TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIA TEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Helle Uibokand

Automatiseeritud PCR praimerite disain bakterite geenidele – genoomipõhine lähenemine

Bakalaureusetöö

Juhendaja Tõnu Margus

TARTU
2007

Sisukord

Lühendid ja mõisted

bp	aluspaar (base pair)
DNA	desoksüribonukleiinhape
gap	järjestuste joondamisel ühte järjestusse insertsioonide/ deletsioonide tõttu tekkiv (tühi) koht
IS	transponeeruv element (Insertion Sequenses)
IUB/IUPAC kood	nomeklatuur nukleiinhapete kirjeldamiseks (International Union of Biochemistry/International Union of Pure and Applied Chemistry)
Mb	10^6 aluspaari
nt	nukleotiid
paraloog	samast liigist pärit järjestused, millel on ühine eellane (tekkinud duplikatsiooni teel)
PCR	polümeraasi ahelreaktsioon (Polymerase Chain Reaction)
16S rRNA	ribosoomi väikese subühiku rRNA sedimentatsiooni koefitsiendiga 16 Svedbergi

Sissejuhatus

Planeedi biomassist suurima osa moodustavad prokarüootsed mikroorganismid. Mikroobiliikide geneetiline, ainevahetuslik ja füsioloogiline mitmekesisus on palju suurem kui taimedel ja loomadel. Tänapäevaks on hinnanguliselt 2 - 3 miljardist bakteriliigist identifitseeritud vähem kui 1%.

DNA sekveeerimise tehnoloogia on teinud kättesaadavaks järjest rohkem bakterite täisgenoome. Sellest tingitult avanevad täiesti uued võimalused, mis nõuavad ka uusi lähenemisi – tervet genoomi hõlmavaid. Üheks enamkasutatavaks meetodiks geenide „õngitsemisel“ ja hübridisatsiooniproovide disainimisel on PCR. Üksikute geenidele PCR praimerite ennustamiseks leidub piisavalt veebi saite ja programme. Sama teenust terve genoomi kõikidele geenidele ei paku aga keegi.

Teoreetilises osas antakse ülevaade bakterigenoomide sekveneerimise olukorrast ja genoomide andmebaasidest. Lisaks antakse ülevaade bakterigenoomide üldistest omadustest ja tutvustatakse PCR praimerite disaini põhimõtteid.

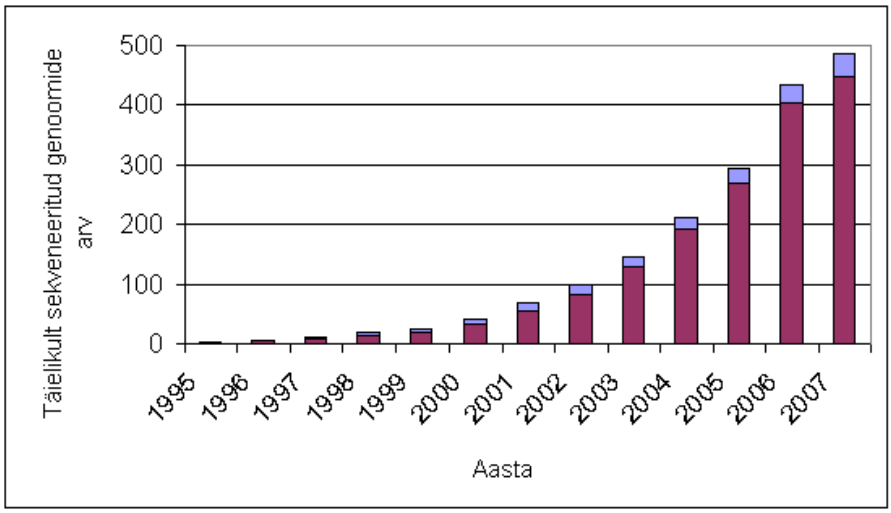
Käesoleva töö praktilise osa peamiseks eesmärgiks oli hinnata PCR praimerite disaini edukust mõjutavaid tegureid. Töö käigus on püütud automaatselt disainida praimereid kõigile bakterite geenidele. Samas pole see 100% õnnestunud ja seega on otsitud ning analüüsitud põhjusid, miks praimerite ennustamine ebaõnnestus. Teises, kvaliteedi hindamise etapis on püütud hinnata disainitud praimerite võimalikku edukust PCR’l, analüüsides praimerite alternatiivseid seondumiskohti ja alternatiivsete produktide tekkimise võimalust.

I Teoreetiline osa

1. Bakterigenoomide sekveneerimine ja genoomide andmebaasid

1.1 Bakterigenoomide DNA primaarjärjestuse määramine

Esimene bakterigenoom, mille DNA primaarjärjestus määrati 1995. aastal oli *Haemophilus influenzae*. Aasta- aastalt on avalikustatud genoomide arv kasvanud (joonis 1) (Celestino *et al*., 2004). 29. aprilli 2007. aasta seisuga oli täielikult sekveneeritud 449 bakteri ja 38 arhe genoomid, mis kuuluvad 348 erinevale liigile (www.ncbi.nlm.nih.gov/).



Joonis 1. Täielikult sekveneeritud genoomide arv aastate lõikes. Punane värv tähistab bakterite ja sinine arhede genoome. Andmed pärinevad NCBI andmebaasist (www.ncbi.nlm.nih.gov/).

Mikroobigenoomide sekveneerimise projekte põhinevad kas ülalt-alla või alt-üles lähenemisel (*Shotgun*). Esimene neist suunatud lähenemine alustab suurte insert-kloonide raamatukogude konstrueerimisest, mis on järjestatud ja sekveneeritud kloon klooni järel. See meetod on ajalooliselt vanem ja ei nõua palju arvutivõimsust, kuid kloonide kaardistamine on aeganõudev. See lähenemine on algselt rakendatud bakteri liikide sekveneerimisel nagu *Mycoplasma pneumoniae* ja *Escherichia coli*. Alternatiivne alt-üles lähenemine põhineb väikeste insert-kloonide raamatukogude sekveneerimisel. “Shotgun” meetodi käigus bakteri genoom purustatakse ca 1 kb pikkusteks juhuslikeks fragmentideks. Need tükid kloneeritakse ja saadakse antud genoomi insertide raamatukogu. Raamatukogust võetakse juhuslikult kloon klooni haaval ja määratakse järjestus. Esimeses faasis järjestuse informatsioon genereeritakse kiiresti, kuid teine faas, kus “gap`id” (kloonide vahelised tühimikud) suletakse, on

aeganõudev. Tüüpiline takistus on suured kordused nagu ribosoomide operonide mitmekordsed koopiad bakteriaalsetes genoomides. Kordustest tingitud probleeme saab lahendada kasutades lisateavet genoomi kaardilt erinevate järjestuste ja markerite vahelise kauguse kohta. kaardi informatsioon toetab järjestuse koostamist ja gapide sulgemist (Weinel *et al.*, 2001).

