 2210633
278. Vibrio vulnificus CMCP6
279. Vibrio vulnificus YJ016
280. Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis
281. Wolbachia endosymbiont of Drosophila melanogaster
282. Wolbachia endosymbiont strain TRS of Brugia malayi
283. Wolinella succinogenes DSM 1740

284. Xanthomonas axonopodis pv. citri str. 306
285. Xanthomonas campestris pv. campestris str. 8004
286. Xanthomonas campestris pv. campestris str. ATCC 33913
287. Xanthomonas campestris pv. vesicatoria str. 85-10
288. Xanthomonas oryzae pv. oryzae KACC10331
289. Xylella fastidiosa 9a5c
290. Xylella fastidiosa Temecula1
291. Yersinia pestis biovar Medievalis str. 91001
292. Yersinia pestis CO92
293. Yersinia pestis KIM
294. Yersinia pseudotuberculosis IP 32953

Lisa 2. Programmi Sig.pl tekst

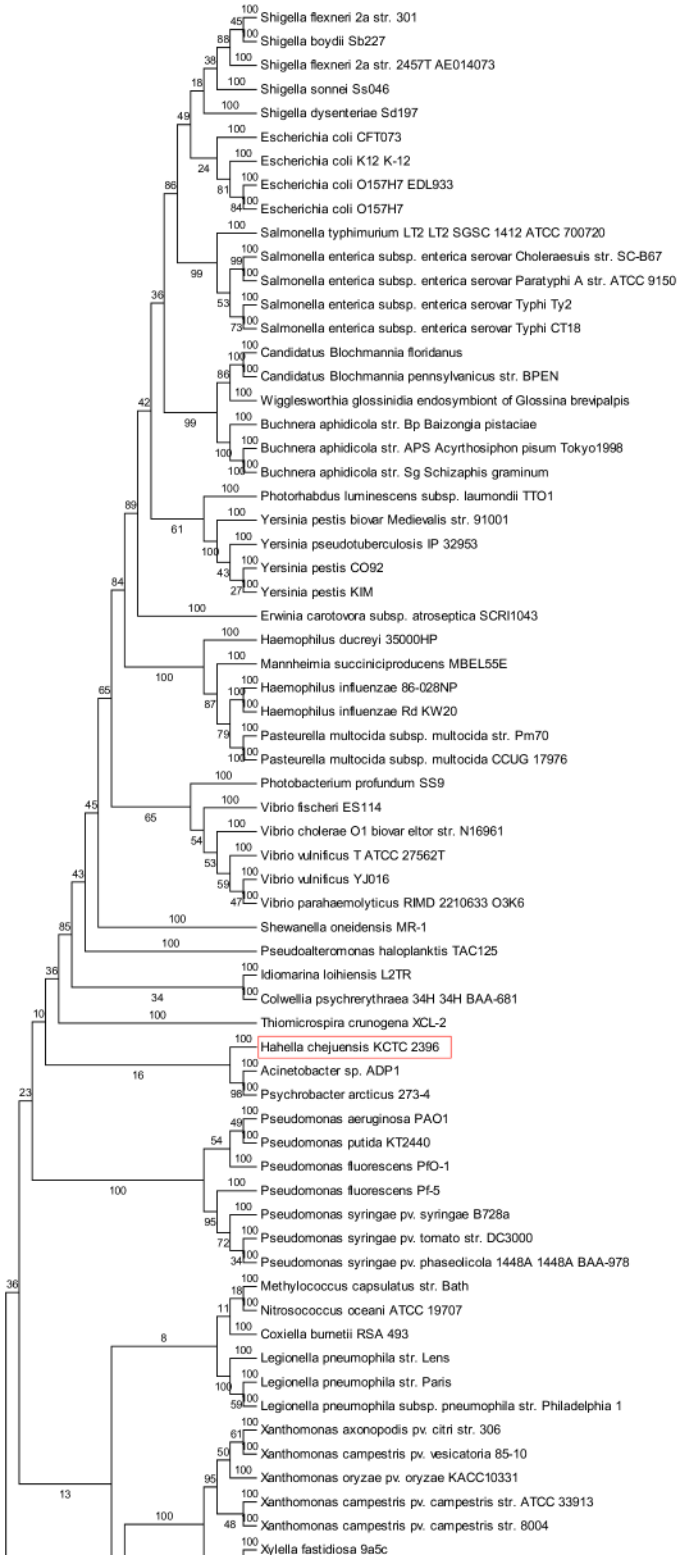
```
#!/usr/bin/perl
#####################################
#
# Author Aleksander Sudakov
# Last edited 13.08.2006
# Version 11
#
#
#
#####################################

use strict;
use warnings;

# MODULES

use Statistics::Basic::Correlation;
use Statistics::Basic::CoVariance;

# VARIABLES

my $pathtodb = "/storage/db/Bacteria_2006/";
my $window = 1000;
my $step = 500;
my $width = 6;
my $smallstep = 1;
my %emptysig = sighash($width); # initialize an empty hash

my @args = @ARGV;
unless ($args[0] && $args[1]) {
    @args = (0, 502);
}

my $dirlist = "keys4b.csv"; # tab-delimited spreadsheet, file must contain
                            # folder name in second column, species name in 3rd
                            # and file names separated by commas in 4th column
                            # for example
                            # B Acinetobacter_sp_ADP1/  Acinetobacter sp. ADP1  NC_005966

# this part is responsible for translating a csv file (from keys4.ods) to get paths, species names and NC_indexes
die "Can't open $dirlist" unless open (DIRLIST, $dirlist);
my $n = 0; # counter
my @paths; # will contain info of species per row

while(<DIRLIST>) {
  $paths[$n] = [parsepath($_)];
  # now we have an array 0-folder,1-name, 2-A or B, 3..n-NC number
  $n++;
}

close DIRLIST;

#-----------------------------------------------------

# now we need to get full path to every fna file
# how many are there?

$n = 0;
my @replicons;

foreach my $row (@paths) {
  my $j = 3;
  # one species has many replicons
  while ($row->[$j]) {
    $replicons[$n] = [$pathtodb.$row->[0], $row->[1], $row->[2], $row->[$j]];
    $j++;
    $n++;
  }
}

#######################################################################

my %sigs;

for (my $n = $args[0]; $n <= $args[1]; $n++) {
  my $row = $replicons[$n];
  close CSV;

  open(CSV, ">CSV/$$row[3].csv"); #result file
  my $filename = $$row[0].$$row[3];
  print "$n: $filename\n";
  $sigs{$$row[3]} = gethash($filename); #kogub
  #print "$$row[3]\n";
}


#######################################################################

sub gethash {

  my $filename = $_[0];
  chomp($filename);

  my $n = 0;
  my @result;
```

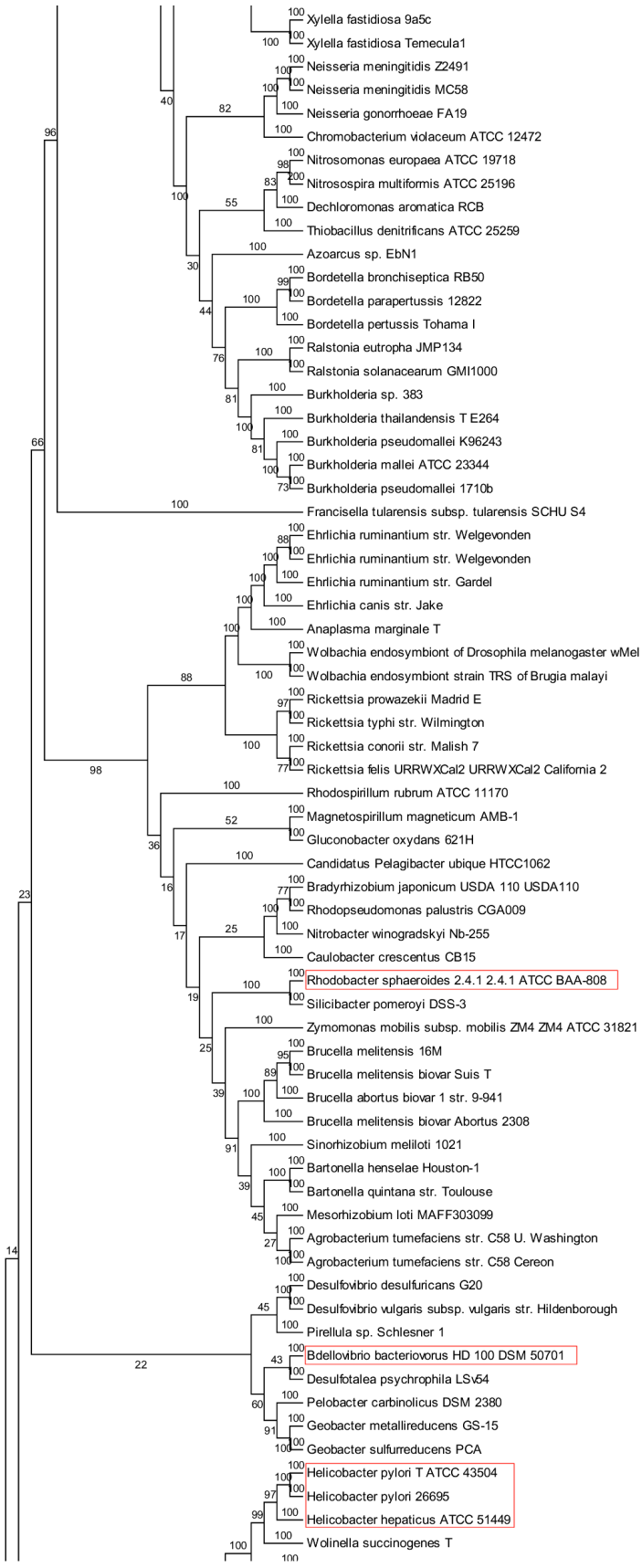
```
 96:   my $subseq;
 97:   my (@vector1, @vector2);
 98:   my $m = 0;
 99:   my $sequence = readfna("$filename.fna");
100:   my $length = length($$sequence);
101:   if ($length == 0) {
102:       return 0;
103:   }
104:
105:   print CSV "Number\tCovariance\tCorrelation\tShort name\tSequence $filename is $length nt long\n";
106:
107:   $result[0] = signature($sequence);
108:   foreach my $key (keys %{$result[0]}) {
109:     $vector1[$m] = ${$result[0]}{$key};
110:     $m++;
111:   }
112:
113:   my $ptt = readpttrnt("$filename.ptt"); #annotation files
114:   my $rnt = readpttrnt("$filename.rnt");
115:
116:   my $pttl = 0; # counters
117:   my $rntl = 0;
118:
119:   while ($n*$step < $length) {
120:
121:     if (($n*$step+$window) > $length) {
122:       $subseq = substr($$sequence, $n*$step, ($length - $n*$step));
123:     } else {
124:       $subseq = substr($$sequence, $n*$step, $window);
125:     }
126:     $n++;
127:     $m = 0;
128:
129:     my $hashref = signature(\$subseq); #get signature for this substring
130:
131:     # converting hash of signature to array vectors
132:     foreach my $key (keys %{$result[0]}) {
133:       $vector2[$m] = $hashref->{$key};
134:       $m++;
135:     }
136:
137:     # See if there is a record in annotation file
138:     while (
139:       $ptt->[$pttl] &&
140:       ($ptt->[$pttl][0] < ($n*$step+$window))
141:       ) {
142:         if (
143:           ($ptt->[$pttl][0] > $n*$step) ||
144:           ($ptt->[$pttl][1] > $n*$step)
145:           ) {
146:             $result[$n][3] .= ", " if ($result[$n][3]);
147:             $result[$n][4] .= ", " if ($result[$n][4]);
148:
149:             $result[$n][3] .= "$ptt->[$pttl][2]";
150:             $result[$n][4] .= "$ptt->[$pttl][3]";
151:       }
152:       $pttl++;
153:     }
154:     # reset counter
155:     if ($pttl > 0) {
156:       if ($pttl < 5) {
157:         $pttl = 0;
158:       } else {
159:         $pttl -= 5;
160:       }
161:     }
162:     # annotation 2
163:     while (
164:       $rnt &&
165:       $rnt->[$rntl] &&
166:       ($rnt->[$rntl][0] < ($n*$step+$window))
167:       ) {
168:         if (
169:           ($rnt->[$rntl][0] > $n*$step) ||
170:           ($rnt->[$rntl][1] > $n*$step)
171:           ) {
172:             $result[$n][3] .= ", " if ($result[$n][3]);
173:             $result[$n][4] .= ", " if ($result[$n][4]);
174:
175:             $result[$n][3] .= "$rnt->[$rntl][2]";
176:             $result[$n][4] .= "$rnt->[$rntl][3]";
177:             $result[$n][6] = "R";
178:       }
179:       $rntl++;
180:     }
181:     #reset counter
182:     if ($rntl > 0) {
183:       if ($rntl < 5) {
184:         $rntl = 0;
185:       } else {
186:         $rntl -= 5;
187:       }
188:     }
189:
190:
```

