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

Aleksander Sudakov

Horisontaalse geeniülekande tuvastamise *in silico* meetodid

Keskastme lõputöö

Juhendaja: Tõnu Margus

TARTU
2005

Sisukord

Sisukord....2
Lühendid ja mõisted....3
1. Sissejuhatus....4
2. Mikroorganismide geneetika....5
2.1 Mikroobide varajane evolutsioon....5
2.2 Süstemaatiline jaotamine....5
2.3 Geneetiline materjal ja selle mitmekesisus....6
2.4 Bakterite muutlikkus....8
3. Horisontaalne geeniülekanne....10
3.1 Ülevaade....10
3.2 HGT ajalugu....10
3.3 Hetkeseis ja ulatus....12
4. In silico meetodid horisontaalse geeniülekande ennustamiseks....14
4.1 Meetodite klassifikatsioon....14
4.2 Parameetrilised meetodid....15
4.3 Fülogeneetilised meetodid....18
4.4 Füleetilise jaotuse meetodid....21
4.5 Kauge liigiga sarnasuse meetodid....22
4.6 Arutelu....24
5. Kokkuvõte....26
6. Summary....27
Kasutatud kirjandus....28

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)
CAI	koodoni adaptatsiooniindeks (Codon Adaptation Index)
ClustalW	järjestuste globaalse joondamise programm
COG	ortoloogsete gruppide klaster (Cluster of Ortologous Groups)
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
MM	Markovi mudel
MSA	mitme järjestuse joondus (Multiple Sequence Alignment)
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)
rRNA	ribosomaalne RNA
RSCU	Suhteline sünonüümsete koodonite kasutuskoefitsient (Relative Synonymous Codon Usage)
SEH	korduva endosümbioosi hüpotees (Serial Endosymbios Hypotyhesis)
SSU	ribosoomi väike subühik (Small Subunit)
XGD	geeni asendamine teisest liigist pärit ortoloogi poolt (Xenologous Gene Displacement)

1. Sissejuhatus

Alates molekulaarbioloogia tekkimisest arvati, et organismidevaheliste geneetiliste sidemete ajalooline muster võib olla rekonstrueeritav raku molekulides sisalduva info abil. Selleks tuleb võrrelda geenide ja valkude järjestusti eri organismidel. Charles Darwini teooria järgi on liigid seotud puukujulise põlvnemisega, igal kahel liigil on ühine eelkäija. Viimase aja genoomikas on pööratud tähelepanu genoomile kui olemile, mis on eristatav geenidest ja organismist (Sicheritz-Pontén, 2000). Geenid on kromosoomil omavahel füüsiliselt seotud, mis omakorda päranduvad koos tütarrakkudesse. Kuid geenide, genoomide ja organismide fülogenees on eri tahud või kirjeldused samast ajaloolisest protsessist. Iga geeni või valgu perekonna fülogeneetiline puu näitab ka organismide sugulust. Kuid informatsioon üksikgeenidel võib olla ebapiisav kirjeldamaks objektiivselt organismide põlvnemist. Sel puhul tuleb kasutada mitme geeni või valgu järjestusi täielikult sekveneeritud genoomidest, et saada parem, ajalooliselt täpsem, puu.

Tänaseks on üldtuntud fakt, et geenid võivad päranduda mitte ainult vertikaalselt, emarakult tütarrakkudele mööda evolutsioonipuu üksikharusid, vaid ka horisontaalselt, otse ühest harust teise. Horisontaalselt ülekantud geeni lähimaid homolooge võib leida evolutsioonipuu doonori, aga mitte retsipiendi harus (Clarke *et al*., 2002). Selle tagajärjena on antud geeni põhjal ehitatud fülogenees väga erinev vertikaalselt pärandunud geenide omast. Kui horisontaalne geeniülekanne (HGT) on sage, teeb anomaalsete geenide puude arv ja kirjusus võimatuks määrata genoomi põhjal organismi õiget evolutsioonipuud.

Selleks, et analüüsimisel mitte kasutada geene, mis on mõjutatud HGT poolt, tuleb need alguses identifitseerida. Viimasel aastakümnel on tehtud mitmeid uuringuid, et hinnata kui suur osa genoomist on mõjutatud sellest fenomenist ning millised tunnused on iseloomulikud horisontaalselt ülekandunud geenidele. Uute genoomide sekveneerimine on suurendanud arvutil tehtava info töötlemise rolli. Kuid usaldusväärset horisontaalselt ülekantud geenide identifitseerimist ei saa veel automaatselt teha. Sellise meetodi väljatöötamine võib olla lähiaastatel saavutatav.

2. Mikroorganismide geneetika

2.1 Mikroobide varajane evolutsioon

Me ei tea kindlalt, millised organismid olid esimesed bioloogilised olemid, kuid selle kohta on mitu teooriat. Paljud teadlased usuvad, et esimene elu vorm oli anaeroobne, sest varajase Maa atmosfäär ei sisaldanud hapnikku (Schopf, 1992). Võimalik, et need mikroorganismid olid termofiilsed, elades kuumaveeallikates või vulkaanide kõrval. Sellele viitab fakt, et elusloodusе evolutsioonipuu kõige vanemad harud sisaldavad termofiilseid eubaktereid ja arheaid (Nelson et al., 1999). Carl Woese pakkus mikroobide esiisa nimeks "*progenote*". Selle tähtsaimaks omaduseks oli rakumembraan, mis eraldab olulisi kemikaale väliskeskkonnast ning millel saab läbi viia lihtsaid reaktsioone (Woese, 1987). *Progenote* populatsioon võis olla jagatud varajases arengus kahte liini, mis viis eubakterite ja arheate tekkele. Algne metabolism oli tõenäoliselt heterotroofne, sest autotroofsed organismid on liiga keerulised, et tekkida spontaanselt. Heterotroofide toiduks olid lihtsad abiootiliselt tekkinud orgaanilised ühendid. Energiat võis saada ka gaasilise vesiniku oksüdeerimise teel. Vesinikukasutajad on teada nii arheate kui eubakterite hulgas. Mõned neist on ka autotroofsed ning kasutavad süsinikdioksiidi ainsa süsinikuallikana, paljud on anaeroobsed, sobides varajase Maa hapnikuta keskkonda. Fotosüntees tekkis peale bakterite ja arheate lahknemist – paljudel eubakteritel on fotosünteesiaparaat, aga ühelgi arheal seda ei esine (Perry *et al.*, 2002).

Tsüanobakterid on ainsad prokarüoodid, mille fotosüntees toodab hapnikku. Fossiilidest saadud tõendusmaterjali põhjal on selgunud, et tsüanobakterid olid olemas vähemalt 3,3 miljonit aastat tagasi. Need bakterid põhjustasid hapniku hulga kasvu atmosfääris ning 2 miljonit aastat tagasi oli hapniku tase piisavalt kõrge, et moodustada aeroobset keskkonda (Schopf, 1992).