Kõik genoomid, mille DNA primaarjärjestuse on määratud, pärinevad bakteriliikidelt, mida saab kasvatada laboratooriumi tingimustes või loomsetes rakkudes. Kuid suurt osa baktereid ei saa üldse kultiveerida. Sekveneerimistehnoloogia areng on andnud võimaluse määrata DNA järjestus ilma mikroobide kultiveerimiseta (Fraser *et al.*, 2000; Kemmer *et al.*, 2002). See tehnoloogia on oluline oganismide puhul, mida ei tunta hästi või elavad väga keerulistes keskkonnatingimustes näiteks meri, inimese seedetrakt (Nelson, 2003).

Sekveneeritud genoomid annoteeritakse ja järjestused avalikustatakse sisestades nad andmebaasidesse või eraldi sellele genoomile mõeldud kodulehel.

1.2 Bakterigenoomide andmebaasid

Bakterigenoomide andmed on loomulik osa nukleiinhapete andmebaasi(de)st. Seoses järjest kasvava täielikult sekveneeritud genoomide hulgaga on tekkinud vajadus andmeid genoomi-põhiselt kättesaadavaks teha. Iga suurema andmebaasi juures on võimalik taksonoomiliselt korrastatud genoomide andmeid leida vastava alajaotuse alt. Tabelis 1 on toodud niisugused andmebaasid ja nende veebi aadressid. Nende hulka paigutasime ka TIGR'i, mis tegeleb uurimistöödega, kuid valdab ka andmebaasi, mis sisaldab nii instituudi enda kui ka teiste asutuste poolt sekveneeritud genoome. Suuremate andmebaaside, GeneBank (mida hooldab NCBI), EMBL ja DDBJ, teevad omavahel koostööd, millesse kuulub ka igapäevane andmete vahetamine (Benson *et al.*, 2004).

Tabel 1. Suuremad andmebaasid ja bekterigenoomi uurimise keskused.

Nimi	Aadress
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp
European Molecular Biology Laboratory (EMBL) EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl
National Center for Biotehnology Information (NCBI)	http://www.ncbi.nlm.nih.gov
The Institute for Genomic Reasearch (TIGR)	http://www.tigr.org

Väiksemad andmebaasid on toodud tabelis 2 ja 3. Need on spetsialiseerunud konkreetsetele organismidele/organismigruppidele või teatud tüüpi andmetele. Enamus tebelis 2 näidatud andmebaase sisaldavad genoomide võrdlemisel saadud andmeid. Väga paljud sellised uurimused on keskendunud *Esherichia coli*le, kui mudelorganismile. Näiteks EcoCys, EcoGene, Colibri, GenoBase, ECDC, EchoBase, RegulonDB, coliBase andmebaasid (Roy *et al.*, 2004). Tabelis 3 toodud andmebaasid on spetsialiseerunud konkreetsetele andmetüüpidele, näiteks VirFact sisaldab virulentsusfaktoreid. Kuna andmebaase on palju, siis siinkohal ei ole võimalik neist kõigist juttu teha.

Tabel 2. Konkreetse organismi või organismirühma andmebaasid

Nimi	Aadress
*Bacillus subtilis**	
BSORF	http://bacillus.genome.ad.jp
DBTBS	http://dbtbs.hgc.jp
NRSub	http://pbil.univ-lyon1.fr/nrsub/nrsub.html
SubtiList	http://genolist.pasteur.fr/SubtiList
Campylobacter spp.	
CampyDB	http://campy.bham.ac.uk
Chlamydia trachomatis ja *Chlamydia pneumoniae*	
CIDB	http://www.it.deakin.edu.au/CIDB
Clostridium spp.	
ClostriDB	http://clostri.bham.ac.uk
Cyanobacteria ja teised fotosünteesivad bakterid	
CyanoBase	http://www.kazusa.or.jp/cyano/
Escherichia coli	
ASAP	https://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm
CCDB	http://redpoll.pharmacy.ualberta.ca/CCDB
Colibri	http://genolist.pasteur.fr/Colibri
DPInteract	http://arep.med.harvard.edu/dpinteract
EchoBASE	http://www.ecoli-york.org/
EcoCys	http://ecocyc.org
EcoGene	http://bmb.med.miami.edu/EcoGene/EcoWeb
Essential genes in *E.coli*	http://magpie.genome.wisc.edu/~chris/essential.html
GenoBase	http://ecoli.aist-nara.ac.jp
GenProtEC	http://genprotec.mbl.edu
Pec	http://shigen.lab.nig.ac.jp/ecoli/pec
PromEC	http://bioinfo.md.huji.ac.il/marg/promec
RegulonDB	http://www.cifn.unam.mx/Computational_genomics/regulondb/
Escherichia coli, Salmonella ja Shigella	
ColiBase	http://colibase.bham.ac.uk
Leptospira interrogans serovar Lai	
LeptoList	http://www.bioinfo.hku.hk/LeptoList
Mollicutes	
MolliGen	http://cbi.labri.fr/outils/molligen/
Mycobacterium tuberculosis	
MtbRegList	http://www.USherbrooke.ca/vers/MtbRegList
Rhodobacteria sphaeroides	
RsGDB	ttp://www-mmg.med.uth.tmc.edu/sphaeroides
Pseudomonas aeruginosa	
PseudoCAP	http://www.pseudomonas.com/
Oblikatoorse parasiitsed bakterid	
Metagrowth	http://igs-server.cnrs-mrs.fr/axenic/

*Andmebaasid on jadatud organismide järgi, kuid võivad sisaldada ka sugulasgenoomide andmeid.