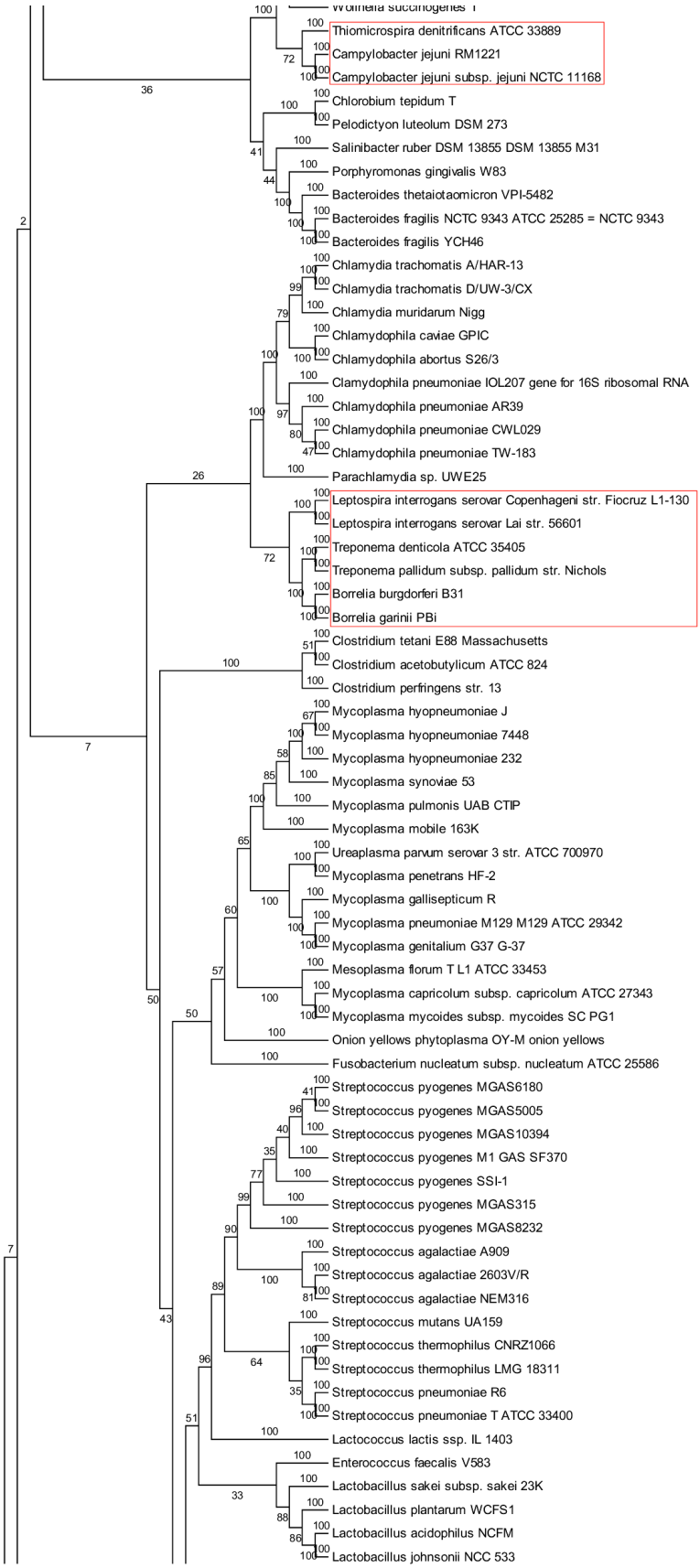
```
191:     # get statistic value (correlation or covariance)
192:     # from statistics package by Jettero Heller
193:
194:     $m = new Statistics::Basic::CoVariance( \@vector1, \@vector2 );
195:     $result[$n][0] = $m->query; # covariance distance to signature(\$subseq);
196:
197:     $m = new Statistics::Basic::Correlation( \@vector1, \@vector2 );
198:     $result[$n][5] = $m->query; # correlation distance to signature(\$subseq);
199:
200:     $result[$n][2] = ($n*$step+$window);
201:
202:     # see that the values are initialized
203:     unless ($result[$n][3] && $result[$n][4]) {
204:       $result[$n][3] = "";
205:       $result[$n][4] = "";
206:     }
207:     unless ($result[$n][6]) {
208:       $result[$n][6] = "";
209:     }
210:     print CSV "$n\t$result[$n][0]\t$result[$n][5]\t$result[$n][3]\t$result[$n][4]\t$result[$n][6]\n";
211:    }
212:
213:   return $result[0]; # signature value for whole sequence
214: }
215:
216: sub parsepath {
217:   chomp; # remove trailing \n   #chomp; # remove trailing \n   #chomp; # remove trailing \n   #chomp; # remove trailing \n
218:   my $path = $_[0];
219:
220:   # returns an array 0-folder,1-name, 2-A or B, 3..n-NC number
221:   my @return;
222:   my @row = split (/\t/, $path);
223:
224:    chop($row[3]); #required on UNIX machines
225:   my @nc = split (/\,/, $row[3]);
226:   @return = ($row[1], $row[2], $row[0], @nc);
227:   #print "@return\n";
228:   return @return;
229: }
230:
231: sub signature {
232:   # returns a hash with a signature for a DNA sequence
233:   my $sequence = $_[0];
234:   my %signature = %emptysig; # initialize an empty hash
235:   my $n = 0; # counter
236:   my $total = 0; # total number of analyzed motifs
237:   my $length = length(${$sequence});
238:   for ($n = 0; $n < ($length-$width+1); $n += $smallstep) {
239:     my $substring = substr(${$sequence}, $n, $width);
240:
241:     # sense and antisense
242:     $signature{$substring} += 1;
243:     $signature{reverse($substring)} += 1;
244:
245:     # sense and antisense - 2, else 1
246:     $total += 2;
247:   }
248:
249:   foreach my $key (keys %signature) {
250:     # to get even proportions on different length fragments
251:     unless ($total == 0) {
252:       $signature{$key} = ($signature{$key}/$total)
253:     }
254:   }
255:
256: return \%signature;
257: }
258:
259: sub sighash {
260:   # uses subroutine addagct to initialize an empty signature hash
261:   my $count = $_[0]; # length of signature word
262:   my %signature;
263:   foreach my $element (addagct($count)) { # gets all combinations
264:     $signature{$element} = "0";
265:   }
266:   return %signature;
267: }
268:
269: sub addagct {
270:   # creates an array of all combinations of four nucleotides
271:
272:   my $count = $_[0];
273:   my @old;
274:   my @new;
275:   for (my $n = 0; $n < $count; $n++) {
276:     if (@new) {
277:       @old = @new;
278:       @new = ();
279:       foreach my $element (@old) {
280:         push (@new, $element."A");
281:         push (@new, $element."G");
282:         push (@new, $element."C");
283:         push (@new, $element."T");
284:         }
285:       } else {
```

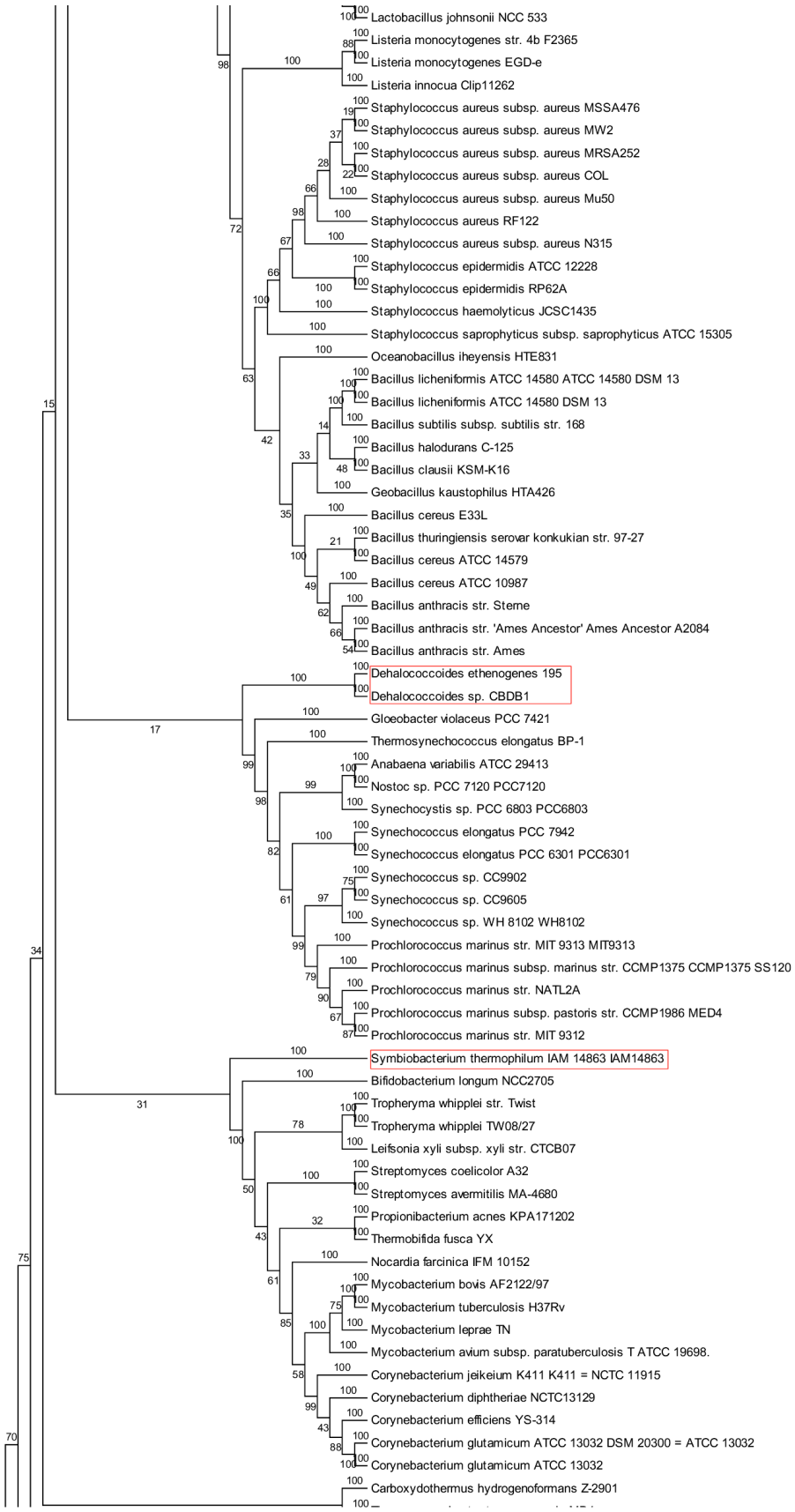
```
286:       @new = ('A', 'G', 'C', 'T');
287:       }
288:     }
289:   return @new;
290: }
291:
292: sub readfna {
293:   # subroutine to get whole sequence in one file without spaces
294:   my $a = "";
295:   unless (open (AFILE, $_[0])) {
296:     print "ERROR on readfile $_[0]!\n";
297:     return 0;
298:   }
299:   $_ = <AFILE>; # no need for title row
300:   while (<AFILE>) {
301:   s/\n//g; # remove all new lines
302:   $a .= $_; # concatenate
303:   }
304:   close (AFILE);
305: return \$a; # returns a reference so memory is used less
306: }
307:
308: sub readpttrnt {
309:   # reads .ptt and ,rnt genbank files and extract data
310:   my @return;
311:   unless (open (AFILE, $_[0])) {
312:     return 0;
313:   }
314:   $_ = <AFILE>;
315:   $_ = <AFILE>;
316:   $_ = <AFILE>; # three first rows are of no use
317:   my $n = 0;
318:   while (<AFILE>) {
319:     chomp;
320:     my @line = split(/\t/);
321:     $return[$n] = [split(/\.\./, $line[0]), $line[4], $line[8]];
322:     # extracts gene start, end coordinates, short and long name of gene
323:     $n++;
324:   }
325:   close (AFILE);
326: return \@return;
327: }
328:
```

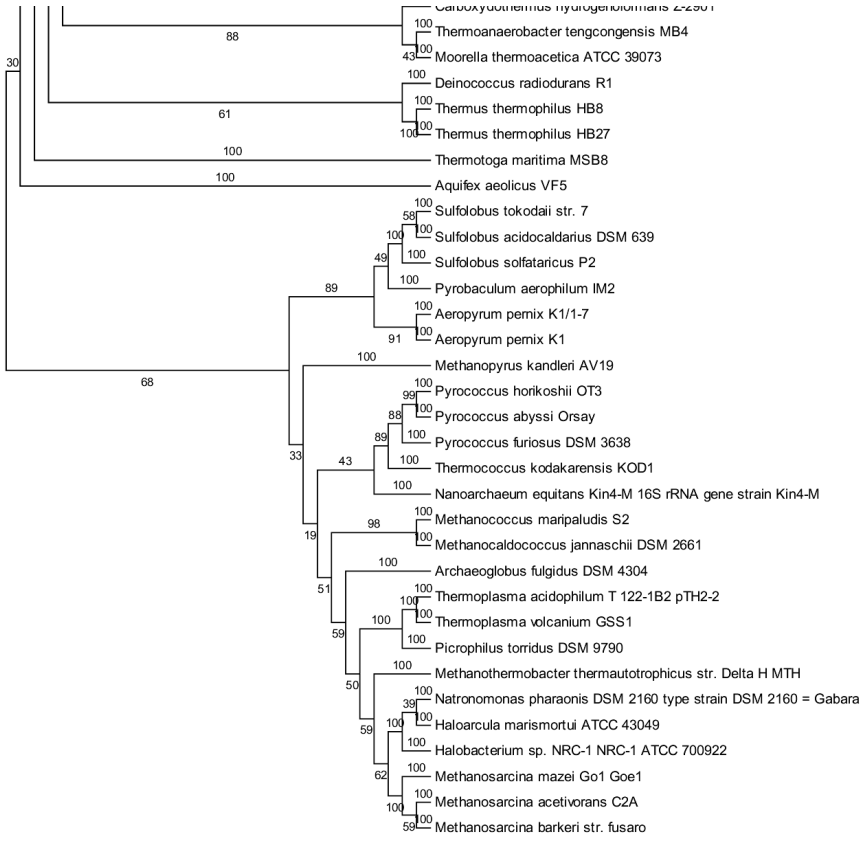
Lisa 3. 16S rRNA järjestuse *maximum likelihood* fülogeneesipuu *bootstrap* väärtustega