2.2 Süstemaatiline jaotamine

Bakterite klassifikatsioon on taksonoomia osa, mis tegeleb bakterite grupeerimisega taksonitesse. Organismide klassifitseerimine võib olla kunstlik, põhinedes organismi fenotüübil, või loomulik, põhinedes organismi arvataval evolutsioonil. Algselt baseerus prokarüootide klassifikatsioon peamiste füsioloogiliste ja morfoloogiliste tunnuste olemasolul. Sellised tunnused on pealtnäha fundamentaalsed ning tihedalt integreeritud raku struktuuri ja

talitusega. Ka fülogeneetilised puud koostati tunnuste tekkimise võimaliku järjekorra põhjal tolleaegsetes bakterirühmades (Boucher et al., 2003). Seega oli enne molekulaarse tüpiseerimise ajastut kogu bakterite klassifitseerimine kunstlik, sest puudus arvestatav alus nende evolutsiooni määramiseks (nagu taimedel ja loomadel on fossiilid). 1960-ndatel avastati, et organismide evolutsioon võib olla salvestatud nende makromolekulide järjestuses. Mõned molekulid, nagu rRNA ja mõned valgud, on evolutsiooni jooksul muutunud piisavalt vähe, et on võimalik eristada nii kaugeid kui ka suhteliselt lähedasi liike. Molekulide järjestused sisaldavad informatsiooni taksonite vahelisest sugulusest. Tulemuseks on 16S rRNA sekveneerimisele tuginev fülogeneetiline klassifikatsioon (Guerrero, 2001).

2.3 Geneetiline materjal ja selle mitmekesisus

Pärilik informatsioon on talletatud kõikide organismide rakkudes korrapärastes DNA segmentides, geenides ja geenidvahelistes regioonides, mis on ühendatud suurtesse molekulidesse – kromosoomidesse. Kogu antud organismi geneetiline materjal moodustab genoomi. Enamike tuntud bakterite genoom on üks ringikujuline kromosoom. Lisaks on mõnedes bakterites olemas väiksemaid ringikujulisi DNA molekule – plasmiide, mis võivad sisaldada eri tingimustes elamiseks vajalikku informatsiooni. Kõige väiksemad genoomid on viirustel, millel võivad esineda nii ühe- kui kaheahelalised DNA või RNA molekulid. Kõrgemad organismid omavad komplekti mitmest lineaarsest kromosoomist.

Enne täisgenoomide sekveneerimist olid meie teadmised prokarüootsest genoomi struktuurist piiratud andmetega, mida sai geel-elektroforeesist ja klassikalisest geenide kaardistamisest. Sekveneerimine on muutunud lihtsamaks ning see lõi aluse võrdleva genoomika tekkele. 22.03.2005 seisuga on sekveneeritud 261 prokarüootset ja eukarüootset genoomi (www.genomesonline.org). Ka üksikute genoomide uurimine lubas teha mitmeid avastusi, võrdlev genoomika aga lisas võimaluse vaadelda genoomide arengu ajalugu. Võrdleva genoomika abil saab uurida järjestuse evolutsiooni ja muutumist igas suurusjärgus alates genoomist kuni üksiku nukleotiidini (Bentley, Parkhill, 2004).

Bakteri genoomi kirjeldamiseks on mitmeid sobivaid parameetreid. Enim kasutatavad on selle suurus, kuju (lineaarne või ringikujuline), replikonide arv, GC protsent, AT/GC *skew*, geenide arv, valkude arv eri klassides. Kõik nimetatud parameetrid kõiguvad liigiti suurtes vahemikes.

Genoomi suurused eri bakterite liikides erinevad kuni kakskümmend korda (490 kbp *Nanoarchaeum equitans´*il kuni 9105 kbp *Bradyrhizobium japonicum'* il). Isegi ühe liigi siseselt võib erinevus olla kuni 10^6 aluspaari (Ussery, Hallin, 2004). Prokarüootses genoomis on genoomi suuruse ja geeniarvu vahel enamasti lineaarne sõltuvus (Bentley, Parkhill, 2004). Seetõttu on väiksem genoom nendel liikidel, kes elavad suhteliselt muutumatutes tingimustes, näiteks parasiteerides peremeesorganismi sees, ja vajavad vähe geene; suur genoom on vajalik muutlikus keskkonnas, kus läheb vaja mitmeid metabolismisüsteeme, näiteks mullas. Suur ja kiire genoomi suuruse ja sisu varieeruvus on võimalik tänu mitmele mehhanismile: geeni duplikatsioon, horisontaalne geeni omandamine, liigisisene geeni kaotus, geenide aktiveerimine ja deaktiveerimine mobiilsete elementidega (Twyman, 1998). See lubab prokarüootidel muutuda ja adapteeruda vastavalt keskkonna survele. Suur genoom on omandanud osa geene, mis on vaid harva kasutatavad, aga nad annavad võimaluse elada laias keskkonnatingimuste kõikumise vahemikus. Sellise genoomiga bakter on näiteks *Streptomyces coelicolor* (Bentley, Parkhill, 2004).

Genoomi vähenemine on tavaliselt tingitud ökoloogilise nišši muutumisest ja adaptatsioonist uues, lihtsamas või stabiilsemas keskkonnas. Geeni kaotamisele eelneb selle inaktiveerimine pseudogeeniks, mistõttu bakteriliikidel, mis on hiljuti muutnud oma evolutsioonilist nišši on paljud geenid inaktiivsed; näiteks *Mycobacterium leprae* 3,4 Mbp genoomis on vaid 1604 funktsionaalset geeni ja 1116 pseudogeeni, mis ei ole veel genoomist eemaldunud (Bentley, Parkhill, 2004).

Ajalooliselt on bakterite genoomi kirjeldatud kui üksikut ringikujulist DNA kromosoomi. Tänapäeval on teada mitmeid kõrvalekaldeid sellest mudelist: mitme kromosoomiga ja lineaarse kromosoomiga bakterid. Lineaarsed kromosoomid on iseseisvalt tekkinud vähemalt kolmes bakteriperekonnas (Bentley, Parkhill, 2004). Ei ole teada, mis põhjustab sellise adaptatsiooni teket või mis on selle kasu evolutsioonis.

GC ja AT aluspaaride osakaal bakterites kõigub 26.5% ja 72.1% vahel. GC osakaal korreleerub genoomi suurusega (suuremas genoomis rohkem GC aluspaare) ja eluviisiga (vabalt elavates bakterites on GC sisaldus keskmiselt 49%, parasiitides – 38%). GC osakaal igas genoomis on piiratud mitme bakterisisese ja keskkonna faktori poolt. Kui bakter omandab võõra DNA on selle GC sisaldus enamasti erinev. Ajapikku läbib võõras DNA amelioratsiooni protsessi: suunatud mutatsioonide mõjul muutub selle GC sisaldus samaks nagu ülejäänud genoomil (Bentley, Parkhill, 2004).

2.4 Bakterite muutlikkus

Bakterite muutlikkus on tihedalt seotud nende geneetilise informatsiooni muutusega, mida põhjustavad mutatsioonid, geneetilise materjali vahetamine ning reparatsiooni protsessid (Twyman, 1998).