Tabel 3. Konkreetsele andmetüübile mõeldud andmebaasid

Nimi	Aadress
GenomeAtlas	http://www.cbs.dtu.dk/services/GenomeAtlas/
EMGlib	http://pbil.univ-lyon1.fr/emglib/emglib.html
FusionDB	http://igs-server.cnrs-mrs.fr/FusionDB/
HGT-DB	http://www.fut.es/~debb/HGT/
Isfinder	http://www-is.biotoul.fr
Islander	http://www.indiana.edu/~islander
ODB	http://odb.kuicr.kyoto-u.ac.jp/
PCTdb	http://pgtdb.csie.ncu.edu.tw/
RRNDB	http://rrndb.cme.msu.edu
VirFact	http://virfact.burnham.org
VFDB	http://zdsys.chgb.org.cn/VFs/main.htm

GenomeAtlas ja HGT-DB sisaldavad sekveneeritud mikroobigenoomide statistilisi parameetreid. HGT-DB's on toodud GC sisaldus, koodoni ja aminohapete kasutus ning lisaks ka informatsioon, millised geenid kalduvad kõrvale nendest parameetritest. GenomeAtlas's on AT sisalduse, geenide arvu, tRNA ja rRNA arvu kohta. rRNA operonide koopiate arv on toodud ka RRNDB andmebaasis.

Islander, VirFact ja VFDB sisaldavad omavahel kattuvaid andmetüüpe: VirFactis ja VFDB bakterite virulentsusfaktorid, VirFactis ja Islanderis on patogeensed saared.

Ülejäänud andmebaasid sisaldavad väga erinevat tüüpi andmeid. EMGlib's leiduvad GeneBank andmebaasi annotatsioonid, FusionDB's Rosetta kivid, Isfinder's IS elemendid, ODB's operonid, PCTdb's prokarüootide kasvutemperatuurid.

2. Bakterigenoomide omadused

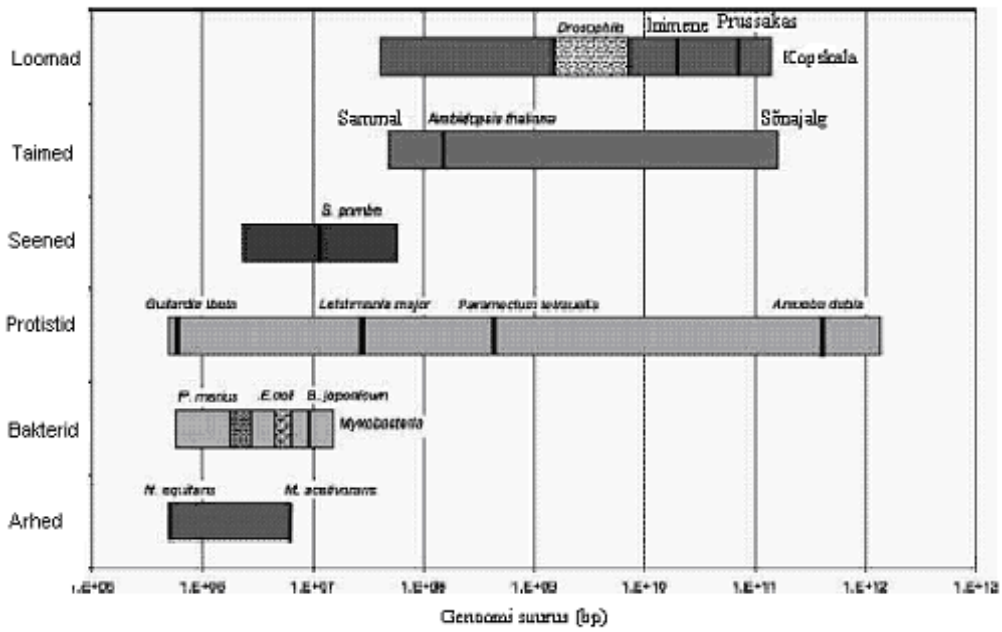
Sekveneeritud järjestused on andnud võimaluse uurida erinevusi genoomide struktuuris ja DNA järjestuse koostises ning neid omavahel võrrelda. Saadud informatsioon lubab paremini aru saada bakterite genoomide korraldusest ja evolutsioonist. Genoome on uuritud mitme omaduse alusel (Liò, 2002). Olulisemad parameetrid on genoomi suurus, replikonide arv, geenide arv ja G+C sisaldus (Bentley *et al.*, 2004).

2.1 Genoomide suurused

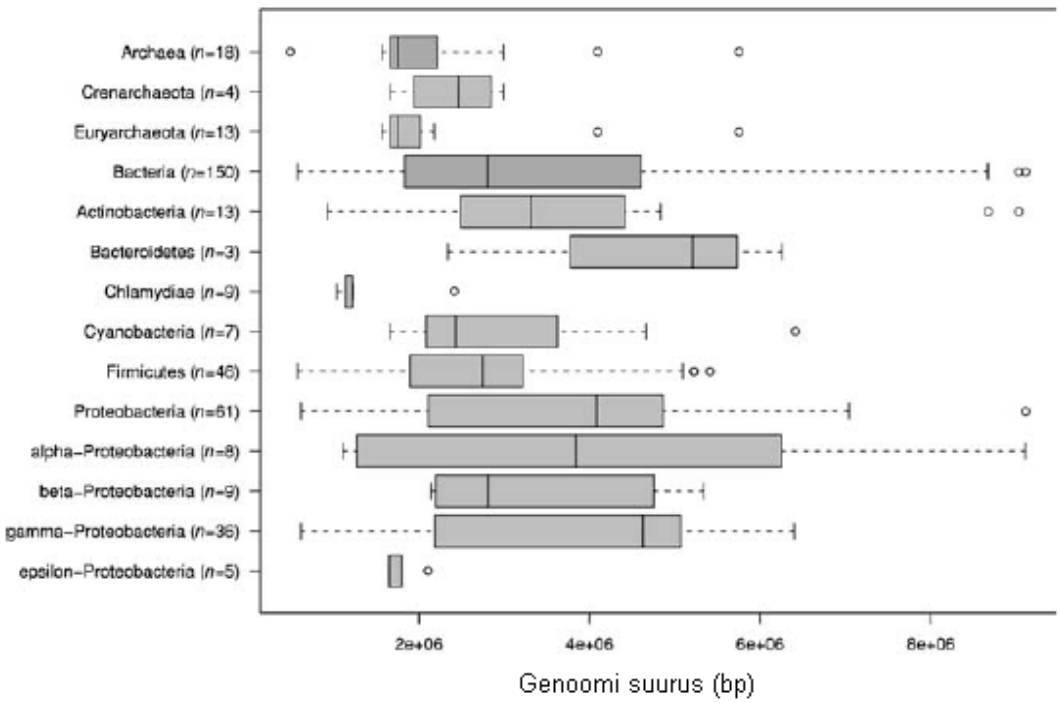
Bakterigenoomide suurused võivad varieeruda kuni 1 suurusjärk. Enamik prokarüootide genoomide suurusi jäävad 0,6 ja 8,6 Mb vahele. Võreldes eukarüootide genoomide suurustega, mis varieeruvad 4 suurusjärku (umbes 10^7 - 10^{11} bp), on see väga kitsas vahemik (Kunin *et al.*, 2003; Mira *et al.*, 2001). Joonis 2 illustreerib kõigi rakuliste