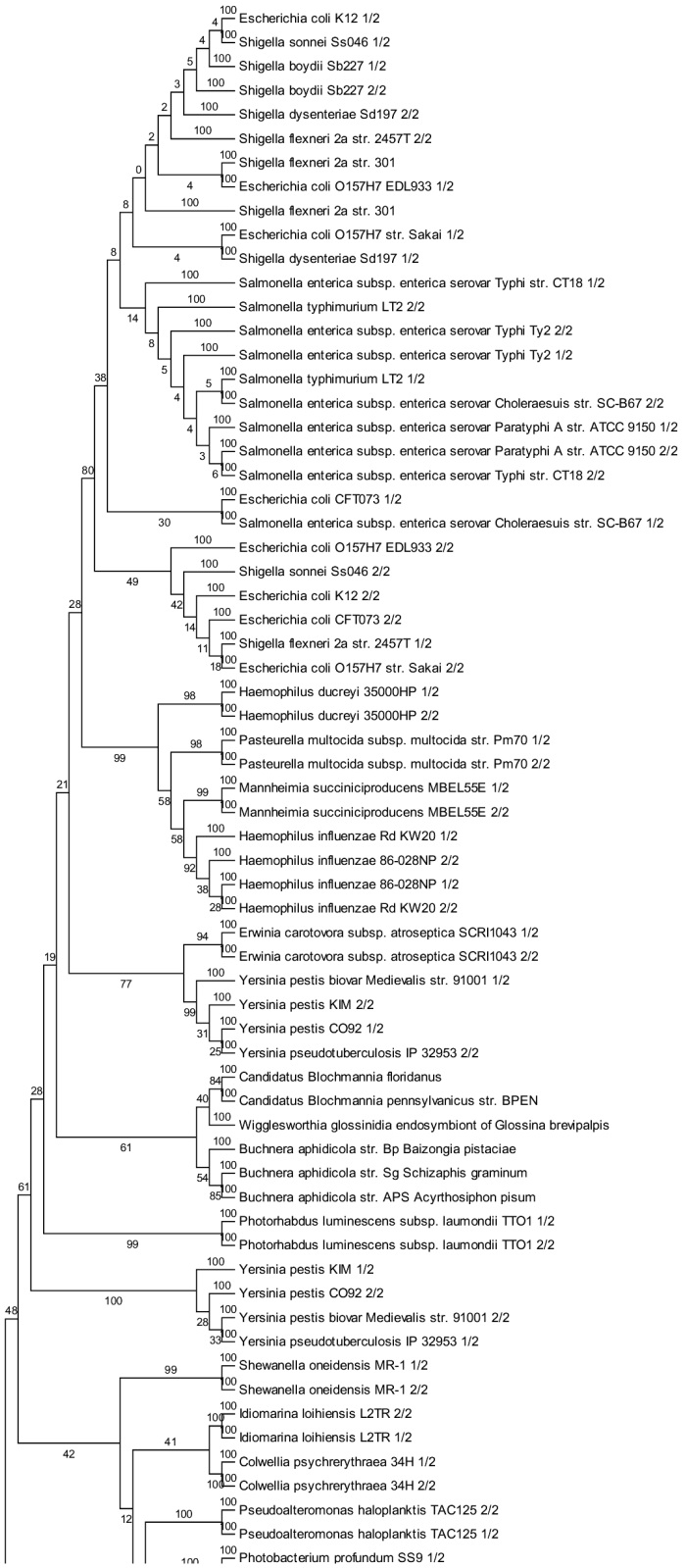

Xylella fastidiosa 9a5c
Xylella fastidiosa Temecula1
Neisseria meningitidis Z2491
Neisseria meningitidis MC58
Neisseria gonorrhoeae FA19
Chromobacterium violaceum ATCC 12472
Nitrosomonas europaea ATCC 19718
Nitrosospira multiformis ATCC 25196
Dechloromonas aromatica RCB
Thiobacillus denitrificans ATCC 25259
Azoarcus sp. EbN1
Bordetella bronchiseptica RB50
Bordetella parapertussis 12822
Bordetella pertussis Tohama I
Ralstonia eutropha JMP134
Ralstonia solanacearum GMI1000
Burkholderia sp. 383
Burkholderia thailandensis T E264
Burkholderia pseudomallei K96243
Burkholderia mallei ATCC 23344
Burkholderia pseudomallei 1710b
Francisella tularensis subsp. tularensis SCHU S4
Ehrlichia ruminantium str. Welgevonden
Ehrlichia ruminantium str. Welgevonden
Ehrlichia ruminantium str. Gardel
Ehrlichia canis str. Jake
Anaplasma marginale T
Wolbachia endosymbiont of Drosophila melanogaster wMel
Wolbachia endosymbiont strain TRS of Brugia malayi
Rickettsia prowazekii Madrid E
Rickettsia typhi str. Wilmington
Rickettsia conorii str. Malish 7
Rickettsia felis URRWXCal2 URRWXCal2 California 2
Rhodospirillum rubrum ATCC 11170
Magnetospirillum magneticum AMB-1
Gluconobacter oxydans 621H
Candidatus Pelagibacter ubique HTCC1062
Bradyrhizobium japonicum USDA 110 USDA110
Rhodopseudomonas palustris CGA009
Nitrobacter winogradskyi Nb-255
Caulobacter crescentus CB15
Rhodobacter sphaeroides 2.4.1 2.4.1 ATCC BAA-808
Silicibacter pomeroyi DSS-3
Zymomonas mobilis subsp. mobilis ZM4 ZM4 ATCC 31821
Brucella melitensis 16M
Brucella melitensis biovar Suis T
Brucella abortus biovar 1 str. 9-941
Brucella melitensis biovar Abortus 2308
Sinorhizobium meliloti 1021
Bartonella henselae Houston-1
Bartonella quintana str. Toulouse
Mesorhizobium loti MAFF303099
Agrobacterium tumefaciens str. C58 U. Washington
Agrobacterium tumefaciens str. C58 Cereon
Desulfovibrio desulfuricans G20
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
Pirellula sp. Schlesner 1
Bdellovibrio bacteriovorus HD 100 DSM 50701
Desulfotalea psychrophila LSv54
Pelobacter carbinolicus DSM 2380
Geobacter metallireducens GS-15
Geobacter sulfurreducens PCA
Helicobacter pylori T ATCC 43504
Helicobacter pylori 26695
Helicobacter hepaticus ATCC 51449
Wolinella succinogenes T

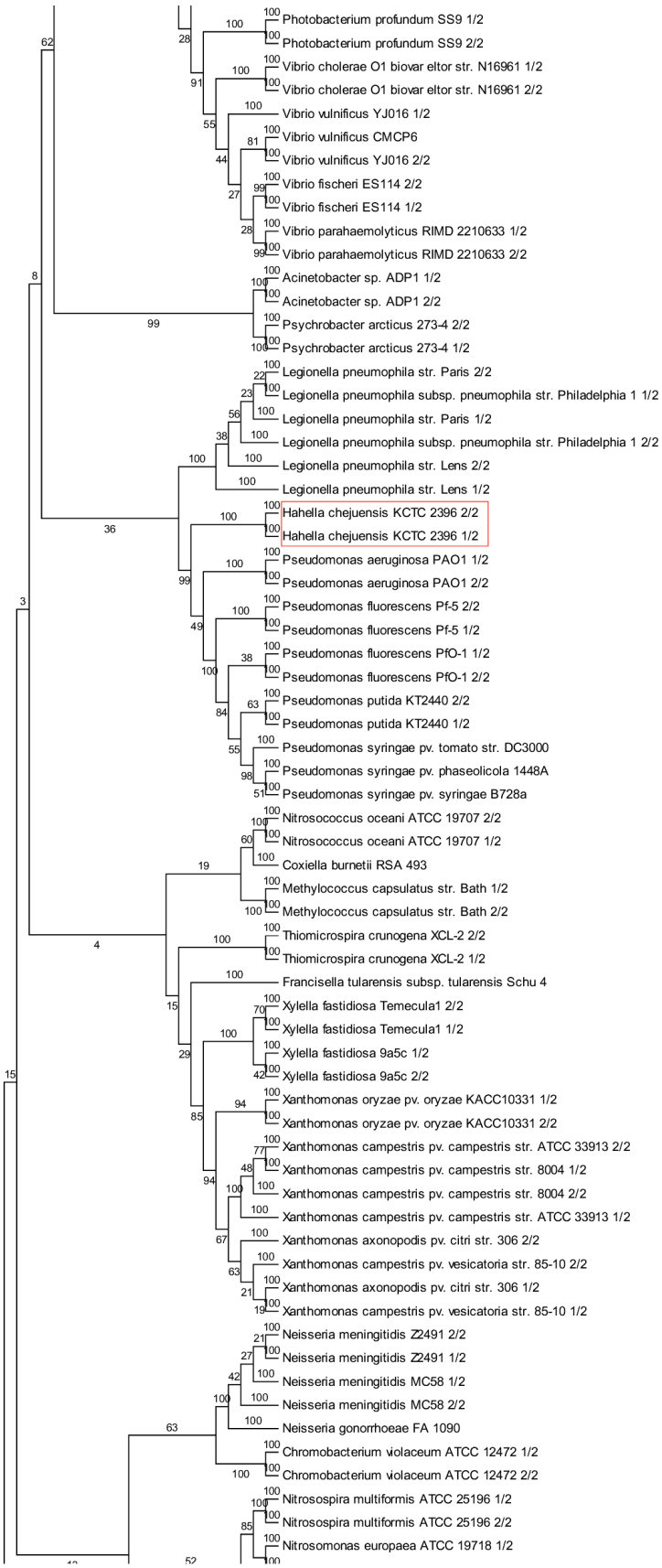

Wolinella succinogenes T
Thiomicrospira denitrificans ATCC 33889
Campylobacter jejuni RM1221
Campylobacter jejuni subsp. jejuni NCTC 11168
Chlorobium tepidum T
Pelodictyon luteolum DSM 273
Salinibacter ruber DSM 13855 DSM 13855 M31
Porphyromonas gingivalis W83
Bacteroides thetaiotaomicron VPI-5482
Bacteroides fragilis NCTC 9343 ATCC 25285 = NCTC 9343
Bacteroides fragilis YCH46
Chlamydia trachomatis A/HAR-13
Chlamydia trachomatis D/UW-3/CX
Chlamydia muridarum Nigg
Chlamydophila caviae GPIC
Chlamydophila abortus S26/3
Clamydophila pneumoniae IOL207 gene for 16S ribosomal RNA
Chlamydophila pneumoniae AR39
Chlamydophila pneumoniae CWL029
Chlamydophila pneumoniae TW-183
Parachlamydia sp. UWE25
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
Leptospira interrogans serovar Lai str. 56601
Treponema denticola ATCC 35405
Treponema pallidum subsp. pallidum str. Nichols
Borrelia burgdorferi B31
Borrelia garinii PBi
Clostridium tetani E88 Massachusetts
Clostridium acetobutylicum ATCC 824
Clostridium perfringens str. 13
Mycoplasma hyopneumoniae J
Mycoplasma hyopneumoniae 7448
Mycoplasma hyopneumoniae 232
Mycoplasma synoviae 53
Mycoplasma pulmonis UAB CTIP
Mycoplasma mobile 163K
Ureaplasma parvum serovar 3 str. ATCC 700970
Mycoplasma penetrans HF-2
Mycoplasma gallisepticum R
Mycoplasma pneumoniae M129 M129 ATCC 29342
Mycoplasma genitalium G37 G-37
Mesoplasma florum T L1 ATCC 33453
Mycoplasma capricolum subsp. capricolum ATCC 27343
Mycoplasma mycoides subsp. mycoides SC PG1
Onion yellows phytoplasma OY-M onion yellows
Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Streptococcus pyogenes MGAS6180
Streptococcus pyogenes MGAS5005
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes M1 GAS SF370
Streptococcus pyogenes SSI-1
Streptococcus pyogenes MGAS315
Streptococcus pyogenes MGAS8232
Streptococcus agalactiae A909
Streptococcus agalactiae 2603V/R
Streptococcus agalactiae NEM316
Streptococcus mutans UA159
Streptococcus thermophilus CNRZ1066
Streptococcus thermophilus LMG 18311
Streptococcus pneumoniae R6
Streptococcus pneumoniae T ATCC 33400
Lactococcus lactis ssp. IL 1403
Enterococcus faecalis V583
Lactobacillus sakei subsp. sakei 23K
Lactobacillus plantarum WCFS1
Lactobacillus acidophilus NCFM
Lactobacillus johnsonii NCC 533