Uue mutatsiooni efekti võib näha bakteris juba mõne tunni jooksul miljonites tütarrakkudes. Kui genoomis olevate geenide ekspressioonist ei piisa vastuseks keskkonna muutustele, siis peab bakter kohanduma uute keskkonnatingimustega muul viisil.

Mutatsioon on päranduv muutus genoomi vastava regiooni nukleotiidses järjestuses. Mutatsioonid võivad olla põhjustatud vigadest DNA replikatsioonis, DNA kahjustamisest, vigadest rekombinatsioonis ning mobiilsete geneetiliste elementide aktiivsusest. Juhuslikud mutatsioonid bakteri populatsioonis tekivad aeglaselt, kuid pidevalt. Mõned mutatsioonid võivad olla kahjulikud ning viia funktsionaalse geeni inaktiveerumisele. Mutatsioonid on üks peamisi muutlikkuse põhjuseid bakteris (Perry *et al.*, 2002).

Bakteri evolutsioon võib olla kiirendatud tänu nende võimele omandada ja ekspresseerida terveid geene, mis on pärit teistest bakteritest. Selline muutlikkus on eriti kiire, kui uus omandatud geen võimaldab ellu jääda potentsiaalselt surmavas keskkonnas. Uued geenid võivad suurendada bakteri metaboolset mitmekülgsust, lubades hõivata uusi ökoloogilisi nišše. Geneetilise info liikumist sellisel viisil nimetatakse horisontaalseks või lateraalseks geeniülekandeks.

On mitmeid viise kuidas teisest rakust pärit geneetiline materjal pääseb bakterisse:

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

Lisaks on *Streptomyces* liikidel kirjeldatud *cell fusion* mehhanismi, mille puhul kahe raku plasmamembraanid sulanduvad kokku (Twyman, 1998).

Uue geneetilise info integreerumine kromosoomi kasutab rekombinatsiooni. Rekombinatsioon on kontrollitud mitme ensüümi poolt ning selle käigus toimub DNA ahelate ümberpaigutumine lõikamise ja ligeerimise abil. Homoloogiline rekombinatsioon vajab osalevate DNA järjestuste sarnasust. Sait-spetsiifilises rekombinatsioonis osaleb järjestus, mis on piiratud kahe lühikese saidiga, mida ensüümid ära tunnevad. Seda rekombinatsioonitüüpi kasutavad mitmed protsessid, näiteks viiruse genoomi integreerumine, transposoonide liikumine, promootoralade inversioon (Bentley, Parkhill, 2004).

Transpositsioon on geeni liikumine ühest kromosoomi saidist teise. Selline liikumine vajab eripärast rekombinatsiooni transponeeruva DNA ja kromosoomi vahel. See on sarnane sait-spetsiifilisele rekombinatsioonile ja kasutab spetsialiseeritud ensüüme ning DNA järjestusi, mis määravad transpositsiooni saidi. Mobiilsete elementide hulka bakterite genoomides kuuluvad IS-elemendid, transposoonid ja transponeeruvad bakteriofaagid (Twyman, 1998). IS-elemendid on lühikesed 700-5000 bp, sisaldavad otstes pöördkordusjärjestusi ja kodeerivad nende liikumist võimaldavaid ensüüme, sh. transposaasi. Transposoonid kannavad lisaks ka ebavajalikke geene ning ulatuvad kuni 14 kbp (Tn7) ja rohkem (Twyman, 1998).

Mobiilsete elementide liikumine genoomis põhjustab insertsioonisaidis raaminihkeid ja geeni inaktiveerumist (Perry *et al.*, 2002). Transposooni väljalõikamine geenist võib selle uuesti aktiveerida.

3. Horisontaalne geeniülekanne

3.1 Ülevaade

Klassikalise evolutsiooniteooria järgi pärandub geneetiline informatsioon evolutsoonipuul emarakult tütarrakkudele vertikaalselt.

Horisontaalne geeniülekanne (HGT) on võõra geneetilise informatsiooni liikumine teise liigi genoomi. See on evolutsiooni fenomen, mille ulatus ja olemasolu on pikka aega olnud vaidlusalune küsimus. HGT on ka raske üheselt tõestada (Daubin *et al*., 2003). Tänu bakterite võimele integreerida võõrast DNA-d enda genoomi, on paljud liigid evolutsioneerunud kiiremini ja omandasid võime asustada uusi ökologilisi nišše. Juba enne täisgenoomide sekveneerimist räägiti üksikutest HGT juhtudest, aga valitseva arvamuse järgi oli fenomen liiga harvaesinev selleks, et mõjutada evolutsiooni kulgu. Ainus evolutsiooni sündmus, kus tunnistati HGT tähtsust, oli geeniülekanne endosümbiontsetest organellidest. Paljud geenid liikusid mitokondritest ja kloroplastidest eukarüootide tuuma genoomi (Woese, 1987).

Sekveneeritud järjestuste genoomika näitas, et tegemist on liiga sagedase nähtusega, et pidada seda tähtsusetuks. Esimene kinnitus 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, millest võis järeldada kõikide taoliste geenide *E. coli* poolt omandamist.

Mitme prokarüootse genoomi võrdlus tõestas, et horisontaalsel ülekandel bakterites on oluline roll (Ochman *et al*., 2000). Suur erinevus ka evolutsiooniliselt lähedaste bakteriliikide geenikomplektides näitas, et genoomide evolutsiooni ei saa kirjeldada ainult vertikaalse põlvnevusega. Osa erinevusest tulenes geenide kadumisest, eriti parasiitides, aga ainult HGT abil sai seletada genoomi ebaühtlast koostist (Ochman *et al*., 2000). Eriti suur on mitmekesisus arheate puhul. Nende geenide sarnasus on jaotatud funktsionaalsusega seoses: info töötlemisega seotud geenid (translatsioon, transkriptsioon ja replikatsioon) on eukarüootse sugulusega, aga metaboolsed, struktuursed ja mõned tundmatud valgud on eubakteriaalsed (Koonin *et al*., 2001).

3.2 HGT ajalugu

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

1943. a. 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 (Koonin *et al*., 2001).

Antibiootikumi resistentsuse levikut plasmiidide jm. mobiilsete elementide abil tunti juba aastakümneid tagasi, kuid tol ajal ei arvatud, et horisontaalse geeniülekande fenomen on laialdaselt levinud. Ka paljud hiljutised genoomianalüüsid on tõestanud, et vertikaalne pärandumine on domineeriv geenide leviku viis (Lawrence, Hendrickson, 2003).

Esimesed molekulaarse fülogeneesi uurimised olid tehtud Zukerandl ja Pauligi poolt. Nad näitasid, et liikide fülogeneesi rekonstrueerimiseks saab kasutada molekulaarseid järjestusi kui tradi?sioonilisi 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 deletsioon ja mitte horisontaalne geeniülekanne (Doolittle *et al*., 2003).