eluvormide genoomide suuruste vahemikke. Mõnede bakterite ja arhede suurused kattuvad väiksemate eukarüootidega ja suuremate viirustega. Prokarüootide seas kattuvad erinevate hõimkondade genoomide suuruste vahemikud üldjoontes (joonis 3), kuid isegi liikide sees võib esineda ulatuslikku varieerumist. Näiteks *Escherichia coli, Prochlorococcus marinus, Streptomyces coelicolor* genoomid varieeruvad rohkem kui 1,000,000 bp. Bakterite genoomi suurus on erinevate geneetiliste sündmuste summa nagu duplikatsioonid, deletsioonid, inversioonid ja horisontaalne geeni ülekanne (Bentley *et al.*, 2004; Mira *et al.*, 2001).

Genoomi suurus ja sisaldus on suurelt osalt tingitud keskkonna poolt (Bentley *et al.*, 2004). Väiksemad genoomid kuuluvad bakteritele, kes elavad stabiilses piiratud keskkonnas ning on sageli seotud peremeesorganismidega. Samal ajal suuremate genoomidega prokarüoodid hõivavad märksa keerukama ja varieeruvama keskkonna nagu muld. Sellises keskkonnas elavatel bakteritel on ka rohkem geene, mis aitavad muutuvates füüsikalistes ja bioloogilistes tingimustes püsima jääda (Bentley *et al.*, 2004; Mira *et al.*, 2001).



Joonis 2. Valik erinevate organismide ja organismirühmade genoomide suuruste vahemikkudest. Lihtsuse huvides sisaldab joonis Protistide riigi rida, mis on defineeritud kui tuuma sisaldavad mikroorganismid ja nende järglased, välja arvatud seened, loomad ja taimed. Joonis on võetud artiklist (Bentley *et al.*, 2004)

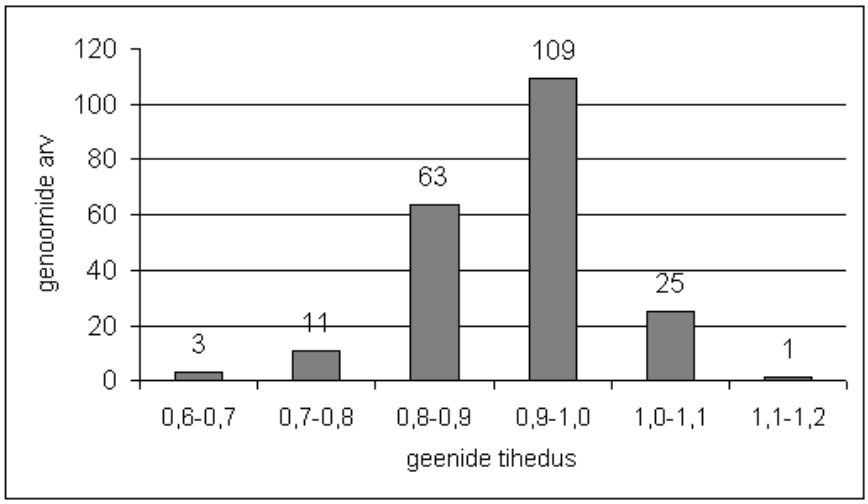


Joonis 3. Sekveneeritud prokarüoodigenoomide suuruste vahemikud. Vahemikud on toodud hõimkondade kaupa. Igal real on märgitud hõimkonna nimi, millele järgneb n väärtus, mis tähistab joonisel kasutatud antud hõimkonna sekveneeritud genoomide arvu. Genoomi suurusevahemike varjutatud osa näitab vaadeldud keskmisi suurusi, vertikaalne joon igas kastis keskmist väärtust ja nn vurrud esindavad pluss või miinus 1,5 kordset keskmiste suuruste vahemikku. Joonis on võetud artiklist (Bentley *et al.*, 2004)

2.2 Geenide arv ja tihedus

Tüüpiliselt sisaldab bakteri genoom 90% kodeerivat järjestust. Võrdluseks võib tuua eukarüootide genoomid, mis sisaldavad <10% kodeerivat järjestust (Salzberg *et al.*, 1998). Geenide arv on prokarüootide seas laialt varieeruv. Keskmine geeni tihedus 1000 nukleotiidi kohta on siiski ~1,0 peaaegu kõigis genoomides (joonis 4) (Rogozin *et al.*, 2002). See näitab selgelt, et genoomi suurus on proportsionaalne geenide arvuga (Bentley *et al.*, 2004). Väiksematel genoomidel on kõrgem geenide osakaal (Ermolaeva *et al.*, 2001) Samas, väiksematel genoomidel on ka vähem geene. Kuid keskkonns, milles vastav mikroob elab, on tema geenide hulk optimaalne (Rogozin *et al.*, 2002).

Geenide arv varieerub liigiti ka + ja – ahela vahel. Enamik geene asuvad + ahelal (McLean *et al.*, 1998). Kõige suuremat asümmeetriat ahelate vahel näitavad madala GC sisaldusega bakterid. Näiteks *Bacillus subtilis´*e 94% olulistest geenidest on kodeeritud + ahelal (Rocha *et al.*, 2003). Geenide arvu erinemine ahelate vahel on tõenäoliselt seotud genoomi stabiilsuse ja PolC erineva kasutusega + ja – ahela sünteesimisel. Liigid, kus puudub PolC näitavad väiksemaid erinevusi ahelate geenide sisalduses (Rocha *et al.*, 2002).