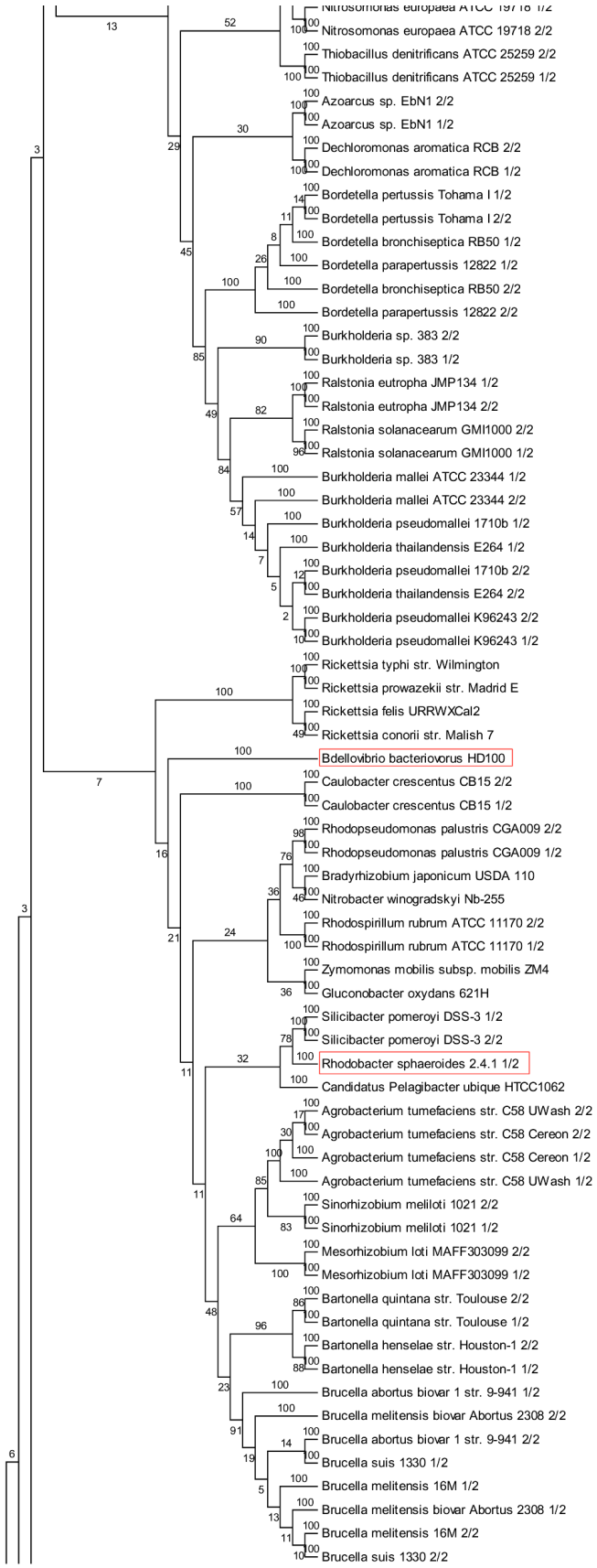

Lactobacillus johnsonii NCC 533
Listeria monocytogenes str. 4b F2365
Listeria monocytogenes EGD-e
Listeria innocua Clip11262
Staphylococcus aureus subsp. aureus MSSA476
Staphylococcus aureus subsp. aureus MW2
Staphylococcus aureus subsp. aureus MRSA252
Staphylococcus aureus subsp. aureus COL
Staphylococcus aureus subsp. aureus Mu50
Staphylococcus aureus RF122
Staphylococcus aureus subsp. aureus N315
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Staphylococcus haemolyticus JCSC1435
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
Oceanobacillus iheyensis HTE831
Bacillus licheniformis ATCC 14580 ATCC 14580 DSM 13
Bacillus licheniformis ATCC 14580 DSM 13
Bacillus subtilis subsp. subtilis str. 168
Bacillus halodurans C-125
Bacillus clausii KSM-K16
Geobacillus kaustophilus HTA426
Bacillus cereus E33L
Bacillus thuringiensis serovar konkukian str. 97-27
Bacillus cereus ATCC 14579
Bacillus cereus ATCC 10987
Bacillus anthracis str. Sterne
Bacillus anthracis str. 'Ames Ancestor' Ames Ancestor A2084
Bacillus anthracis str. Ames
Dehalococcoides ethenogenes 195
Dehalococcoides sp. CBDB1
Gloeobacter violaceus PCC 7421
Thermosynechococcus elongatus BP-1
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7120 PCC7120
Synechocystis sp. PCC 6803 PCC6803
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301 PCC6301
Synechococcus sp. CC9902
Synechococcus sp. CC9605
Synechococcus sp. WH 8102 WH8102
Prochlorococcus marinus str. MIT 9313 MIT9313
Prochlorococcus marinus subsp. marinus str. CCMP1375 CCMP1375 SS120
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus subsp. pastoris str. CCMP1986 MED4
Prochlorococcus marinus str. MIT 9312
Symbiobacterium thermophilum IAM 14863 IAM14863
Bifidobacterium longum NCC2705
Tropheryma whipplei str. Twist
Tropheryma whipplei TW08/27
Leifsonia xyli subsp. xyli str. CTCB07
Streptomyces coelicolor A32
Streptomyces avermitilis MA-4680
Propionibacterium acnes KPA171202
Thermobifida fusca YX
Nocardia farcinica IFM 10152
Mycobacterium bovis AF2122/97
Mycobacterium tuberculosis H37Rv
Mycobacterium leprae TN
Mycobacterium avium subsp. paratuberculosis T ATCC 19698.
Corynebacterium jeikeium K411 K411 = NCTC 11915
Corynebacterium diphtheriae NCTC13129
Corynebacterium efficiens YS-314
Corynebacterium glutamicum ATCC 13032 DSM 20300 = ATCC 13032
Corynebacterium glutamicum ATCC 13032
Carboxydothermus hydrogenoformans Z-2901

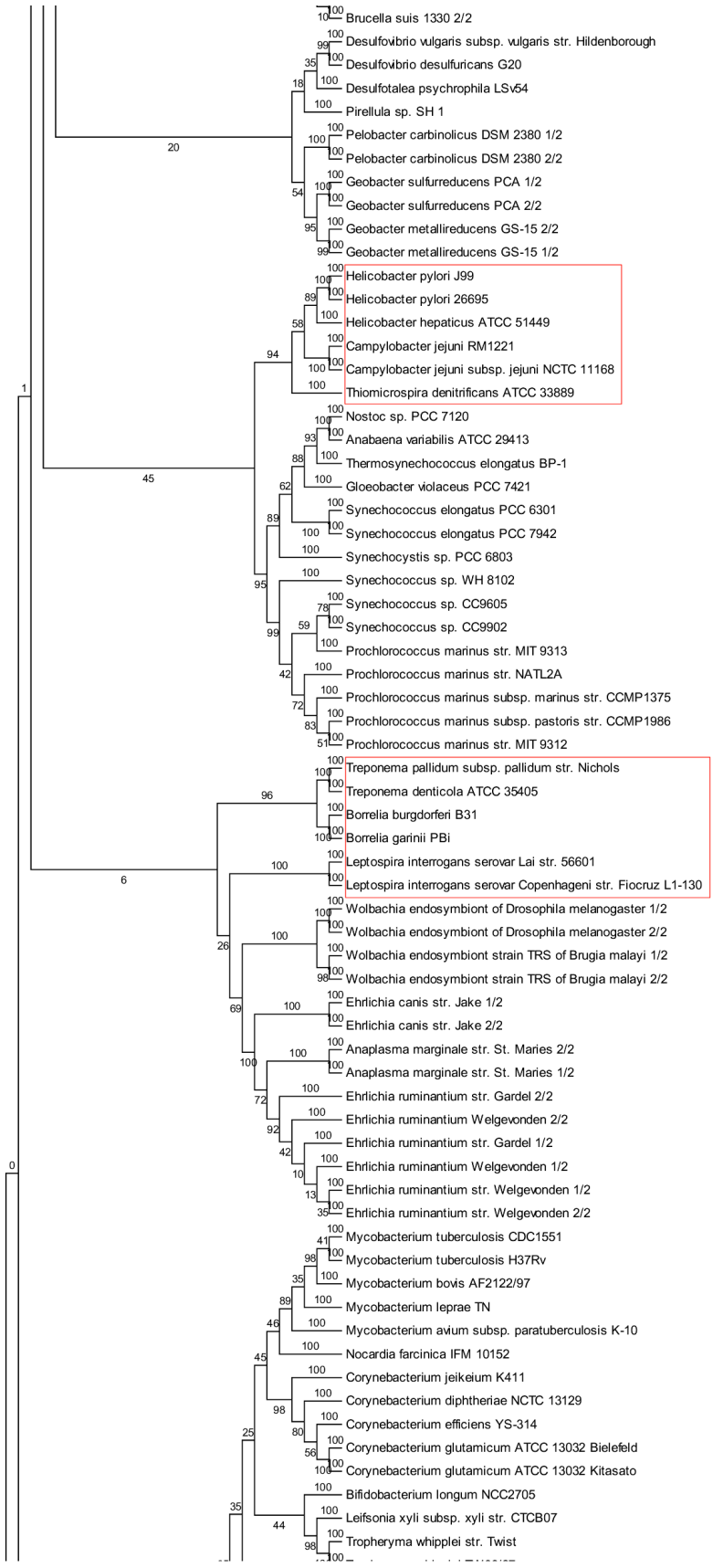

Thermoanaerobacter tengcongensis MB4
Moorella thermoacetica ATCC 39073
Deinococcus radiodurans R1
Thermus thermophilus HB8
Thermus thermophilus HB27
Thermotoga maritima MSB8
Aquifex aeolicus VF5
Sulfolobus tokodaii str. 7
Sulfolobus acidocaldarius DSM 639
Sulfolobus solfataricus P2
Pyrobaculum aerophilum IM2
Aeropyrum pernix K1/1-7
Aeropyrum pernix K1
Methanopyrus kandleri AV19
Pyrococcus horikoshii OT3
Pyrococcus abyssi Orsay
Pyrococcus furiosus DSM 3638
Thermococcus kodakarensis KOD1
Nanoarchaeum equitans Kin4-M 16S rRNA gene strain Kin4-M
Methanococcus maripaludis S2
Methanocaldococcus jannaschii DSM 2661
Archaeoglobus fulgidus DSM 4304
Thermoplasma acidophilum T 122-1B2 pTH2-2
Thermoplasma volcanium GSS1
Picrophilus torridus DSM 9790
Methanothermobacter thermautotrophicus str. Delta H MTH
Natronomonas pharaonis DSM 2160 type strain DSM 2160 = Gabara
Haloarcula marismortui ATCC 43049
Halobacterium sp. NRC-1 NRC-1 ATCC 700922
Methanosarcina mazei Go1 Goe1
Methanosarcina acetivorans C2A
Methanosarcina barkeri str. fusaro

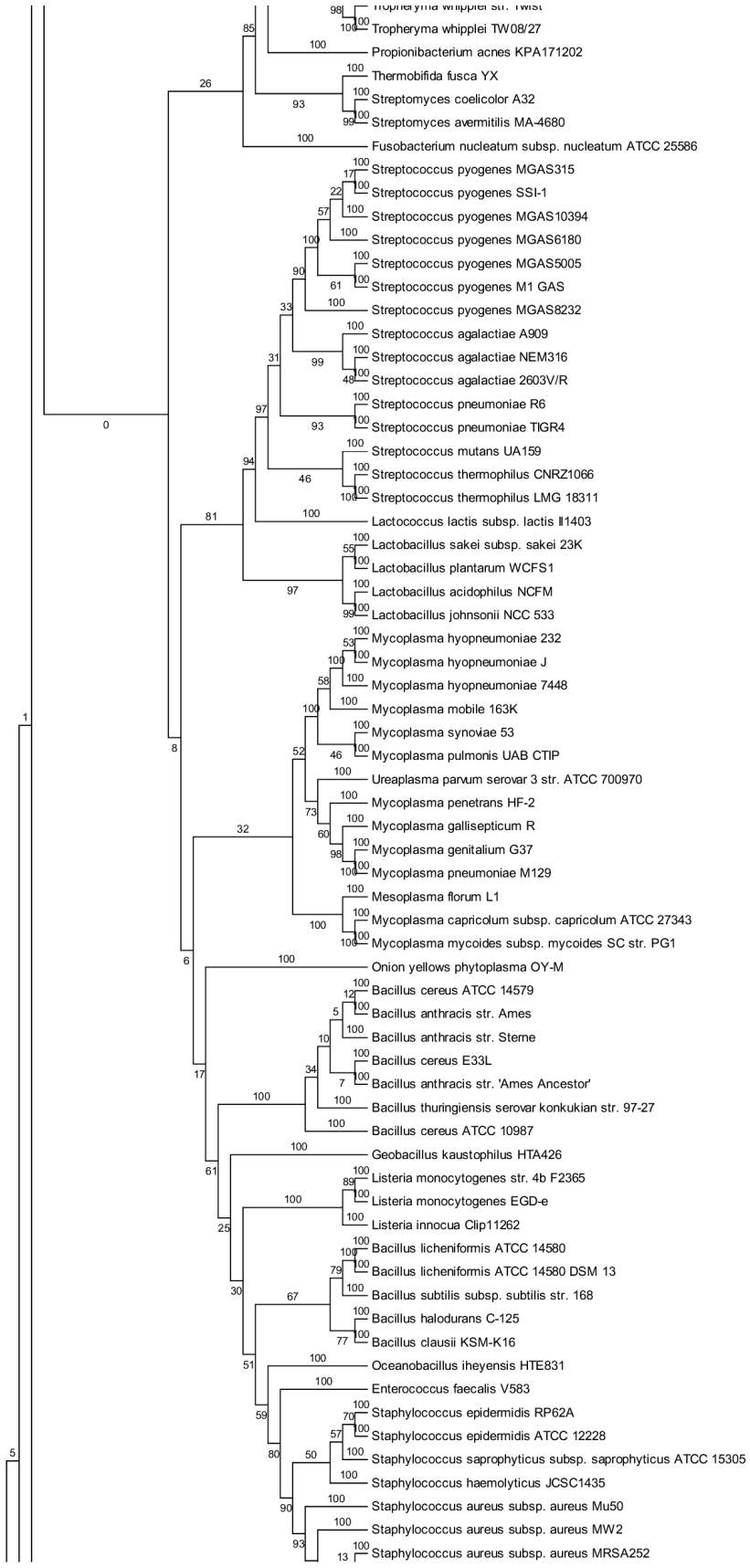
Lisa 4. Ef-Tu valgujärjestuse *maximum likelihood* fülogeneesipuu bootstrap väärtustega