Carl Woese pooldas 1970-ndatel SSU rRNA valimist üldiseks molekulaarseks kellaks (Woese, 1987). Sellel on homoloogid kõikides prokarüootide, mitokondrite, kloroplastide ja eukarüootide tuuma genoomides. Lisaks interakteerub rRNA paljude teiste molekulidega ning

on väga konserveerunud. Järgmise kümne aasta jooksul tõestasid SSU rRNA andmed SEH hüpoteesi (Doolittle *et al*., 2003).

SSU rRNA abil oli võimalik lihtsalt paigutada bakteriliigid taksonoomilistesse rühmadesse, mis üldjoontes vastasid Bergey's klassifikatsioonile. Osutus võimalikuks uurida loodusliku keskkonna mitmekesisust ilma vajaduseta isoleerida, kultiveerida, ega isegi näha ühtegi organismi. 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 klaadi. Üldine fülogeneesipuu SSU rRNA põhjal on kolmeharulise tähe kujuline (*unresolved polychotomy*), st. kõik suured bakterirühmad tekkisid samaaegselt. See võib olla SSU rRNA vähese eraldusvõime tõendiks (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. Mitmes töös tõestati fülogeneetiliste puude abil HGT tähtsust fotosünteesi, aeroobse hingamise, lämmastiku fikseerimise, sulfaadi redutseerimise, isoprenoidide biosünteesi, signaali ülekande, gaasivesiikuli abil pinnal hõljumise ja muude mehhanismide evolutsioonis eubakterites ja arheates (Boucher *et al*., 2003).

3.3 Hetkeseis ja ulatus

Tänu viimasel ajal saadud horisontaalse geeniülekande aktuaalsust tõestavatele faktidele on huvi HGT fenomeni uurimise vastu oluliselt kasvanud.

On loodud genoomide andmebaas HGT-DB, mis sisaldab geenide statistilisi parameetreid nagu G+C sisaldus, koodon- ja aminohapete kasutus (http://www.fut.es/~debb/HGT/). Lisaks sisaldab andmebaas informatsiooni geenide kohta, mis erinevad nendes parameetrites genoomi keskmisest – tõenäolise horisontaalse geeniülekande tagajärg. Seisuga 22.03.2005. on andmebaasis 94 eubakteri ja arhea genoomi. Andmebaas sisaldab otsingumootorit ja organismide nimekirja ning võimalust pääseda ligi iga geeni statistilistele parameetritele (Garcia-Vallvé, 2003). Anomaalsete geenide detekteerimise protseduur põhineb Garcia-Vallvé et al. (2000) artiklil.

A. Tsirigos ning I. Rigoutsos (2005) töö tulemused on saadaval aadressil http://cbcsrv.watson.ibm.com/HGT/. Andmebaas sisaldab 123 prokarüootse liigi genoomide

horisontaalselt ülekandunud geenide andmeid. Kasutatud on genoomi signatuuril põhinevat parameetrilist meetodit “libiseva akna” põhimõttel.

Viimaste aastate kirjanduses on palju diskuteeritud nn. *core* geenide olemasolus. So. genoomi geenide kogum, millel ei ole toimunud HGT (Doolittle *et al*., 2003). Arvatakse, et infotöötlemise mehanismide geenid (replikatsioon, transkriptsioon ja translatsioon) on vähem vastuvõtlikud horisontaalsele ülekandele; metabolismi ja struktuurgeenid – rohkem (Nelson *et al*., 1999; Boucher *et al*., 2003). Osade tööde tulemused näitavad aga, et selline tuum leidub ainult ligidaste liikide vahel (Nesbø *et al*., 2001). Tuumikgeenide puudumisel ei omaks klassikaline evolutsioonipuu mõtet ning evolutsiooni tuleks kujutada võrgustikuna. Siiski Carl Woese (1987) arvamus, et SSU rRNA ei ole vastuvõtlik HGT protsessidele on ilmselt õige ning selle molekuli fülogenees peegeldab tõest evolutsiooni.

4. *In silico* meetodid horisontaalse geeniülekande ennustamiseks

4.1 Meetodite klassifikatsioon

HGT detekteerimise meetodid võib jagada neljaks klassiks (Ragan, 2001a).

- **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 positioonis, 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 HGT tulemuse. Meetodite eelis on võimalus analüüsida eraldiseisvaid genoome suhteliselt vähese arvutamisega. Sellistel meetoditel on suur valepositiivsete ja valenegatiivsete tulemuste protsent (Koski *et al*., 2001), leitud geene tuleb enne horisontaalse ülekande järeldamist analüüsida fülogeneetiliste meetodite abil.
- **fülogeneetilised meetodid** on ajalooliselt peamine viis evolutsiooni käigu vaatlemiseks. 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 sündmuseid. Meetodite puuduseks on arvutustöö mahukus, käsitsi võrdlemise vajadus ning kaugete liikide vaheliste puude ebatäpsus.
- **füleetiline kokkusobimatus**. See meetod on seotud fülogeneesipuude täpsuse suurendamisega. Andmebaasi otsingu põhjal identifitseeritakse nn. füleetiliselt kokkusobimatud järjestused ning eemaldatakse nad fülogeneesipuu analüüsist. Füleetilise jaotuse erinevused võivad olla tingitud horisontaalselt ülekandunud geenist, aga ka liigispetsiifilisest geeni deletsioonist ja duplikatsioonist.
- **ebatavaline sarnasus kaugest liigist pärit geeniga**. 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.

Lisaks eristatakse 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).

4.2 Parameetrilised meetodid

Bakteri genoomi alad võivad ülejäänud genoomist erineda oma omadustelt. Nendeks võib olla aluseline koostis, puriinide ja pürimidiinide suhe, oligonukleotiidide sagedus või koodonkasutus (Ragan, 2001a). Erinevuse põhjuseks võib olla näiteks geeni ebatavaline funktsioon või ekspressioon. Ka horisontaalselt ülekandunud aladel on sageli genoomist statistiliselt usaldusväärne erinevus, mis lubab neid detekteerida.

1. 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 kandidaat 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).

- Võõrad geenid genoomis leitakse tänu anomaalsustele nukleotiidide tasakaalus või koodonkasutuses (*compositional difference* – CD). Normaalsetes geenides on need näitajad tasakaalus genoomi keskmisega. Lisaks esineb amelioratsiooni protsess, mis muudab uued geenid järjest sarnasemaks "omade" geenidega.
- Kodeerivad ja mittekodeerivad alad genoomis erinevad oma ehituselt ja koostiselt, mis lubab neid eristada Markovi mudelite (MM) abil. See võimaldab tuvastada valke kodeerivaid alasid mikroobsetes ja muudes genoomides. Samasuguse mudeli võib teha ainult atüüpiliste geenide leidmiseks.
- Kolmas meetod on põhimõtteliselt erinev. Iga geeni kohta tehakse BLASTP otsing GenBank andmebaasist, tulemused sorteeritakse fülogeneetilise kuuluvuse alusel. Nende geenide tulemused, mille lähedased liigid erinevad keskmisest mustrist, nimetatakse fülogeneetiliselt kokkusobimatuteks (*phylogenetically discordant* - PD).