Joonis 4.Genoomide jaotus geenitiheduse lõikes. Geenide tihedus on geenide arv 1000bp kohta. Kokku vaadeldi 212 bakterigenoomi. Väikesema tihedusega genoomid kuuluvad *Ehrlichia ruminantium*ile (tüved: Welgevonden-1, 2 ja Gardel) ning suurema tihedusega genoom oli *Bacillus anthracis*´l (tüvi Ames05819). Joonis põhineb GenomeAtlas (http://www.cbs.dtu.dk/services/GenomeAtlas/) andmetel.

2.3 Replikonide arv

Suuremal osal bakteritel on üks tsirkulaarne kromosoom. Kuigi on leitud ka lineaase ja/või mitme replikoniga baktereid (Bentley *et al.*, 2004). Lineaarne kromosoom on näiteks *Agrobacterium tumefaciens, Streptomyces coelicolos, Borrelia burgdoferi* genoomidel. Kui *streptomyces*´i liikidel on üks lineaarne kromosoom, siis *agrobacteria* ja *borreelia* genoomidel on lisaks lineaarsele ka tsirkulaarne replikon. Lineaarseid plasmiide on isoleeritud eelmainitud gruppidest kui ka teistest ainult tsirkulaarse kromosoomiga bakteritest (Bentley *et al.*, 2004). Enamik mitme replikoniga oganismid kuuluvad *α-Proteobacteria* või *β-Proteobacteria* hõimkonda (Coenye *et al.*, 2003).

Mitme replikoni olemasolu selgitamiseks bakterites on loodud mitmeid hüpoteese. Mitu replikoni võivad olla tekkinud vajadusest saavutada kõrgem üldine replikatsiooni kiirus (Cole *et al.*, 1999). Erinev replikoni koopiate arv võib olla vajalik kindlates tingimustes spetsiifilise geeni ekspressooni tasandil (Heidelberg *et al.*, 2000).

2.4 GC sisaldus

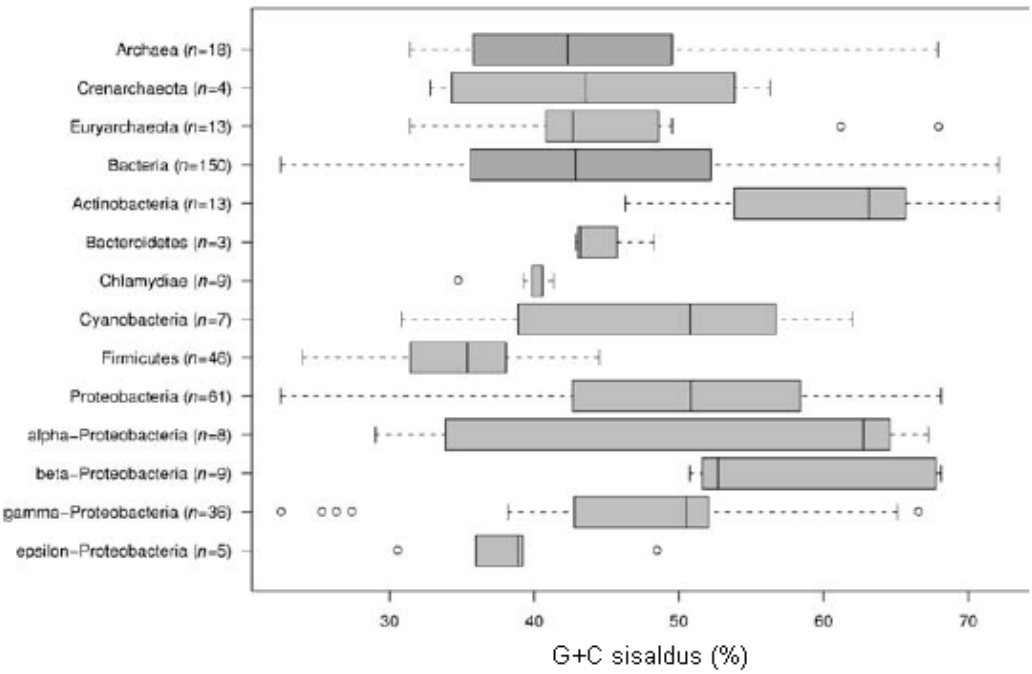
Guaniini ja tsütosiini (GC) sisaldus on oluline parameeter genoomide struktuuri ja evolutsiooni uurimisel ning võrdlemisel. Igale erinevasse liiki kuuluvale organismile on iseloomulik kindel GC nukleotiidide sisaldus. Bakterite genoomide keskmine GC sisaldus

erineb liikide läbilõikes (Liò 2002). Töö kirjutamise ajaks sekveneeritud genoomide GC sisaldus jääb 16,6% [Casonella rudii PV (Nakabchi *et al.*, 2006)] ja 74,9% [Anaeromysobaxter dehalogenen 2CP-C (www.ncbi.nlm.nih.gov/)] vahele. Erinevusi on näha ka hõimkondade kaupa (joonis 5). *Firmicutes* on enamasti AT rikkad, kuid *actinobacteria* GC rikkad (Bentley *et al.*, 2004).

Paljudes genoomides esineb regioone, mis erinevad oluliselt keskmisest GC väärtusest (Ussey *et al.*, 2004). Organismi tavalisest GC väärtusest erineva GC sisaldusega geenid reedavad nende võõrast päritolu (Liò, 2002). Kuid mõnel nukleotiidide sisalduse einevusel on leitud bioloogiline tähendus. Genoomi AT rikkamad alad paiknevad replikatsiooni alguspunktis (*origin*) ja terminuses. Replikatsiooni alguspunktis asub väike AT rikas regioon, mis peab avanema replikatsiooni algatamiseks, kuid laiemal skaalal on alguspunkt GC rikkam (Ussery *et al.*, 2004).

Nukleotiidide sisaldus on erinev ka DNA ahelate vahel. Üldiselt on guaniini (ja tavaliselt tümiini) sisaldus suurem + ahelal (Lobory *et al.*, 1996). Asümmeetria tekkepõhjuseks peetakse ahela- spetsiifilisi mutatsioone, kuid kindlat mehhanismi pole leitud (McLean *et al.*, 1998).