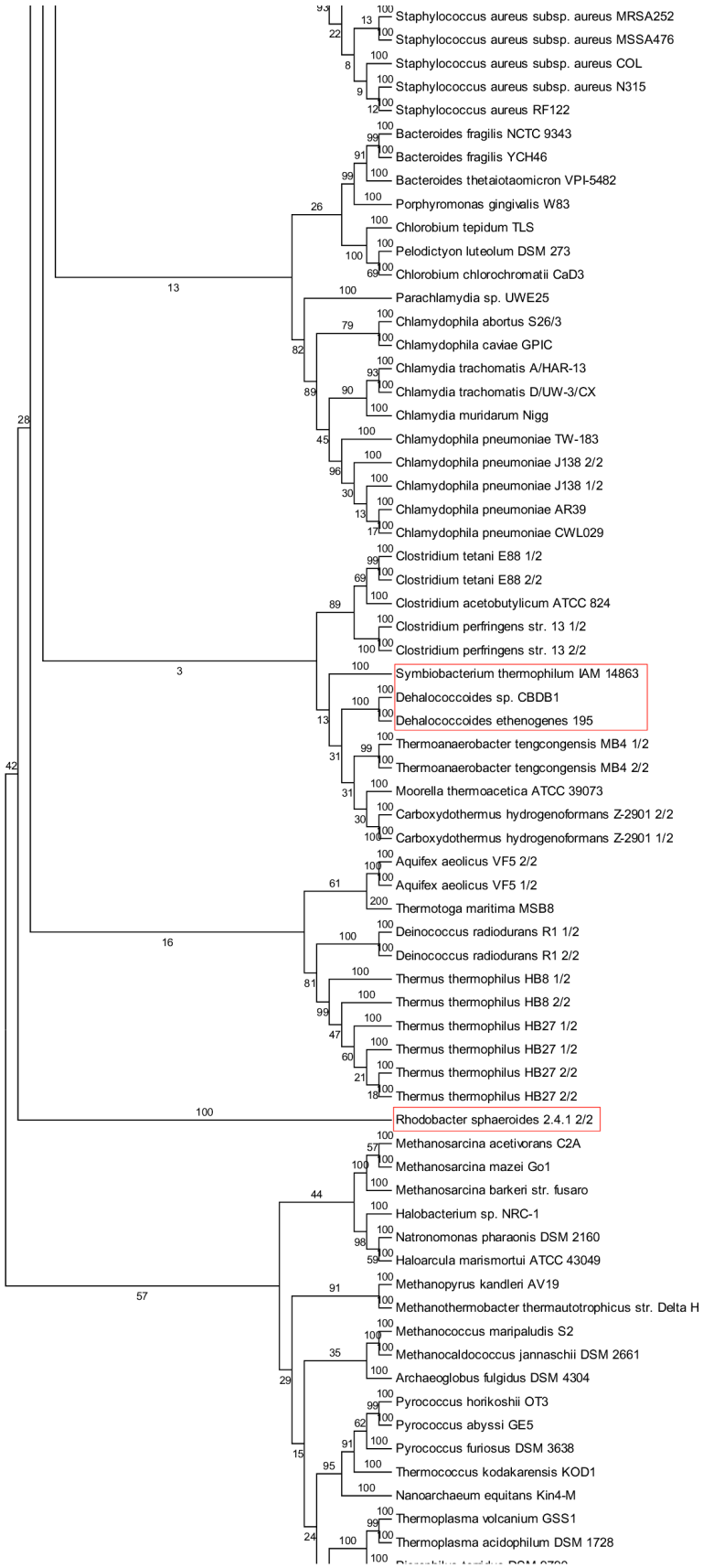


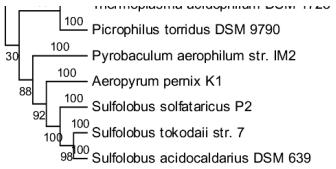

Photobacterium profundum SS9 1/2
Photobacterium profundum SS9 2/2
Vibrio cholerae O1 biovar eltor str. N16961 1/2
Vibrio cholerae O1 biovar eltor str. N16961 2/2
Vibrio vulnificus YJ016 1/2
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016 2/2
Vibrio fischeri ES114 2/2
Vibrio fischeri ES114 1/2
Vibrio parahaemolyticus RIMD 2210633 1/2
Vibrio parahaemolyticus RIMD 2210633 2/2
Acinetobacter sp. ADP1 1/2
Acinetobacter sp. ADP1 2/2
Psychrobacter arcticus 273-4 2/2
Psychrobacter arcticus 273-4 1/2
Legionella pneumophila str. Paris 2/2
Legionella pneumophila subsp. pneumophila str. Philadelphia 1 1/2
Legionella pneumophila str. Paris 1/2
Legionella pneumophila subsp. pneumophila str. Philadelphia 1 2/2
Legionella pneumophila str. Lens 2/2
Legionella pneumophila str. Lens 1/2
Hahella chejuensis KCTC 2396 2/2
Hahella chejuensis KCTC 2396 1/2
Pseudomonas aeruginosa PAO1 1/2
Pseudomonas aeruginosa PAO1 2/2
Pseudomonas fluorescens Pf-5 2/2
Pseudomonas fluorescens Pf-5 1/2
Pseudomonas fluorescens PfO-1 1/2
Pseudomonas fluorescens PfO-1 2/2
Pseudomonas putida KT2440 2/2
Pseudomonas putida KT2440 1/2
Pseudomonas syringae pv. tomato str. DC3000
Pseudomonas syringae pv. phaseolicola 1448A
Pseudomonas syringae pv. syringae B728a
Nitrosococcus oceani ATCC 19707 2/2
Nitrosococcus oceani ATCC 19707 1/2
Coxiella burnetii RSA 493
Methylococcus capsulatus str. Bath 1/2
Methylococcus capsulatus str. Bath 2/2
Thiomicrospira crunogena XCL-2 2/2
Thiomicrospira crunogena XCL-2 1/2
Francisella tularensis subsp. tularensis Schu 4
Xylella fastidiosa Temecula1 2/2
Xylella fastidiosa Temecula1 1/2
Xylella fastidiosa 9a5c 1/2
Xylella fastidiosa 9a5c 2/2
Xanthomonas oryzae pv. oryzae KACC10331 1/2
Xanthomonas oryzae pv. oryzae KACC10331 2/2
Xanthomonas campestris pv. campestris str. ATCC 33913 2/2
Xanthomonas campestris pv. campestris str. 8004 1/2
Xanthomonas campestris pv. campestris str. 8004 2/2
Xanthomonas campestris pv. campestris str. ATCC 33913 1/2
Xanthomonas axonopodis pv. citri str. 306 2/2
Xanthomonas campestris pv. vesicatoria str. 85-10 2/2
Xanthomonas axonopodis pv. citri str. 306 1/2
Xanthomonas campestris pv. vesicatoria str. 85-10 1/2
Neisseria meningitidis Z2491 2/2
Neisseria meningitidis Z2491 1/2
Neisseria meningitidis MC58 1/2
Neisseria meningitidis MC58 2/2
Neisseria gonorrhoeae FA 1090
Chromobacterium violaceum ATCC 12472 1/2
Chromobacterium violaceum ATCC 12472 2/2
Nitrosospira multiformis ATCC 25196 1/2
Nitrosospira multiformis ATCC 25196 2/2
Nitrosomonas europaea ATCC 19718 1/2


Nitrosomonas europaea ATCC 19718 2/2
Thiobacillus denitrificans ATCC 25259 2/2
Thiobacillus denitrificans ATCC 25259 1/2
Azoarcus sp. EbN1 2/2
Azoarcus sp. EbN1 1/2
Dechloromonas aromatica RCB 2/2
Dechloromonas aromatica RCB 1/2
Bordetella pertussis Tohama I 1/2
Bordetella pertussis Tohama I 2/2
Bordetella bronchiseptica RB50 1/2
Bordetella parapertussis 12822 1/2
Bordetella bronchiseptica RB50 2/2
Bordetella parapertussis 12822 2/2
Burkholderia sp. 383 2/2
Burkholderia sp. 383 1/2
Ralstonia eutropha JMP134 1/2
Ralstonia eutropha JMP134 2/2
Ralstonia solanacearum GMI1000 2/2
Ralstonia solanacearum GMI1000 1/2
Burkholderia mallei ATCC 23344 1/2
Burkholderia mallei ATCC 23344 2/2
Burkholderia pseudomallei 1710b 1/2
Burkholderia thailandensis E264 1/2
Burkholderia pseudomallei 1710b 2/2
Burkholderia thailandensis E264 2/2
Burkholderia pseudomallei K96243 2/2
Burkholderia pseudomallei K96243 1/2
Rickettsia typhi str. Wilmington
Rickettsia prowazekii str. Madrid E
Rickettsia felis URRWXCal2
Rickettsia conorii str. Malish 7
Bdellovibrio bacteriovorus HD100
Caulobacter crescentus CB15 2/2
Caulobacter crescentus CB15 1/2
Rhodopseudomonas palustris CGA009 2/2
Rhodopseudomonas palustris CGA009 1/2
Bradyrhizobium japonicum USDA 110
Nitrobacter winogradskyi Nb-255
Rhodospirillum rubrum ATCC 11170 2/2
Rhodospirillum rubrum ATCC 11170 1/2
Zymomonas mobilis subsp. mobilis ZM4
Gluconobacter oxydans 621H
Silicibacter pomeroyi DSS-3 1/2
Silicibacter pomeroyi DSS-3 2/2
Rhodobacter sphaeroides 2.4.1 1/2
Candidatus Pelagibacter ubique HTCC1062
Agrobacterium tumefaciens str. C58 UWash 2/2
Agrobacterium tumefaciens str. C58 Cereon 2/2
Agrobacterium tumefaciens str. C58 Cereon 1/2
Agrobacterium tumefaciens str. C58 UWash 1/2
Sinorhizobium meliloti 1021 2/2
Sinorhizobium meliloti 1021 1/2
Mesorhizobium loti MAFF303099 2/2
Mesorhizobium loti MAFF303099 1/2
Bartonella quintana str. Toulouse 2/2
Bartonella quintana str. Toulouse 1/2
Bartonella henselae str. Houston-1 2/2
Bartonella henselae str. Houston-1 1/2
Brucella abortus biovar 1 str. 9-941 1/2
Brucella melitensis biovar Abortus 2308 2/2
Brucella abortus biovar 1 str. 9-941 2/2
Brucella suis 1330 1/2
Brucella melitensis 16M 1/2
Brucella melitensis biovar Abortus 2308 1/2
Brucella melitensis 16M 2/2
Brucella suis 1330 2/2


Brucella suis 1330 2/2
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
Desulfovibrio desulfuricans G20
Desulfotalea psychrophila LSv54
Pirellula sp. SH 1
Pelobacter carbinolicus DSM 2380 1/2
Pelobacter carbinolicus DSM 2380 2/2
Geobacter sulfurreducens PCA 1/2
Geobacter sulfurreducens PCA 2/2
Geobacter metallireducens GS-15 2/2
Geobacter metallireducens GS-15 1/2
Helicobacter pylori J99
Helicobacter pylori 26695
Helicobacter hepaticus ATCC 51449
Campylobacter jejuni RM1221
Campylobacter jejuni subsp. jejuni NCTC 11168
Thiomicrospira denitrificans ATCC 33889
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Thermosynechococcus elongatus BP-1
Gloeobacter violaceus PCC 7421
Synechococcus elongatus PCC 6301
Synechococcus elongatus PCC 7942
Synechocystis sp. PCC 6803
Synechococcus sp. WH 8102
Synechococcus sp. CC9605
Synechococcus sp. CC9902
Prochlorococcus marinus str. MIT 9313
Prochlorococcus marinus str. NATL2A
Prochlorococcus marinus subsp. marinus str. CCMP1375
Prochlorococcus marinus subsp. pastoris str. CCMP1986
Prochlorococcus marinus str. MIT 9312
Treponema pallidum subsp. pallidum str. Nichols
Treponema denticola ATCC 35405
Borrelia burgdorferi B31
Borrelia garinii PBi
Leptospira interrogans serovar Lai str. 56601
Leptospira interrogans serovar Copenhageni str. Fiocruz L1-130
Wolbachia endosymbiont of Drosophila melanogaster 1/2
Wolbachia endosymbiont of Drosophila melanogaster 2/2
Wolbachia endosymbiont strain TRS of Brugia malayi 1/2
Wolbachia endosymbiont strain TRS of Brugia malayi 2/2
Ehrlichia canis str. Jake 1/2
Ehrlichia canis str. Jake 2/2
Anaplasma marginale str. St. Maries 2/2
Anaplasma marginale str. St. Maries 1/2
Ehrlichia ruminantium str. Gardel 2/2
Ehrlichia ruminantium Welgevonden 2/2
Ehrlichia ruminantium str. Gardel 1/2
Ehrlichia ruminantium Welgevonden 1/2
Ehrlichia ruminantium str. Welgevonden 1/2
Ehrlichia ruminantium str. Welgevonden 2/2
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis H37Rv
Mycobacterium bovis AF2122/97
Mycobacterium leprae TN
Mycobacterium avium subsp. paratuberculosis K-10
Nocardia farcinica IFM 10152
Corynebacterium jeikeium K411
Corynebacterium diphtheriae NCTC 13129
Corynebacterium efficiens YS-314
Corynebacterium glutamicum ATCC 13032 Bielefeld
Corynebacterium glutamicum ATCC 13032 Kitasato
Bifidobacterium longum NCC2705
Leifsonia xyli subsp. xyli str. CTCB07
Tropheryma whipplei str. Twist