Selliste geenide arvatav fülogeneetiline puu oleks tunduvalt erinev ülejäänud genoomi puust. Nende eemaldamisel on ülejäänud geenide puu usaldusväärsus tugevam kogu-genoomi kauguste mõõtmistel. Ning HGT on kõige tõenäolisem põhjus, miks geeni topoloogia on erinev ülejäänud genoomist (Clarke et. al 2002).

- Neljas meetod on geenide ebatavaline jaotumine eri organismides (*distribution profile* - DP). Selle põhjuseks võib olla nii HGT kui geenikaotus või erinev järjestuse muutumise kiirus antud liigis. Et eristada HGT sündmusi, rakendati kahepoolse läve kriteeriumit BLASTP tulemustele. Tulemused, mis jäid kahe läve vahele olid tõenäoliselt horisontaalselt ülekandunud geenid.

Mark Ragani töö eesmärk oli võrrelda antud nelja meetodi abil leitud ORF-e. Mitme meetodi abil identifitseeritud geenid on palju tõenäolisemalt horisontaalselt ülekandunud. Juhul kui märkimisväärset kattuvust ei leita, tähendaks see surrogaatsete meetodite vähest väärtust HGT leidmiseks.

Kasutati GenBank *E. coli* K12 järjestusi, J. Lawrence, M. Borodovsky ja R. Charlebois eelnevatest töödest pärit andmeid.

Eri meetodite abil leitud ORFide arv erineb oluliselt (5%-10%), sõltuvalt kasutatud künnise (*threshold*) väärtustest. Iga meetod leiab oma mittejuhusliku komplekti ORFe. Meetodite omavahelist kattuvust võrreldi statistiliselt juhusliku võimaliku kattuvusega. Kõige suurem kattuvus oli CD ja MM meetoditel. Ülejäänud meetodite paarid leiavad tunduvalt vähem ORFe oletatavast juhusliku kattuvuse numbrist. Lisaks leitakse ainult osa transposoonidest ja insertsioonidest. Selle põhjuseks võivad olla amelioratsiooni protsessid, mis ajapikku kaotavad nähtavad erinevused uute ja vanade järjestuste vahel. Seetõttu leiab CD meetod ainult hiljutised horisontaalsed ülekanded, sh. transponeeruvad elemendid. Atüüpiline Markovi mudel leiab osa samadest geenidest kui CD meetud, aga lisaks ka teised "oma" sihtmärgid. Ülejäänud kaks meetodit keskenduvad liikidevahelistele ja domeenidevahelistele erinevustele ja võivad detekteerida vanemaid sündmusi.

Autor järeldab, et selliste meetodite tulemuste ühisosa ei peegelda horisontaalselt ülekandunud geene. Kokku leiavad need meetodid 1691 ORF-i – 39% koguarvust (Ragan, 2001b).

2. Garcia-Vallvé kaasautoritega (2000) analüüsisid 24 täisgenoomi mitme parameetrilise meetodiga. Igale genoomile arvutati keskmine väärtus ning igale geenile standard hälbe (?) väärtus järgmistes parameetrites: koodonkasutus, suhteline sünonüümsete koodonite kasutus

(*relative synonymous codon usage*, RSCU), üldine G+C sialdus (G+C[T]), G+C sisaldus igas koodonipositsioonis (G+C[1], G+C[2], G+C[3]), aminohappeline koostis. Alla 300 bp pikkusega geenid võivad anda kõrvalekalduvaid tulemusi ning neid ei analüüsitud.

G+C sisalduse puhul võõrasteks loeti need geenid, mille puhul hälve oli suurem kui 1,5 standardhälvet genoomi G+C[T] või G+C[1] ja G+C[2] keskmisest väärtusest. Lisaks tehti libiseva akna otsing 11 geeni suuruse aknaga, et leida järjestikku asetsevad võõrad geenid – võõra genoomi piirkond. Kui geen asub kahe võõra geeni vahel, ja selle G+C sisalduse hälbe märk on sama, siis loeti ka see geen võõraks.

Koodonkasutuse analüüsil kasutati Mahalanobis'e kauguse valemit. Võõraks loeti geenid, mille koodonkasutus erines keskmisest üle 2 standardhälbe.

Aminohappelise koostise puhul on ebaselge, mida tähendab kõrvalekalle keskmisest väärtusest. Kõrvalekalle võib olla tingitud geeni funktsioonist või horisontaalselt ülekandunud geeni erinevast koodonkasutusest ja G+C sisaldusest. Seetõttu geenid, mille aminohappeline koostis vähemalt ühe aminohappe puhul oli üle 3 ? eemaldati HGT ennustatud geenide nimekirjast.

Ennustatud HGT geenide hulka kuuluvad need geenid, mis on üle 300 bp pikad, eristuvad võõrana nii G+C sisalduses kui koodonkasutuses ning ei kaldu kõrvale aminohappe sisalduses; ka võõra genoomi alasse kuuluvad geenid ennustati horisontaalselt ülekandunud operoniks.

Autorid said tulemuseks 24 genoomis HGT osakaaluks 1,56% *Borrelia burgdorfer 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). Seisuga 22.03.05 oli andmebaasis 94 liiki.

3. Ühendades *E. coli* MG1655 genoomi järjestuse omaduste põhjal saadud HGT tulemused nelja lähedase liigi fülogeneetilise analüüsiga, märkisid J.G. Lawrence ja H. Ochman (2002) tulemuste ulatuslikku kattumist. Ortolooge määrati >60% järjestuse identsuse ja genoomi konteksti (kõrvalasuvate ORF-ide sarnasus) põhjal. 755-st atüüpilisest geenist olid 627 (83%) HGT-le viitava fülogeneetilise levikuga; ülejäänud 128 geeni olid keskmiselt vaid 628 bp pikad. Kokku leiti 1052 omandatud ORF-i. Neist 425, mis ei ole identifitseeritud atüüpilistena omasid G+C sisaldust, mis on *E. coli* genoomi keskmine (± 3,6%). 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).

4. Genoomi signatuur (GS) on lühikeste oligonukleotiidide järjestuses esinemise sageduste kogum, mis peegeldab mitmeid järjestuse omadusi (Sandberg *et al*., 2001). Oligonukleotiidide sagedused leitakse libiseva akna abil geenile, genoomile, või vaid mõnele regioonile. Seejärel võrreldakse regiooni signatuuri ühe või mitme genoomi signatuuriga. Sageduste sarnasuse hindamiseks kasutatakse kovariatsioon, korrelatsioon analüüsi, arvutatakse Mahalanobis'e ja Eukleides'e kaugus ning tehakse χ^2 test (Tsirigos, Rigoutsos, 2005).