GC sisaldus on üks parameetreid, mis mõjutab küllaltki palju parimerite disaini ja praimerite headust (Haas *et al.*, 1998; Kämpke *et al.*, 2001).



Joonis 5. Sekveneeritud prokarüoodigenoomide G+C sisalduse vahemikud. Vahemikud on toodud hõimkondade kaupa. Igal real on märgitud hõimkonna nimi, millele järgneb n väärtus, mis tähistab joonisel kasutatud antud hõimkonna sekveneeritud genoomide arvu. Vahemike varjutatud osa näitab vaadeldud keskmisi suurusi, vertikaalne joon igas kastis keskmist väärtust ja nn vurrud esindavad pluss või miinus 1,5 kordset keskmiste suuruste vahemikku. Joonis on võetud artiklist (Bentley *et al.*, 2004)

2.5 Kordusjärjestused

DNA kordusi võib määratleda kui järjestusi, mis on sarnased teise järjestusega samast genoomist. Arvatavasti tekivad kordused õnnestunud duplikatsiooni või mõne muu mehhanismi tagajärjel näiteks insertsioon või transpositsioon.(Achaz *et al.*, 2002)

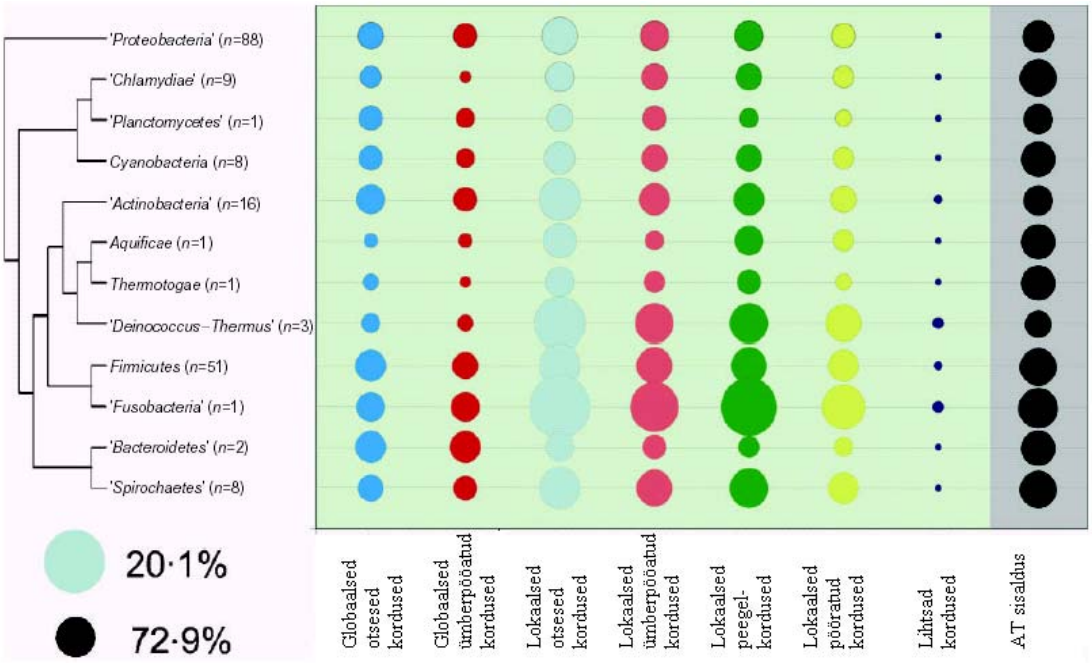
Bakterite puhul on kordused jagatud tavaliselt kahte alamklassi: madala keerukusega kordused ja pikad kordused. Esimese kategooria moodustavad lühikesed tandeemsed kordused (mikrosatelliidid). Teise klassi kuuluvad pikad tandeemsed kordused, speisser kordused, transponeeruvad elemendid. Mikrosatelliidid on väga levinud eukarüootides, kuid bakterite ja arhede genoomides on neid palju vähem (joonisel 6 ‘lihtsad kordused’) (Achaz *et al.*, 2002; Ussery,Binnewies *et al.*, 2004).

Kordusjärjestusi võib jagada ka otsesteks kordusteks (*direct repeats*), ümberpööratud kordusteks (*inverted repeats*) (kus DNA järjestus on ümberpööratud ja asub vastasahelal), peegelkordusteks (*mirror repeats*) (kus DNA järjestus on ümber pööratud ning asub samal ahelal) ja pööratud kordusteks (*everted repeats*) (kus 5′→3′ DNA järjestus on korduv teisel ahelal 3′→5′ suunas). Korduseid jaotatakse ka globaalseteks ja lokaalseteks (Jensen *et al.*, 1999). Joonisel 6 on näidatud erinevate korduste tüüpide keskmised väärtused 12 hõimkonna

kohta. Vaatamata vähestele genoomidele mõnes hõimkonnas, on näha üldisi arengusuundi. Globaalsete korduste arv bakterite seas on päris väike. Ainult kahe hõimkonna genoomidel on rohkem kui 5% kordusi, mis kattuvad 80% või rohkem 100 bp aknas. Paljudes eukarüoodigenoomides on see arv suurem kui 50% (Ussey, Binnewies *et al.*, 2004). Ümberpööratud järjestusi esineb enamikes kromosoomides vähem kui otseseid kordusi. Seega suhe ümberpööratud ja otseste korduste vahel on peaaegu alati < 1. Erandiks on näiteks Neisseria genoom. Selle põhjuseks võib olla selektsiooni ebaefektiivsus otseste ja ümberpööratud korduste vahel, sest korduste tihedus on kõrge.(Achaz *et al.*, 2003)

Kordusi uurides on leitud positiivne korrelatsioon koopiate arvu ja kromosoomi suuruse vahel ning kordusjärjestuse suuruse ja kromosoomi suuruse vahel. Seega, mida pikem kromosoom, seda rohkem kordusi ja seda pikemad nad on.(Achaz *et al.*, 2002)

Kordused on töös olulised, kuna nad mõjutavad PCR praimerite headust (Li *et al.*, 1997).



Joonis 6. DNA koduste võrdlus 189 bakterigenoomis. Ringi pindala näitab koduste osa genoomis. Suur värviline ring tähistab ligikaudu 20% kordusi st. 20% genoomil on kordused, mis kattuvad genoomse järjestusega antud aknas vähemalt 80%. Viimases tulbas on toodud mustade ringidega AT sisaldus. Võrdlus on esitatud 12 hõimkonna kohta. Hõimkonna nimede ees on 16S rRNA fülogeneetiline puu ning järel joonisel kasutatud genoomide arv. Joonis on võetud artiklist (Ussery, Binnewies *et al.*, 2004).