Tropheryma whipplei TW08/27
Propionibacterium acnes KPA171202
Thermobifida fusca YX
Streptomyces coelicolor A32
Streptomyces avermitilis MA-4680
Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Streptococcus pyogenes MGAS315
Streptococcus pyogenes SSI-1
Streptococcus pyogenes MGAS10394
Streptococcus pyogenes MGAS6180
Streptococcus pyogenes MGAS5005
Streptococcus pyogenes M1 GAS
Streptococcus pyogenes MGAS8232
Streptococcus agalactiae A909
Streptococcus agalactiae NEM316
Streptococcus agalactiae 2603V/R
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus mutans UA159
Streptococcus thermophilus CNRZ1066
Streptococcus thermophilus LMG 18311
Lactococcus lactis subsp. lactis Il1403
Lactobacillus sakei subsp. sakei 23K
Lactobacillus plantarum WCFS1
Lactobacillus acidophilus NCFM
Lactobacillus johnsonii NCC 533
Mycoplasma hyopneumoniae 232
Mycoplasma hyopneumoniae J
Mycoplasma hyopneumoniae 7448
Mycoplasma mobile 163K
Mycoplasma synoviae 53
Mycoplasma pulmonis UAB CTIP
Ureaplasma parvum serovar 3 str. ATCC 700970
Mycoplasma penetrans HF-2
Mycoplasma gallisepticum R
Mycoplasma genitalium G37
Mycoplasma pneumoniae M129
Mesoplasma florum L1
Mycoplasma capricolum subsp. capricolum ATCC 27343
Mycoplasma mycoides subsp. mycoides SC str. PG1
Onion yellows phytoplasma OY-M
Bacillus cereus ATCC 14579
Bacillus anthracis str. Ames
Bacillus anthracis str. Sterne
Bacillus cereus E33L
Bacillus anthracis str. 'Ames Ancestor'
Bacillus thuringiensis serovar konkukian str. 97-27
Bacillus cereus ATCC 10987
Geobacillus kaustophilus HTA426
Listeria monocytogenes str. 4b F2365
Listeria monocytogenes EGD-e
Listeria innocua Clip11262
Bacillus licheniformis ATCC 14580
Bacillus licheniformis ATCC 14580 DSM 13
Bacillus subtilis subsp. subtilis str. 168
Bacillus halodurans C-125
Bacillus clausii KSM-K16
Oceanobacillus iheyensis HTE831
Enterococcus faecalis V583
Staphylococcus epidermidis RP62A
Staphylococcus epidermidis ATCC 12228
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
Staphylococcus haemolyticus JCSC1435
Staphylococcus aureus subsp. aureus Mu50
Staphylococcus aureus subsp. aureus MW2
Staphylococcus aureus subsp. aureus MRSA252


Staphylococcus aureus subsp. aureus MRSA252
Staphylococcus aureus subsp. aureus MSSA476
Staphylococcus aureus subsp. aureus COL
Staphylococcus aureus subsp. aureus N315
Staphylococcus aureus RF122
Bacteroides fragilis NCTC 9343
Bacteroides fragilis YCH46
Bacteroides thetaiotaomicron VPI-5482
Porphyromonas gingivalis W83
Chlorobium tepidum TLS
Pelodictyon luteolum DSM 273
Chlorobium chlorochromatii CaD3
Parachlamydia sp. UWE25
Chlamydophila abortus S26/3
Chlamydophila caviae GPIC
Chlamydia trachomatis A/HAR-13
Chlamydia trachomatis D/UW-3/CX
Chlamydia muridarum Nigg
Chlamydophila pneumoniae TW-183
Chlamydophila pneumoniae J138 2/2
Chlamydophila pneumoniae J138 1/2
Chlamydophila pneumoniae AR39
Chlamydophila pneumoniae CWL029
Clostridium tetani E88 1/2
Clostridium tetani E88 2/2
Clostridium acetobutylicum ATCC 824
Clostridium perfringens str. 13 1/2
Clostridium perfringens str. 13 2/2
Symbiobacterium thermophilum IAM 14863
Dehalococcoides sp. CBDB1
Dehalococcoides ethenogenes 195
Thermoanaerobacter tengcongensis MB4 1/2
Thermoanaerobacter tengcongensis MB4 2/2
Moorella thermoacetica ATCC 39073
Carboxydothermus hydrogenoformans Z-2901 2/2
Carboxydothermus hydrogenoformans Z-2901 1/2
Aquifex aeolicus VF5 2/2
Aquifex aeolicus VF5 1/2
Thermotoga maritima MSB8
Deinococcus radiodurans R1 1/2
Deinococcus radiodurans R1 2/2
Thermus thermophilus HB8 1/2
Thermus thermophilus HB8 2/2
Thermus thermophilus HB27 1/2
Thermus thermophilus HB27 1/2
Thermus thermophilus HB27 2/2
Thermus thermophilus HB27 2/2
Rhodobacter sphaeroides 2.4.1 2/2
Methanosarcina acetivorans C2A
Methanosarcina mazei Go1
Methanosarcina barkeri str. fusaro
Halobacterium sp. NRC-1
Natronomonas pharaonis DSM 2160
Haloarcula marismortui ATCC 43049
Methanopyrus kandleri AV19
Methanothermobacter thermautotrophicus str. Delta H
Methanococcus maripaludis S2
Methanocaldococcus jannaschii DSM 2661
Archaeoglobus fulgidus DSM 4304
Pyrococcus horikoshii OT3
Pyrococcus abyssi GE5
Pyrococcus furiosus DSM 3638
Thermococcus kodakarensis KOD1
Nanoarchaeum equitans Kin4-M
Thermoplasma volcanium GSS1
Thermoplasma acidophilum DSM 1728


100
Picrophilus torridus DSM 9790
100
30
Pyrobaculum aerophilum str. IM2
100
Aeropyrum pernix K1
88
100
Sulfolobus solfataricus P2
92
100
Sulfolobus tokodaii str. 7
100
98
100
Sulfolobus acidocaldarius DSM 639

Lisa 5. Bdellovibrio bacteriovorus HD100 tüüpilisuse väärtused (Joonis 3).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	50S ribosomal protein L7/L12, 50S ribosomal protein L10	1,71E-08
2	50S ribosomal protein L10, 50S ribosomal protein L1	1,12E-08
3	50S ribosomal protein L1, 50S ribosomal protein L11	7,8E-09
4	50S ribosomal protein L1, 50S ribosomal protein L11, transcription antitermination protein	1,21E-08
5	50S ribosomal protein L11, transcription antitermination protein, hypothetical protein	1,27E-08
6	transcription antitermination protein, hypothetical protein, translation elongation factor Tu, -	1,36E-08
7	hypothetical protein, translation elongation factor Tu, -	1,02E-08
8	translation elongation factor Tu	7,45E-09
9	translation elongation factor Tu, -, -, -	1,12E-08
10	translation elongation factor Tu, RNA methyltransferase, TrmH family, group, -, -, -, -	8,84E-09
11	RNA methyltransferase, TrmH family, group, -, -	6,78E-09
12	RNA methyltransferase, TrmH family, group, orotidine 5'-phosphate decarboxylase	8,37E-09
13	RNA methyltransferase, TrmH family, group, orotidine 5'-phosphate decarboxylase, hypothetical protein	1,16E-08
14	orotidine 5'-phosphate decarboxylase, hypothetical protein, hypothetical protein	9,9E-09
15	hypothetical protein, hypothetical protein, hypothetical protein	1,26E-08
	Genoomi keskmine	1,35E-08
	Ala keskmine−σ	8,38E-09

Lisa 6. *Campylobacter jejuni* RM1221 tüüpilisuse väärtused (Joonis 4a).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	transcriptional regulator, putative, amino acid ABC transporter, permease protein, His/Glu/Gln/Arg/opine family	2,12E-07
2	amino acid ABC transporter, permease protein, His/Glu/Gln/Arg/opine family, amino acid ABC transporter, permease protein, His/Glu/Gln/Arg/opine family	1,64E-07
3	amino acid ABC transporter, permease protein, His/Glu/Gln/Arg/opine family, amino acid ABC transporter, permease protein, His/Glu/Gln/Arg/opine family, amino acid ABC transporter, ATP-binding protein	1,06E-07
4	amino acid ABC transporter, permease protein, His/Glu/Gln/Arg/opine family, amino acid ABC transporter, ATP-binding protein	1,15E-07
5	amino acid ABC transporter, ATP-binding protein, Thr tRNA, Tyr tRNA, Gly tRNA, Thr tRNA	1,57E-07
6	elongation factor Tu, Thr tRNA, Tyr tRNA, Gly tRNA, Thr tRNA	1,1E-07
7	elongation factor Tu	6,23E-08
8	elongation factor Tu, 50S ribosomal protein L33, Trp tRNA	6,82E-08
9	elongation factor Tu, 50S ribosomal protein L33, translocase, transcription antitermination protein NusG, Trp tRNA	9,01E-08
10	translocase, transcription antitermination protein NusG, 50S ribosomal protein L11, Trp tRNA	1,08E-07
11	transcription antitermination protein NusG, 50S ribosomal protein L11, 50S ribosomal protein L1	9,16E-08
12	50S ribosomal protein L11, 50S ribosomal protein L1	1E-07
13	50S ribosomal protein L1, 50S ribosomal protein L10	9,5E-08
14	50S ribosomal protein L1, 50S ribosomal protein L10, 50S ribosomal protein L7/L12	1,02E-07
	Genoomi keskmine	1,35E-07
	Ala keskmine−σ	8,42E-08

Lisa 7. *Helicobacter pylori* 26695 tüüpilisuse väärtused (Joonis 4b).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	50S ribosomal protein L7/L12, 50S ribosomal protein L10, 50S ribosomal protein L1	6,62E-08
2	50S ribosomal protein L10, 50S ribosomal protein L1	5,98E-08
3	50S ribosomal protein L1, 50S ribosomal protein L11, transcription antitermination protein NusG	6,02E-08
4	50S ribosomal protein L1, 50S ribosomal protein L11, transcription antitermination protein NusG	6,11E-08
5	transcription antitermination protein NusG, translocase, 50S ribosomal protein L33, Trp tRNA	6,04E-08
6	transcription antitermination protein NusG, translocase, 50S ribosomal protein L33, elongation factor Tu, Trp tRNA	5,48E-08
7	50S ribosomal protein L33, elongation factor Tu	4,75E-08
8	elongation factor Tu	3,55E-08
9	elongation factor Tu, Thr tRNA, Gly tRNA, Tyr tRNA, Thr tRNA	4,16E-08
10	multidrug resistance protein (hetA), Thr tRNA, Gly tRNA, Tyr tRNA, Thr tRNA	3,58E-08
11	multidrug resistance protein (hetA)	5,51E-08
12	multidrug resistance protein (hetA)	7,28E-08
13	multidrug resistance protein (hetA), hypothetical protein	6,11E-08
14	multidrug resistance protein (hetA), hypothetical protein	5,61E-08
	Genoomi keskmine	6,42E-08
	Ala keskmine−σ	4,78E-08

Lisa 8. *Campylobacter jejuni subsp. jejuni* NCTC 11168 tüüpilisuse väärtused (Joonis 4c).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	putative transcriptional regulator, amino-acid ABC transporter integral membrane protein	1,93E-07
2	putative transcriptional regulator, amino-acid ABC transporter integral membrane protein, amino-acid ABC transporter integral membrane protein	2,05E-07
3	amino-acid ABC transporter integral membrane protein, amino-acid ABC transporter integral membrane protein	1,51E-07
4	amino-acid ABC transporter integral membrane protein, amino-acid ABC transporter ATP-binding protein	1,09E-07
5	amino-acid ABC transporter integral membrane protein, amino-acid ABC transporter ATP-binding protein, Thr tRNA, Tyr tRNA	1,1E-07
6	amino-acid ABC transporter ATP-binding protein, elongation factor Tu, Thr tRNA, Tyr tRNA, Gly tRNA, Thr tRNA	1,43E-07
7	elongation factor Tu, Tyr tRNA, Gly tRNA, Thr tRNA	1,18E-07
8	elongation factor Tu	6,06E-08
9	elongation factor Tu, 50S ribosomal protein L33, translocase, Trp tRNA	5,3E-08
10	elongation factor Tu, 50S ribosomal protein L33, translocase, transcription antitermination protein NusG, Trp tRNA	9,76E-08
11	translocase, transcription antitermination protein NusG, 50S ribosomal protein L11	1,1E-07
12	transcription antitermination protein NusG, 50S ribosomal protein L11, 50S ribosomal protein L1	8,68E-08
13	50S ribosomal protein L11, 50S ribosomal protein L1	9,94E-08
14	50S ribosomal protein L1, 50S ribosomal protein L10	1,03E-07
15	50S ribosomal protein L10, 50S ribosomal protein L7/L12, DNA-directed RNA polymerase beta chain	1,06E-07
	Genoomi keskmine	1,32E-07
	Ala keskmine−σ	8,28E-08