Sõltuvalt oligonukleotiidi pikkusest (2 kuni 9), võib GS sisaldada rohkem informatsiooni järjestuse omadustest, kui muud parameetrilised meetodid. Seetõttu on genoomi signatuuri rakendus horisontaalse geeniülekande detekteerimises potentsiaalselt suurem. Erinevate meetodite võrdlust on läbi viinud A. Tsirigos ja I. Rigoutsos (2005), kes leidsid originaalse testimismeetodi abil, et GS meetod leiab HGT suurema täpsusega kui muud parameetrilised meetodid: G+C sisaldus, koodonkasutus või koodoni adaptatsiooni indeks (CAI). Ka sünonüümsete koodonite kasutuse (RSCU) meetod identifitseerib järjestusi väiksema täpsusega, kui genoomi signatuur (Sandberg *et al*., 2003). Lisaks näitas GS häid tulemusi katsetes, kus tuli leida varem kirjeldatud horisontaalse ülekande juhtumid (Dufraigne *et al*., 2005; Tsirigos, Rigoutsos, 2005).

GS meetodi oluliseks eeliseks on võimalus võrrelda ning klassifitseerida järjestusi vähema arvutusmahuga, kui järjestuste joondamise abil. Horisontaalse ülekande puhul lubab GS leida organismi, millest pärineb ülekantud geen (Sandberg *et al*., 2001; Dufraigne *et al*., 2005). Seejuures ei ole vajalik doonororganismi täisgenoomi või selle konkretse valgu järjestuse olemasolu. Evolutsiooniliselt kaugete HGT sündmuste analüüsimisel GS meetodi täpsus langeb.

4.3 Fülogeneetilised meetodid

Fülogeneetilisteks meetoditeks võib nimetada selliseid meetodeid, mis 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 aeglasem ning täpsem, hinnatakse nukleotiidide või aminohapete asenduse tõenäosust. Üle kümne järjestuse analüüsil muutub arvutamine väga aeglaseks (arvutusmahukus kasvab logaritmiliselt).

Maximum parsimony võtab aluseks minimaalse nukleotiidide asendamise arvu.

1. Carl Woese ja kaasautorite (2000) arvamusel on aminoatsüül-tRNA süntetaasid ilmselt kogenud horisontaalset ülekannet. Nendel geenidel on universaalne funktsioon ja modulaarne ehitus, mis teeb ülekande lihtsamaks. Kuigi ensüümide funktsioon on olemuselt sama, ei ole nende horisontaalse leviku profiil sarnane. Nimelt Asp, Glu, Phe, Leu, Tyr ja Trp aminohapete t-RNA 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).

Autorid kasutasid nelja lähenemisviisi:

a) tavalised fülogeneesi puud. Valkude joondatud järjestustest leiti 1000 tõenäolisemat puud BLOSUM 45 maatriksi järgi. Valiti kõige suurema *maximum likelihood* (Nei, 1996) väärtusega puu, kasutades MOLPHY programmi.

b) visuaalne joondamise hinnang

c) dipeptiidide sarnasuse maatriksid

d) signatuuri analüüs. Signatuuriks nimetatakse antud juhul joonduse positsioone, mis on konserveerunud vähemalt 80% rühma liikmetest.

Leitud HGT sündmused toimusid kaugete liikide vahel varajases evolutsiooni staadiumis.

2. V. Daubin ja kaasautorid (2003) otsisid seost DNA omandamise ja fülogeneetilise puude sobimatuse vahel. Kasutatakse korraga nelja suguluses oleva liigi genoomi järjestust. Iga neliku omavaheline nö. etalon fülogeneesipuu põhineb SSU rRNA puul. Neliku geenid klassifitseeritakse fülogeneetilise jaotumise järgi hiljuti omandatud ja hiljuti kaotatud

geenideks. Identifitseeriti ortoloogid, millel on ainult üks koopia igas genoomis. Testitakse, millist kolmest võimalikust puust toetab iga ortoloogide nelik. Tulemused paigutati kolme kategooriasse järgmiselt:

1) toetab ribosoomi RNA puud;

2) toetab alternatiivset topoloogiat (võimalik HGT);

3) ei anna olulist toetust ühelegi topoloogiale.

Kõigi liikide jaoks on HGT toetavate nelikute 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.

Fülogeneetiline analüüs põhineb *four-cluster minimum evolution* meetodil (Rzhetsky *et al*., 1995). Lihtsa statistilise testiga määratakse nelja monofüleetilise liigi fülogeneesi hargnemise järjekorda. Valitakse kolmest võimalikust juureta puust see, mille harude pikkus on minimaalne.

Tulemused on piiratud suhteliselt ligidaste liikidega, kuid tõstavad esile meetodite artefakte eelnevates HGT töödes. Järeldatakse, et fülogeneesipuude ehitamisel sageli kasutatav *likelihood-mapping* meetod ei anna õiget hinnangut HGT kohta kaugete liikide vahel.

3. Novichkov kaasautoritega (2004) pakuvad uut fülogeneetilist meetodit, mis ei vaja fülogeneesipuude ehitamist või käsitsi võrdlemist.

Töö kirjeldab metoodikat selliste ortoloogsete geenide identifitseerimiseks, mille evolutsiooni kiirus erineb keskmisest. Võrreldakse ortoloogsete geenide vahelisi evolutsioonilisi kauguseid standartsete genoomidevaheliste kaugustega (kahe genoomi kõikide üks-ühele ortoloogide vaheliste kauguste keskmine). Statistilise arvutuse abil leitakse kõrvalekalded molekulaarse kella mudelist. 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 (XGD) – tulemus. Ülejäänud geenide erinevus null-hüpoteesist võib olla seletatud liigispetsiifilise muutusega evolutsiooni kiiruses.

Uuringus osales 21 täisgenoomi, sh. ?-Proteobakterid (6), ?-Proteobakterid (7), *Bacillus*/*Clostridium* (8). Igal rühmal on oma COG komplekt, igas liigis ainult ühe koopiana

esinevad geenid, mille valgu pikkus on vähemalt 60 aminohapet. ?-Proteobakterid – 563 geeni, ?-Proteobakterid – 274 geeni, *Bacillus*/*Clostridium* – 234 geeni. 114 neist on ühised kõigile.

MSA koostamise jaoks kasutati programme MAP ning T-Coffee. Ainult väga konserveerunud osad joondamistest olid kasutatud edasises analüüsis. *Maximum-likelihood* (Nei, 1996) meetodil PAML programmipaketis arvutati kaugused iga COG jaoks ning nende põhjal 114 COG jaoks ehitati fülogeneesipuud Tree-Puzzle programmis. Genoomide omavahelised kaugused olid kasutatud evolutsioonipuude tuletamiseks *neighbor-joining* meetodil (Saitou, Nei, 1987) ning *least-squares* meetodil (Nei, 1996) paketi PHYLIP programmides NEIGHBOR ning FITCH.

4.4 Füleetilise jaotuse meetodid

Fülogeneetiliselt kokkusobimatud järjestused (PDS), mis on Clarke ja kaasautorite (2002) artikli põhiteema, on genoomi need ORF-id, millel on statistiliselt erinev sarnaste järjestuste liikide muster (*similarity relationship pattern*). Suure tõenäosusega on nad horisontaalselt ülekandunud.

HGT genoomis võib moonutada evolutsioonipuu topoloogiat. Kui leida ja eemaldada mittevastavad geenid analüüsist, siis on võimalik ülejäänute põhjal ehitada täpsema puu. See sobib paremini organismide rühmadega ja puudega, mis on tuletatud rRNA geenidest.