2.6 Genoomsed ümberkorraldused

Mikroobide genoomid on läbi teinud korduvad ümberkorraldused. Sama bakteri 2 tüve võivad omada olulise arvu geene, mis on ühes aga mitte teises tüves (Ermolaeva *et al.*,

2001). Genoomid on kujunenud läbi evolutsiooniprotsesside nagu geenide teke, horisontaalne ülekanne ja geenide kadumine (Kunin *et al.*, 2003).

Geeni kaotus prokarüootides võib vähendada genoomi suurust. See on hästi näha, kui võrrelda parasiitset bakterit vabalt elava sugulasega (Rogozin *et al.*, 2002). Vähendatud genoomil on eelistatult kadunud geenid, mis on seotud aminohapete, nukleotiidide ja vitamiinide biosünteesiga, kuna neid aineid saab üle võtta peremeesorganismist (Konstantinidis *et al.*, 2004). Näiteks *Mycoplasma* liike võrrelda *Bacillus- Clostridium* gruppiga. Genoomide redutseerumisel võib olla ka teisi põhjusi peale parasiitse eluviisi (Rogozin *et al.*, 2002).

Geenide duplikatsioon on tähtis roll uute geenide ja uute biokeemiliste funktsioonide evolutsioonis (Jordan *et al.*, 2001; Kondrashov *et al.*, 2002). Vastavalt loodud hüpoteesile on igal geenil optimaalne koopiate arv genoomi kohta, mis võib varieeruda sõltuvalt keskkonnast. Leiti, et suuremal osal paraloogidel (sugulasgeenid samast genoomist), mis on fikseerunud populatsioonis, on otsene mõju kohanemusele erinevates keskkonnatingimustes (Kondrashov *et al.*, 2002).

Geenijärjestuste ülekanded ühest liigist teise on sageli reegel kui erand (Bansal *et al.*, 2002). Osad prokarüoodi geenidest kuuluvad operonidesse. Täieliku operoni horisontaalne ülekanne on valiku poolt eelistatud üksikute geenide ülekandele (Rogozin *et al.*, 2002). Geeni horisontaalne ülekanne on mänginud olulist rolli patogeensetes bakterites. Paljudele patogeensetele bakteritele on omased nn. „patogeensed saarekesed". Need saarekesed eristuvad oma nukleotiidse koostise poolest (GC%, dinukleotiidide jaotus, koodi kasutus) ülejäänud genoomsest DNA'st. Arvatakse, et need piirkonnad on omastatud horisontaalse ülekande teel. Kuid horisontaalselt ülekantud geenide osakaal on patogeensetel bakteritel väiksem, kuna patogeensete bakterite genoomide vähendamine on evolutsiooniprotsessis domineeriv. Teine erinevus patogeensete ja mittepatogeensete bakterite vahel võib olla elupaik. Liikidel, mis on laia levialaga nagu *Escherichia coli* ja *Bacillus subtilis* on suurem võimalus geenide vahetamiseks (Garcia-Vallvé *et al.*, 2000). Samal ajal toimub rakusisestes bakterites homoloogsete geenide vaheline rekombinatsioon. Sellega võib seletada genoomi vähenemist evolutsiooni jooksul (Achaz *et al.*, 2003).

Hoolimata kõigist ümberkorraldustest mõni bakterigenoom on suhteliselt stabiilne. Buchnera sp. genoom esindab üht stabiilsemat genoomijärjestust (Kunin *et al.*, 2003).

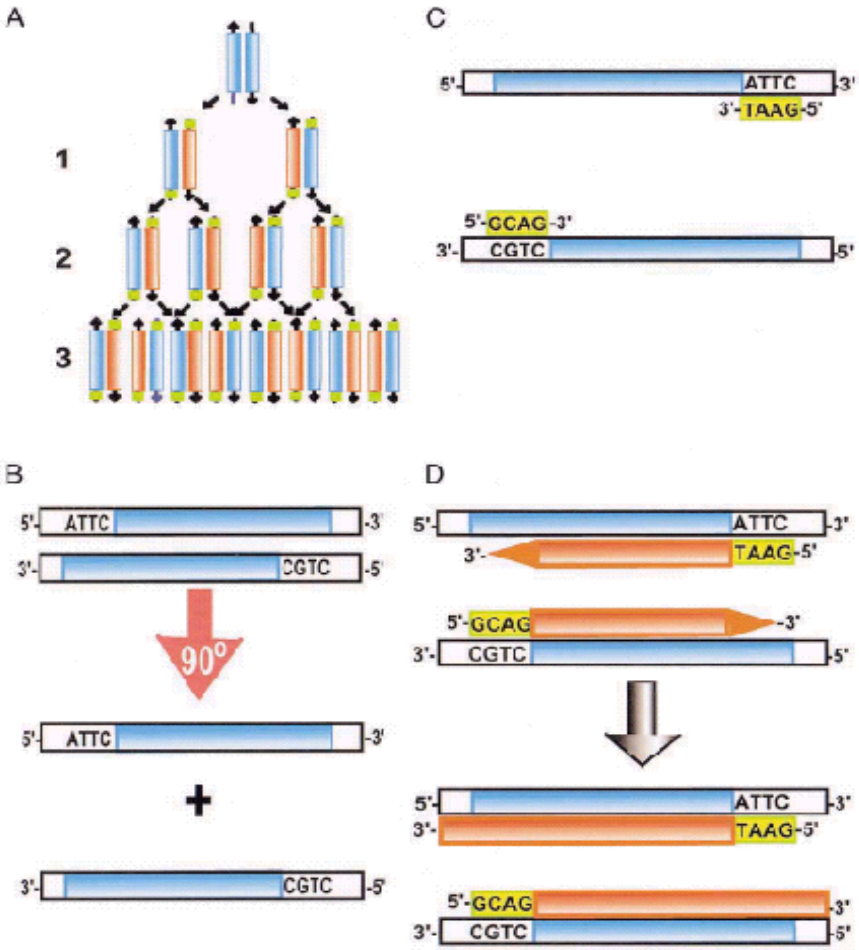
3. PCR praimerite disain

3.1 PCR põhimõte

Aastal 1985 esitleti polümeraasi ahelreaktsiooni (PCR) tehnoloogiat, mis on muutunud molekulaarteaduse lahutamatuks osaks (Kim *et al.*, 2002). PCR'i eesmärk on kordistada spetsiifilist DNA fragmenti (Kämpke *et al.*, 2001). Identseid koopiaid saab kasutada mitmeti. Kõige sagedamini kasutatakse PCR, et amplifitseerida kindel kogus spetsiifilist geeni, mille külgnev järjestus on teada (Kim *et al.*, 2002). PCR'i rakendatakse DNA sekveneerimisel, genoomse DNA kloonimisel, mutatsioonide ja geeni ekspressiooni uurimisel (Bermingham *et al.*, 2003).