Lisa 9. *Helicobacter pylori* J99 tüüpilisuse väärtused (Joonis 4d).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	50S ribosomal protein L10, 50S ribosomal protein L1	5,65E-08
2	50S ribosomal protein L10, 50S ribosomal protein L1, 50S ribosomal protein L11	6,31E-08
3	50S ribosomal protein L1, 50S ribosomal protein L11, transcription antitermination protein NusG	6,59E-08
4	50S ribosomal protein L11, transcription antitermination protein NusG, translocase	6,11E-08
5	transcription antitermination protein NusG, translocase, 50S ribosomal protein L33	5,2E-08
6	translocase, 50S ribosomal protein L33, elongation factor Tu	5,41E-08
7	elongation factor Tu	4,79E-08
8	elongation factor Tu	2,77E-08
9	elongation factor Tu, abc transporter, ATP-binding protein	3,15E-08
10	abc transporter, ATP-binding protein	4,64E-08
11	abc transporter, ATP-binding protein	6,2E-08
12	abc transporter, ATP-binding protein	6,9E-08
13	abc transporter, ATP-binding protein, hypothetical protein	6,96E-08
14	abc transporter, ATP-binding protein, hypothetical protein	6,1E-08
	Genoomi keskmine	6,33E-08
	Ala keskmine−σ	4,58E-08

Lisa 10. *Thiomicrospira denitrificans* ATCC 33889 tüüpilisuse väärtused (Joonis 4e).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	Auxin Efflux Carrier, thioredoxin-like, tRNA/rRNA methyltransferase (SpoU)	7,52E-08
2	thioredoxin-like, tRNA/rRNA methyltransferase (SpoU)	7,06E-08
3	tRNA/rRNA methyltransferase (SpoU), Protein of unknown function DUF328	8,39E-08
4	tRNA/rRNA methyltransferase (SpoU), Protein of unknown function DUF328	9,94E-08
5	Protein of unknown function DUF328, Thr tRNA, Tyr tRNA, Gly tRNA, Thr tRNA	9,96E-08
6	Translation elongation factor Tu, Thr tRNA, Tyr tRNA, Gly tRNA, Thr tRNA	6,32E-08
7	Translation elongation factor Tu, Thr tRNA	2,72E-08
8	Translation elongation factor Tu, Ribosomal protein L33	3,13E-08
9	Translation elongation factor Tu, Ribosomal protein L33, SecE subunit of protein translocation complex, NusG antitermination factor, Trp tRNA	4,01E-08
10	Ribosomal protein L33, SecE subunit of protein translocation complex, NusG antitermination factor, Trp tRNA	5E-08
11	NusG antitermination factor, Ribosomal protein L11	6,27E-08
12	Ribosomal protein L11, Ribosomal protein L1	6,61E-08
13	Ribosomal protein L11, Ribosomal protein L1, Ribosomal protein L10	5,82E-08
14	Ribosomal protein L1, Ribosomal protein L10, Ribosomal protein L7/L12	5,33E-08
	Genoomi keskmine	6,78E-08
	Ala keskmine−σ	4,92E-08

Lisa 11. *Helicobacter hepaticus* ATCC 51449 tüüpilisuse väärtused (Joonis 4f).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	50S ribosomal protein L7/L12, 50S ribosomal protein L10, 50S ribosomal protein L1	4,98E-08
2	50S ribosomal protein L7/L12, 50S ribosomal protein L10, 50S ribosomal protein L1	4,98E-08
3	50S ribosomal protein L10, 50S ribosomal protein L1	4,27E-08
4	50S ribosomal protein L1, 50S ribosomal protein L11	4,24E-08
5	50S ribosomal protein L1, 50S ribosomal protein L11, transcription antitermination protein NusG	4,64E-08
6	50S ribosomal protein L11, transcription antitermination protein NusG, translocase, 50S ribosomal protein L33, Trp tRNA	5,03E-08
7	transcription antitermination protein NusG, translocase, 50S ribosomal protein L33, elongation factor Tu, Trp tRNA	5,05E-08
8	50S ribosomal protein L33, elongation factor Tu	4,58E-08
9	elongation factor Tu, Thr tRNA	3,6E-08
10	elongation factor Tu, Thr tRNA, Gly tRNA, Tyr tRNA, Thr tRNA	3,62E-08
11	hypothetical protein, Gly tRNA, Tyr tRNA, Thr tRNA	3,91E-08
12	hypothetical protein, Ser tRNA	5,92E-08
13	hypothetical protein, hypothetical protein, Ser tRNA, Cys tRNA, Leu tRNA, Gly tRNA	7,47E-08
14	hypothetical protein, hypothetical protein, Ser tRNA, Cys tRNA, Leu tRNA, Gly tRNA	4,39E-08
15	hypothetical protein, hypothetical protein	4,31E-08
	Genoomi keskmine	5,3E-08
	Ala keskmine−σ	4,03E-08

Lisa 12. *Treponema pallidum subsp. pallidum str. Nichols* tüüpilisuse väärtused (Joonis 5a).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	hypothetical protein, SsrA-binding protein	2,19E-08
2	hypothetical protein, SsrA-binding protein, signal peptidase I (sip)	2,3E-08
3	signal peptidase I (sip), oxygen-independent coproporphyrinogen III oxidase, putative	2,03E-08
4	signal peptidase I (sip), oxygen-independent coproporphyrinogen III oxidase, putative	1,8E-08
5	oxygen-independent coproporphyrinogen III oxidase, putative	2,11E-08
6	oxygen-independent coproporphyrinogen III oxidase, putative, translation elongation factor TU (tuf)	2,44E-08
7	oxygen-independent coproporphyrinogen III oxidase, putative, translation elongation factor TU (tuf)	2,45E-08
8	translation elongation factor TU (tuf)	2,15E-08
9	translation elongation factor TU (tuf), 30S ribosomal protein S10, ribosomal protein L3 (rplC)	1,54E-08
10	translation elongation factor TU (tuf), 30S ribosomal protein S10, ribosomal protein L3 (rplC)	1,54E-08
11	ribosomal protein L3 (rplC), 50S ribosomal protein L4	1,66E-08
12	ribosomal protein L3 (rplC), 50S ribosomal protein L4, ribosomal protein L23 (rplW)	1,48E-08
13	50S ribosomal protein L4, ribosomal protein L23 (rplW), 50S ribosomal protein L2	1,73E-08
14	ribosomal protein L23 (rplW), 50S ribosomal protein L2, ribosomal protein S19 (rpsS)	1,87E-08
15	50S ribosomal protein L2, ribosomal protein S19 (rpsS), 50S ribosomal protein L22	1,7E-08
	Genoomi keskmine	1,94E-08
	Ala keskmine−σ	1,43E-08

Lisa 13. *Treponema denticola* ATCC 35405 tüüpilisuse väärtused (Joonis 5b).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	protease complex-associated polypeptide	1,8E-08
2	protease complex-associated polypeptide	4,06E-08
3		4,53E-08
4		2,26E-08
5	hypothetical protein, hypothetical protein	2,51E-08
6	hypothetical protein, hypothetical protein, translation elongation factor Tu	3,75E-08
7	hypothetical protein, translation elongation factor Tu	5,2E-08
8	translation elongation factor Tu	3,74E-08
9	translation elongation factor Tu, 30S ribosomal protein S10, ribosomal protein L3	2,03E-08
10	translation elongation factor Tu, 30S ribosomal protein S10, ribosomal protein L3	2,91E-08
11	ribosomal protein L3, 50S ribosomal protein L4	3,61E-08
12	ribosomal protein L3, 50S ribosomal protein L4, ribosomal protein L23	3,97E-08
13	50S ribosomal protein L4, ribosomal protein L23, 50S ribosomal protein L2	4,2E-08
14	ribosomal protein L23, 50S ribosomal protein L2, ribosomal protein S19	4,06E-08
15	50S ribosomal protein L2, ribosomal protein S19, 50S ribosomal protein L22	3,08E-08
	Genoomi keskmine	6,37E-08
	Ala keskmine−σ	3,88E-08

Lisa 14. *Borrelia burgdorferi* B31 tüüpilisuse väärtused (Joonis 5c).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	hypothetical protein	1,61E-07
2	hypothetical protein	2,06E-07
3	hypothetical protein, hypothetical protein	1,62E-07
4	hypothetical protein, hypothetical protein, hypothetical protein	1,69E-07
5	hypothetical protein, hypothetical protein	1,84E-07
6	hypothetical protein, translation elongation factor TU (tuf)	2,16E-07
7	translation elongation factor TU (tuf)	1,67E-07
8	translation elongation factor TU (tuf), 30S ribosomal protein S10	8,27E-08
9	translation elongation factor TU (tuf), 30S ribosomal protein S10, ribosomal protein L3 (rplC)	8,12E-08
10	30S ribosomal protein S10, ribosomal protein L3 (rplC), 50S ribosomal protein L4	8,1E-08
11	ribosomal protein L3 (rplC), 50S ribosomal protein L4, ribosomal protein L23 (rplW)	7,96E-08
12	50S ribosomal protein L4, ribosomal protein L23 (rplW), 50S ribosomal protein L2	1,07E-07
13	ribosomal protein L23 (rplW), 50S ribosomal protein L2	1,2E-07
14	50S ribosomal protein L2, ribosomal protein S19 (rpsS), 50S ribosomal protein L22	8,23E-08
	Genoomi keskmine	1,4E-07
	Ala keskmine−σ	1,05E-07

Lisa 15. *Borrelia garinii* Pbi tüüpilisuse väärtused (Joonis 5d).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	UDP-N-acetylglucosamine 1-carboxyvinyltransferase, hypothetical protein	1,12E-07
2	hypothetical protein	2,2E-07
3	hypothetical protein	2,26E-07
4	hypothetical protein	1,7E-07
5	hypothetical protein, lipoprotein, putative	1,95E-07
6	lipoprotein, putative, translation elongation factor TU	2,08E-07
7	translation elongation factor TU	2,24E-07
8	translation elongation factor TU	1,62E-07
9	translation elongation factor TU, 30S ribosomal protein S10, ribosomal protein L3	7,8E-08
10	translation elongation factor TU, 30S ribosomal protein S10, ribosomal protein L3	8,74E-08
11	ribosomal protein L3, 50S ribosomal protein L4	8,89E-08
12	ribosomal protein L3, 50S ribosomal protein L4, hypothetical protein, ribosomal protein L23	9,09E-08
13	50S ribosomal protein L4, hypothetical protein, ribosomal protein L23, 50S ribosomal protein L2	1,15E-07
14	ribosomal protein L23, 50S ribosomal protein L2, ribosomal protein S19	1,11E-07
15	50S ribosomal protein L2, ribosomal protein S19, 50S ribosomal protein L22	9,03E-08
	Genoomi keskmine	1,53E-07
	Ala keskmine−σ	1,13E-07

Lisa 16. *Leptospira interrogans serovar Lai str*. 56601 tüüpilisuse väärtused (Joonis 5e).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	hypothetical protein	9,71E-08
2	hypothetical protein, Translation elongation and release factor	1,04E-07
3	Translation elongation and release factor	9,8E-08
4	Translation elongation and release factor	8,57E-08
5	Translation elongation and release factor	5,61E-08
6	Translation elongation and release factor, Elongation factor Tu	4,17E-08
7	Translation elongation and release factor, Elongation factor Tu	3,55E-08
8	Elongation factor Tu, ribosomal protein S10	2,47E-08
9	Elongation factor Tu, ribosomal protein S10, ribosomal protein L3	2,74E-08
10	ribosomal protein S10, ribosomal protein L3, 50S ribosomal protein L4	3,51E-08
11	ribosomal protein L3, 50S ribosomal protein L4	4,17E-08
12	50S ribosomal protein L4, ribosomal protein L23, 50S ribosomal protein L2	5,66E-08
13	50S ribosomal protein L4, ribosomal protein L23, 50S ribosomal protein L2	5,78E-08
14	50S ribosomal protein L2, ribosomal protein S19	4,93E-08
	Genoomi keskmine	7,47E-08
	Ala keskmine−σ	4,1E-08

Lisa 17. *Leptospira interrogans serovar Copenhageni str. Fiocruz* L1-130 tüüpilisuse väärtused (Joonis 5f).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	30S ribosomal protein S3, 50S ribosomal protein L22, 30S ribosomal protein S19, 50S ribosomal protein L2	5,2E-08
2	30S ribosomal protein S19, 50S ribosomal protein L2	4,34E-08
3	50S ribosomal protein L2, 50S ribosomal protein L23, 50S ribosomal protein L4	3,97E-08
4	50S ribosomal protein L2, 50S ribosomal protein L23, 50S ribosomal protein L4	5,5E-08
5	50S ribosomal protein L4, 50S ribosomal protein L3	6,08E-08
6	50S ribosomal protein L3, 30S ribosomal protein S10, elongation factor Tu	5,29E-08
7	50S ribosomal protein L3, 30S ribosomal protein S10, elongation factor Tu	4,15E-08
8	elongation factor Tu	3,12E-08
9	elongation factor Tu, elongation factor G	2,31E-08
10	elongation factor Tu, elongation factor G	2,31E-08
11	elongation factor G	3,71E-08
12	elongation factor G	4,74E-08
13	elongation factor G	5,62E-08
14	elongation factor G, hypothetical protein	9,12E-08
15	hypothetical protein	1,01E-07
	Genoomi keskmine	7,44E-08
	Ala keskmine−σ	4,24E-08