Selles töös analüüsisid autorid 28 eubakteri, 8 arhea ja ühe eukarüoodi (*Saccharomyces cerevisiae*) genoomide ORF-e (kokku 81160). Meetod vajab ortoloogsete geenide täpset identifitseerimist. Umbes pooltele ORF-idele (42925) leiti vähemalt 4 BLASTP tulemust teisest liigist (*reciprocal best match*, RBM). Vertikaalselt pärandunud geenid omavad enamasti sarnaseid RBM liikide nimistut ning järjekorda. Kui antud ORF-il ei ole sarnane RBM muster, on ta kokkusobimatu e. PDS. Selliseid järjestusi oli eri genoomidel 6% kuni 16,8%, keskmiselt 10,8%.

42925 ORF-i osales puu ehitamises Fitch-Margoliash *least-squares* (Nei, 1996) paarikaupa kauguste arvutamise algoritmil. Kuigi *pair-wise* BLAST maatriks ei garanteeri homoloogide täpset identifitseerimist, oli tulemuseks üldisele klassifikatsioonile vastav puu. Kasutati FITCH ja CONSENCE programme PHYLIP tarkvarapaketist. Puud ehitati TREEVIEW programmiga.

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 (99% enne, 100% peale eemaldamist). Võib järeldada, et kokkusobimatud ORF-id ei toeta kindlat alternatiivset puu topoloogiat, kuid nende vertikaalne põlvnemine ei ole tõestatud (Clarke et al., 2002).

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

4.5 Kauge liigiga sarnasuse meetodid

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

1. *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. Tal on unikaalsed metaboolsed omadused, mis võivad leida kasutust biotehnoloogilistes ülesannetes (Nelson *et al*., 1999).

Nelson ja kaasautorid (1999) käsitlevad peamiselt genoomi kirjeldamist, leitud geenide tõenäolist funktsiooni ja bakteri metaboliitilisi võimeid. Kõikidele geenidele tehti BLASTX andmebaasiotsing kõikide sekveneeritud genoomide vastu. Tulemused olid üllatavad, sest 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 fuktsiooniga valkudest on sarnasemad arheast pärit homoloogiga. Genoomi järjestuse ?[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).

Tehti ka fülogeneesi uuriming 33 geenile, mille konserveerunud homolooge leiti kõigis 16 sekveneeritud genoomis. Nende põhjal moodustati mitme järjestuse joondus ja fülogeneetilised puud, kasutades *parsimony* ja kauguse meetodeid.

Tulemused näitavad mitut huvitavat mustrit. Saadud puudes langeb enamik geenidest kokku SSU rRNA fülogeneesiga järgmistes punktides: arhead on monofüleetiline grupp (analüüsiti ainult ühte arheate rühma – *Euryarchaeota*), eubakterite rühmad grupeeruvad kokku, pärmi geenid asuvad õigel kohal eubakterite ja arheate suhtes. Samas mitmes kohas on erinevused. Näiteks võrreldes SSU rRNA andmetega ei lange kokku *Aquifex aeolicus* ja *T. maritima* asukoht puul.

2. Paradoxal best hit meetod oli kasutatud Koonini poolt (2001), et detekteerida HGT kandidaadid ja hinnata nende hulka eubakterite ja arhede genoomides kvantitatiivselt.

31 prokarüootse genoomi valkudele tehti anmebaasiotsing BLASTP programmiga. Otsingu tulemuste alusel määrati kolm *paradoxical best hits* kogumit, kasutades SEALS programmi Tax Collector.

Esimene kogum sisaldas HGT kandidaate, mis on fülogeneetiliselt kõige kaugemad. Näiteks 9 arhea liigil need valgud, mille *best hit* on eubakteri või eukarüoodi valk ja selle expect (E)-value on suurem kui arhea valgu *best hit*. Samamoodi olid 22 eubakteri proteoomi jaoks määratud *paradoxical best hits* arheate ja eukarüoodi valkudesse. Eraldi lisati sellesse kogumisse valgud, mille arvestatav *hit* (E<0,001) oli ainult väljaspool oma taksonit.

Vabalt elavates eubakterites on selliste HGT kandidaatide hulk ~3% geenidest. Parasiitsetes eubakterites on see näitaja tunduvalt väiksem, erandiks on *Chlamydia* ja *Rickettsia*. Arheates on see suurus keskmiselt suurem, tüüpiliselt 4% ja 8% vahel. Peab arvestama, et siia satuvad ainult sellised interdomeensed HGT sündmused, mis toimusid evolutsiooniliselt hiljuti, pärast vaadeldavate sama klassi liikide divergentsi. Seega kaugete domeenide vahel on geeni ülekande tõenäosus üsna kõrge.

Teine grupp koosnes HGT kandidaatidest peamiste eubakteri klasside vahel (*Proteobacteria*, *Firmicutes*/gram-positiivsed, *Spirochaetales*).

Kolmas kogum on hiljutiste HGT sündmuste *paradoxical best hits.* Selleks võrreldi lähedalt suguluses olevaid *Chlamydophyla pneumoniae*/*Chlamydia trachomatis* ja *Mycoplasma genitalium*/*Mycoplasma pneumonidae*.

Paradoxical best hits, on isegi konservatiivse künnisega ainult umbkaudne hinnang HGT sündmuste toimumise kohta. Et saada paremat ülevaadet, tehti detailne HGT kandidaatide

analüüs nelja genoomi jaoks: eubakterid *Haemophilus influenzae*, *Vibrio cholerae*, *A. aeolicus* ja arhea *Methanobacterium thermoautotrophicum.* Iga *paradoxical best hit* jaoks määrati antud valgu füleetiline muster ja ehitati fülogeneetilne puu. Kasutati programme ClustalW – järjestuste joondamiseks, ning NEIGHBOR programm PHYLIP paketist *neighbor-joining* meetodil (Saitou, Nei, 1987) – fülogeneesipuude ehitamiseks.

Osadel puudel oli topoloogia liiga keeruline, et saaks eristada XGD (*xenologous gene displacement*) ja uue geeni omandamise sündmuseid, aga paljudel juhtudel oli hästinähtav erinevus.

Suhteline panus XGD ja paraloogsete geenide vahel erines tunduvalt igas neljast genoomist. Esimest on rohkem hüpertermofiilses *A. aeolicus* ja arheas *Mycobacterium thermoautotrophicum*, aga teist parasiitses *V. cholerae* genoomis.

Eukarüootsete geenide osakaal bakterite genoomides oli tüüpiliselt 1% ringis. Erandiks olid mõned parasiidid, nt *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa*, *Xylella fastidosa*, *C. pneumoniae*, kus eukarüootseid geene on keskmiselt rohkem. *Synechocystis* (tsüanobakter, suguluses taimede kloroplastidega) omab suurt hulka taimse päritoluga geene, aga ka looma-spetsiifilisi.