Lisa 18. *Rhodobacter sphaeroides* 2.4.1 Ef-Tu gi|77462243 tüüpilisuse väärtused (Joonis 6a).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	hypothetical protein, DNA topoisomerase IV subunit A	1,07E-07
2	DNA topoisomerase IV subunit A	8,49E-08
3	DNA topoisomerase IV subunit A	9,19E-08
4	DNA topoisomerase IV subunit A	9,62E-08
5	DNA topoisomerase IV subunit A, hypothetical protein	1,06E-07
6	hypothetical protein, Elongation factor Tu (EF-Tu)	9,97E-08
7	hypothetical protein, Elongation factor Tu (EF-Tu)	8,66E-08
8	Elongation factor Tu (EF-Tu)	8,56E-08
9	Elongation factor Tu (EF-Tu), Putative acetyltransferase	8,68E-08
10	Putative acetyltransferase, Trp tRNA	8,86E-08
11	Putative acetyltransferase, Putative preprotein translocase, SecE subunit, Probable transcription antitermination protein NusG, Trp tRNA	9,52E-08
12	Putative preprotein translocase, SecE subunit, Probable transcription antitermination protein NusG	8,6E-08
13	Probable transcription antitermination protein NusG, Ribosomal protein L11, 50S ribosomal protein L1	7,16E-08
14	Ribosomal protein L11, 50S ribosomal protein L1	7,12E-08
	Genoomi keskmine	9,89E-08
	Ala keskmine−σ	6,86E-08

Lisa 19. *Rhodobacter sphaeroides* 2.4.1 Ef-Tu gi|77462257 tüüpilisuse väärtused (Joonis 6b).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	Possible transporter, RhaT family, DMT superfamily, 30S ribosomal protein S12, 30S ribosomal protein S7	8,48E-08
2	30S ribosomal protein S12, 30S ribosomal protein S7, Elongation factor G	5,49E-08
3	30S ribosomal protein S7, Elongation factor G	6,19E-08
4	Elongation factor G	7,08E-08
5	Elongation factor G	6,87E-08
6	Elongation factor G, Elongation factor TU	7,28E-08
7	Elongation factor G, Elongation factor TU	7,22E-08
8	Elongation factor TU	7,34E-08
9	Elongation factor TU, 30S ribosomal protein S10	8,43E-08
10	Elongation factor TU, 30S ribosomal protein S10, 50S ribosomal protein L3	7,47E-08
11	30S ribosomal protein S10, 50S ribosomal protein L3, 50S ribosomal protein L4	6,64E-08
12	50S ribosomal protein L3, 50S ribosomal protein L4, 50S ribosomal protein L23	8,17E-08
13	50S ribosomal protein L4, 50S ribosomal protein L23, 50S ribosomal protein L2	7,56E-08
14	50S ribosomal protein L23, 50S ribosomal protein L2	5,76E-08
15	50S ribosomal protein L2, Ribosomal protein S19, 50S ribosomal protein L22	7,03E-08
	Genoomi keskmine	9,89E-08
	Ala keskmine−σ	7,02E-08

Lisa 20. *Rhodobacter sphaeroides* 2.4.1 gi|77464017 tüüpilisuse väärtused (Joonis 6c).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	hypothetical protein, hypothetical protein, isocitrate dehydrogenase	7,82E-08
2	hypothetical protein, isocitrate dehydrogenase	8,46E-08
3	isocitrate dehydrogenase	8,67E-08
4	isocitrate dehydrogenase, hypothetical protein	8,63E-08
5	isocitrate dehydrogenase, hypothetical protein	8,11E-08
6	hypothetical protein, EF-Tu; elongation factor Tu	8,02E-08
7	hypothetical protein, EF-Tu; elongation factor Tu	7,5E-08
8	EF-Tu; elongation factor Tu	9,17E-08
9	EF-Tu; elongation factor Tu, putative glutamine amidotransferase	1,05E-07
10	EF-Tu; elongation factor Tu, putative glutamine amidotransferase	9,14E-08
11	putative glutamine amidotransferase, hypothetical protein	9,62E-08
12	putative glutamine amidotransferase, hypothetical protein, alanyl-tRNA synthetase	9,53E-08
13	hypothetical protein, alanyl-tRNA synthetase	1,08E-07
14	alanyl-tRNA synthetase	1,28E-07
15	alanyl-tRNA synthetase	1,16E-07
	Genoomi keskmine	9,89E-08
	Ala keskmine−σ	8,5E-08

Lisa 21. *Hahella chejuensis* KCTC 2396 tüüpilisuse väärtused (Joonis 7).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	ribosomal protein S19, ribosomal protein L2	5,5E-09
2	ribosomal protein L2, Ribosomal protein L23, Ribosomal protein L4	9,03E-09
3	ribosomal protein L2, Ribosomal protein L23, Ribosomal protein L4, Ribosomal protein L3	7,19E-09
4	Ribosomal protein L4, Ribosomal protein L3	7,19E-09
5	Ribosomal protein L3, ribosomal protein S10	8,99E-09
6	Ribosomal protein L3, ribosomal protein S10, translation elongation factor Tu	5,07E-09
7	translation elongation factor Tu	5,05E-09
8	translation elongation factor Tu, translation elongation factor G	6,4E-09
9	translation elongation factor Tu, translation elongation factor G	6,15E-09
10	translation elongation factor G	6,24E-09
11	translation elongation factor G	3,93E-09
12	translation elongation factor G	2,07E-09
13	translation elongation factor G, ribosomal protein S7	3,79E-09
14	ribosomal protein S7, ribosomal protein S12	5,59E-09
15	ribosomal protein S12, DNA-directed RNA polymerase, beta' subunit/160 kD subunit	6,83E-09
16	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	8,67E-09
17	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	5,81E-09
18	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	2,24E-09
19	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	5,11E-09
20	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	7,61E-09
21	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	7,07E-09
22	DNA-directed RNA polymerase, beta' subunit/160 kD subunit	7,21E-09
23	DNA-directed RNA polymerase, beta' subunit/160 kD subunit, DNA-directed RNA polymerase, beta subunit	6,5E-09
24	DNA-directed RNA polymerase, beta' subunit/160 kD subunit, DNA-directed RNA polymerase, beta subunit	4,69E-09
25	DNA-directed RNA polymerase, beta subunit	6,25E-09
26	DNA-directed RNA polymerase, beta subunit	1,03E-08
27	DNA-directed RNA polymerase, beta subunit	1,14E-08
28	DNA-directed RNA polymerase, beta subunit	1,03E-08
29	DNA-directed RNA polymerase, beta subunit	9,11E-09
30	DNA-directed RNA polymerase, beta subunit	8,89E-09
31	DNA-directed RNA polymerase, beta subunit	8,26E-09
32	DNA-directed RNA polymerase, beta subunit, ribosomal protein L7/L12	5,84E-09
33	ribosomal protein L7/L12, Ribosomal protein L10	9,2E-09
34	ribosomal protein L7/L12, Ribosomal protein L10, ribosomal protein L1	1,12E-08
35	Ribosomal protein L10, ribosomal protein L1, ribosomal protein L11	6,52E-09
36	ribosomal protein L1, ribosomal protein L11	6,65E-09
37	ribosomal protein L11, transcription termination/antitermination factor NusG	5,11E-09
38	transcription termination/antitermination factor NusG, preprotein translocase, SecE subunit, Trp tRNA	3,61E-09
39	transcription termination/antitermination factor NusG, preprotein translocase, SecE subunit, translation elongation factor Tu, Trp tRNA	5,79E-09
40	translation elongation factor Tu, Trp tRNA	6,46E-09

41	translation elongation factor Tu, hypothetical protein, Thr tRNA, Gly tRNA, Tyr tRNA	7,08E-09
42	translation elongation factor Tu, hypothetical protein, hypothetical protein, Thr tRNA, Gly tRNA, Tyr tRNA, Thr tRNA	5,81E-09
43	hypothetical protein, putative transcriptional regulator, Tyr tRNA, Thr tRNA	2,99E-09
44	hypothetical protein, putative transcriptional regulator	2,66E-09
45	putative transcriptional regulator, Biotin-(acetyl-CoA carboxylase) ligase	1,94E-09
46	putative transcriptional regulator, Biotin-(acetyl-CoA carboxylase) ligase	-1,2E-10
47	Biotin-(acetyl-CoA carboxylase) ligase, IstB helper protein, IS21 family	-3,4E-10
	Genoomi keskmine	1,34E-08
	Ala keskmine−σ	4,82E-09

Lisa 22. *Dehalococcoides sp*. CBDB1 tüüpilisuse väärtused (Joonis 8a).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	ribosomal protein L7, ribosomal protein L10, 50S ribosomal protein L1	1,33E-08
2	ribosomal protein L10, 50S ribosomal protein L1	1,56E-08
3	50S ribosomal protein L1, ribosomal protein L11	1,65E-08
4	50S ribosomal protein L1, ribosomal protein L11, transcription antitermination protein NusG	1,37E-08
5	ribosomal protein L11, transcription antitermination protein NusG, preprotein translocase, SecE subunit, ribosomal protein L33	1,37E-08
6	transcription antitermination protein NusG, preprotein translocase, SecE subunit, ribosomal protein L33, translation elongation factor Tu	1,37E-08
7	ribosomal protein L33, translation elongation factor Tu	1,03E-08
8	translation elongation factor Tu, Thr tRNA, Tyr tRNA	6,59E-09
9	translation elongation factor Tu, hypothetical protein, Thr tRNA, Tyr tRNA, Thr tRNA	7,6E-09
10	hypothetical protein, glycosyl transferase, group 1 family protein, Thr tRNA	1,58E-08
11	hypothetical protein, glycosyl transferase, group 1 family protein	1,85E-08
12	glycosyl transferase, group 1 family protein	1,9E-08
13	glycosyl transferase, group 1 family protein, hypothetical protein	1,91E-08
14	glycosyl transferase, group 1 family protein, hypothetical protein	1,85E-08
	Genoomi keskmine	1,61E-08
	Ala keskmine−σ	1,33E-08

Lisa 23. *Dehalococcoides ethenogenes* 195 tüüpilisuse väärtused (Joonis 8b).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	ribosomal protein L7/L12, ribosomal protein L10, 50S ribosomal protein L1	1,63E-08
2	ribosomal protein L10, 50S ribosomal protein L1	1,45E-08
3	50S ribosomal protein L1, ribosomal protein L11	1,29E-08
4	50S ribosomal protein L1, ribosomal protein L11, transcription antitermination protein NusG	1,44E-08
5	ribosomal protein L11, transcription antitermination protein NusG, preprotein translocase, SecE subunit, ribosomal protein L33	1,61E-08
6	transcription antitermination protein NusG, preprotein translocase, SecE subunit, ribosomal protein L33, translation elongation factor Tu	1,23E-08
7	ribosomal protein L33, translation elongation factor Tu	9,56E-09
8	translation elongation factor Tu, Thr tRNA, Tyr tRNA	7,31E-09
9	translation elongation factor Tu, hypothetical protein, Thr tRNA, Tyr tRNA, Thr tRNA	8,64E-09
10	hypothetical protein, Tyr tRNA, Thr tRNA	1,34E-08
11	hypothetical protein, glycosyl transferase, group 1 family protein	1,7E-08
12	hypothetical protein, glycosyl transferase, group 1 family protein	1,85E-08
13	glycosyl transferase, group 1 family protein, hypothetical protein	1,69E-08
14	glycosyl transferase, group 1 family protein, hypothetical protein	1,82E-08
	Genoomi keskmine	1,70E-08
	Ala keskmine−σ	1,4E-08

Lisa 24. *Symbiobacterium thermophilum* IAM 14863 tüüpilisuse väärtused (Joonis 8c).

Segment	Segmendi kodeeritud valkude nimetused	Tüüpilisuse väärtus
1	30S ribosomal protein S19, 50S ribosomal protein L2	4,28E-08
2	50S ribosomal protein L2, 50S ribosomal protein L23, 50S ribosomal protein L4	6,48E-08
3	50S ribosomal protein L2, 50S ribosomal protein L23, 50S ribosomal protein L4	6,3E-08
4	50S ribosomal protein L4, 50S ribosomal protein L3	5,77E-08
5	50S ribosomal protein L4, 50S ribosomal protein L3, 30S ribosomal protein S10	7,17E-08
6	50S ribosomal protein L3, 30S ribosomal protein S10, translation elongation factor Tu	6,72E-08
7	30S ribosomal protein S10, translation elongation factor Tu	5,91E-08
8	translation elongation factor Tu	6,67E-08
9	translation elongation factor Tu, protein translation elongation factor G	7,12E-08
10	protein translation elongation factor G	6,42E-08
11	protein translation elongation factor G	7,33E-08
12	protein translation elongation factor G	7,61E-08
13	protein translation elongation factor G, 30S ribosomal protein S7	6,14E-08
14	protein translation elongation factor G, 30S ribosomal protein S7, 30S ribosomal protein S12	5,66E-08
	Genoomi keskmine	8,10E-08
	Ala keskmine−σ	5,64E-08