4.6 Arutelu

Vaadeldud horisontaalse geeniülekande detekteerimise meetodide aluseks on väga erinevad põhimõtted. Enamasti otsitakse geene, mis on antud genoomi lõikes ebatüüpilised. Tulemused kõiguvad suures ulatuses sõltuvalt ülekande toimumise ajast ning doonori genoomi omadustest. Geeni järjestuse omadused muutuvad ajapikku ülejäänud genoomi sarnaseks; kui doonorliigi genoom on retsipiendi genoomi sarnane, on ülekantud geen ka sarnaste omadustega. Seetõttu iga meetod identifitseerib oma tunnustega HGT kogumi, mis sisaldab teatud hulga valepositiivseid tulemusi, ning tõenäoliselt jätab välja osa geene mis on antud tunnuse poolest muutunud tüüpiliseks.

Fülogeneesimeetodid sõltuvad fülogeneesipuude kvaliteedist. Samas kasutatakse protsessi kiirendamiseks vähem arvutusi nõudvaid algoritme, mis annavad ligikaudsema tulemuse. Peab märkima, et ka arvutusmahukad algoritmid ei pruugi alati anda õiget tulemust (Nei, 1996). Enamus geene ei sisalda piisavalt evolutsioonilist signaali, mis võib moonutada puu konfiguratsiooni (Daubin *et al*., 2004). Kuid vaid fülogeneetiline uuring võib üheselt tõestada horisontaalse ülekande toimumist.

Viimase aja kirjandus horisontaalse ülekande *in silico* detekteerimiseks on pühendatud uutele meetoditele, mis ei vaja töömahukat fülogeneesipuude uurimist. Laialdaselt kasutatakse otsingut andmebaasidest, mis annab seda paremaid tulemusi, mida rohkem genoome on sekveneeritud.

Parameetrilised HGT detekteerimise meetodid lubavad teha ainult esmaseid järeldusi geeni päritolu kohta, ning leiavad evolutsiooniliselt hilisemaid HGT sündmusi. Samas ei vaja nad arvutusmahukat järjestuste joondamist ja puude ehitamist. Parameetrilisi meetodeid võib kasutada ka sellistel juhtudel, mis ei sobi teistele meetoditele: geeni ortoloogide puudumine, duplitseeritud geenid genoomis. Samuti järjestuste puhul, mis ei ole piiratud geeni mõõtmetega, näiteks on suuremad või väiksemad kui geen. Taoline lähenemine on evolutsioonili seisukohast õige – ülekantud geneetiline materjal võib olla suurem või väiksem kui geen.

Esimesed HGT pühendatud andmebaasid kasutavad just parameetrilisi meetodeid HGT detekteerimiseks. Garcia-Vallvé (2003) tehtud andmebaas on positiivne signaal. Kui toetada selle andmeid teiste (erite fülogeneetiliste) meetoditega saab sellest tähtis ja kasulik genoomika tööriist. Lawrence ja Ochman (2002) märkisid ulatuslikku kattuvust ja tulemuste omavahelist täiendamist parameetrilise (G+C sisaldus) ja fülogeneetilise meetodite vahel.

5. Kokkuvõte

Antud horisontaalse geeniülekande tuvastamise meetodite publikatsioonide ülevaate põhjal võib teha selliseid järeldusi:

- Nüüdseks on horisontaalse geeniülekande tähtsus bakterite evolutsioonis üldtunnustatud, kuid fülogeneetiliste puude ehitamisel sellega ei arvestata.
- Uute genoomide sekveneerimine annab *in silico* meetoditele parema võimaluse horisontaalse geeniülekande tuvastamiseks ja analüüsiks.
- Olemasolevaid HGT detekteerimise meetodeid võib jagada kolmeks grupiks: fülogeneesipuudel põhinevad, andmebaasiotsingul põhinevad ning parameetrilised meetodid.
- HGT detekteerimise tulemusi mõjutavad horisontaalse ülekande aeg, doonori genoomi omadused, parameetrilise detekteerimise puhul erinevad piirväärtused. Seetõttu leiavad eri meetodid "oma" geenide või genoomi alade kogumi ning sellised kogumid kattuvad suhteliselt vähe.
- Tõenäoliste HGT geenide analüüsist eemaldamine fülogeneetiliste puude ehitamisel ei muuda selle topoloogiat, kuid parandab alampuude usaldusväärsuskoefitsienti.
- Võimalik, et standardse horisontaalselt ülekantud geenide *in silico* detekteerimise metoodika väljatöötamisel tuleb kasutada mitme meetodi kombineerimist.
- Senised HGT tööd on olnud suhteliselt eraldiseisvad ja vähese ulatusega. Vajalik on luua neid ühendav andmebaas, kuhu koguda kõigi uuringute tulemused.

6. Summary

Current review of publications about detection of horizontal gene transfer allows to come to following conclusions:

- Importance of horizontal gene transfer in evolution of bacteria is now widely accepted, but is not taken into consideration in construction and analysis of phylogenetic trees.
- Sequencing of new genomes improves ability to detect and analyze horizontal gene transfer with means of *in silico* methods.
- All available HGT detection methods can be divided into three groups: ones that are based on phylogenetic trees, ones that utilize database searches, and parametric methods.
- Results of HGT detection are affected by evolutionary time of gene transfer, by properties of donor genome and by different cutoff values of parametric methods. Different methods find "own" sets of genes or genomic regions and overlapping of such sets is relatively small.
- Removing of potential horizontally transferred genes from construction of phylogenetic tree does not change its topology, but does improve bootstrap values of subtrees.
- To create a standard *in silico* method for detection of HGT combination of several methods may be needed.
- Work on HGT to date was independent and had relatively low span. A database must be created to gather all data of analysis in field of HGT.

Kasutatud kirjandus

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.

Bentley, S.D., Parkhill J. 2004. Comparative genomic struckture of prokaryotes. Annu.Rev.Genet. 38:771-791.

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.

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.J. of Bacteriol. 184:2072-2080.

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.

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

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

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

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. mol. Biol. Evol. 18:404-412.

Lawrence, J.G., Hendrickson, H. 2003. Lateral gene transfer: when will adolescence end? Mol. Microbiol. 50:739-749.

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

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.

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

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.

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.

Ochman, H., Lawrence, J.G., Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299-304.

Perry, J.J., Staley J.T., Lory S. 2002. Microbial Life. Sunderland. Sinauer Associates. p. 253-281, 311-330.

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

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

Rzhetsky, A., Kumar, S., Nei, M. 1995. Four-cluster analysis: a simple method to test phylogenetic hypotheses. Mol. Biol. Evol. 12:163-167.

Saitou, N., Nei, M. 1987. The Neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406-425.

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

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.

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.

Sicheritz-Pontén, T., 2000. Bioinformatic tools for analysis of microbial genomes. Acta Universitatis Upsaliensis, Uppsala. p. 26-27.

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/.

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

Ussery, D.W., Hallin, P.F. 2004. Genome update: length distributions of sequenced prokaryotic genomes. Microbiology 150: 513-516.

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

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.

Woese, C.R., 2002. On the evolution of cells. PNAS 99:8742-8747.

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