PCR koosneb kolmest korduvast reaktsioonist. Esimeses reaktsioonis kaheahelaline DNA lahutatakse üheahelaliseks (joonis 7B). Tavaliselt toimub denaturatsioon 90^{o}C ja kestab 1-2 minutit. Järgmises reaktsioonis jahutatakse DNA 50-60^{o}C –ni, et saaks toimuda praimerite seostumine järjestusele (joonis 7C). Kolmandas rektsioonis sünteesitakse uus DNA (joonis 7D). Harilikult kulgeb see reaktsioon 72^{o}C tingimustes (Kim *et al.*, 2002).

Kui PCR on 100% efektiivne, siis pärast n tsüklit saab ühest molekulist/järjestusest 2^{n} DNA koopiat (Mullis, 1990). Praktikas kasutatakse tavaliselt 20-40 tsüklit (Bermingham *et al.*, 2003).



Joonis 7. Polümeraasi ahelreaktsioon. A. DNA(sinine) amplifitseeritakse igas PCR tsüklis. Geenispetsiifiliste praimeritega(kollane) sünteesitakse uus DNA(punane). B. Iga tsükkel algab kaheahelalise DNA lahutamisega üheahelaliseks. C. Järgmises reaktsioonis seostuvad praimerid DNA'le. D. Kolmandas reaktsioonis sünteesitakse uus DNA. Joonis on võetud artiklist (Kim *et al.*, 2002).

3.2 PCR praimerite ennustamisel arvestatavad parameetrid

PCR's kasutatavad praimerid peavad vastama kindlatele kriteeriumitele. Olulisemad parameetrid on praimerite sulamistemperatuur, praimerite pikkus ja GC nukleotiidide sisaldus. Lisaks jälgitakse, et praimer ei kleepuks iseenda külge, teiste praimerite külge, ega omaks alternatiivseid seondumiskohti genoomis (Li . *et al.*, 1997).

3.2.1 Praimerite sulamistemperatuur

Sulamistempeatuur (T_m) on defineeritud kui temperatuur, mille juures pool DNA dupleksist on paardunud ja pool on mitte (ahelad on üksteisest lahus) (Wetmur *et al.*, 1991). Sulamistemperatuuri arvutamiseks on mitmeid võimalusi. Kõige lihtsaim neist on valem, kus G või C nukleotiid lisab 4^o C ning A või T 2^o C. Valem: (Suggs *et al*., 1981)

$$T_m = 4 * (G + C) + 2 * (A + T)$$

Täpsemaks arvutamiseks kasutatakse *Nearest-Neigbor* meetodit. Valem:

$$T_m = \Delta H / [\Delta S + R * \ln(c/4)] - 273{,}15\ ^oC + 16{,}6 \log_{10}[K^+],$$

kus ΔH ja ΔS on paimer-DNA dupleksi moodustumise entalpia ja entroopia, mis on arvutatud *nearest neighbour* skeemi järgi, R = 1.987 (cal/^{o}C*mol) = 8,31 (J/mol) on universaalne molaarse gaasi konstant, c on oligonukleotiidide molaarne kontsentratsioon lahuses (Breslauer *et al.*, 1986).

3.2.2 Praimerite pikkus ja GC sisaldus

Praimeri pikkus ja GC sisaldus on positiivses korrelatsioonis sulamistemperatuuriga (Haas *et al.*, 1998). PCR praimeri pikkus peaks olema 14-40 nukleotiidi ja GC sisaldus ligilähedane 50% (Bermingham *et al.*, 2003). GC sisalduse arvutamise teeb oluliseks asjaolu, et GC nukleotiidide vahel on kolm vesiniksidet samal ajal AT nukleotiidide vahel on kaks vesiniksidet (Kämpke *et al.*, 2001).

Praimeri seondumine DNA järjestusega sõltub tema 3´ otsa seondumise stabiilsusest. On leitud, et stabiilse 3´ otsa tagab G või C nukleotiidid (Li *et al.*, 1997). Kuid mõned testid on näidanud, et praimerite 3´ otsa nukleotiidide vahel pole kvalitatiivseid erinevusi (Haas *et al.*, 1998).

3.2.3 Praimerite seondumised

PCR edukust mõjutavad oluliselt praimerite dimeeride ja sekundaarstruktuuride moodustumine. Iga praimeri on vaja testida, et nad ei hübridiseeruks iseendaga,

vastaspraimeriga, praimeri 3` otsa seostumist enda ja vastaspraimeri 3` järjestusega (Kämpke *et al.*, 2001).

Samuti on tähtis, et praimeritel oleks unikaalne seondumiskoht ning ei tekiks lisaprodukte genoomse DNA peal. Alternatiivsete seondumiskohtade põhjuseks on praimerid, mis sisaldavad DNA kordusjärjestuste motiive. Praimerite unikaalsuse tagamiseks tuleks vältida korduvaid elemente (Li *et al.*, 1997).

3.3 Praimereid ennustavad programmid

Praimerite disain on paljudes eksperimentides üheks oluliseks aeganõudvaks etapiks. Ennustamise protsessi optimeerimiseks on loodud mitmeid programme (van Baren *et al.*, 2004). Suuremahulistes projektides on rakendatud PRIMER3 (Rozen *et al.*, 2000), PRIDE (Haas *et al.*, 1998), GST-PRIME (Varotto *et al.*, 2001), PRIMER MASTER (Proutski *et al.*, 1996), PRIMO (Li *et al.*, 1997) jt. Väikese arvu praimerite disainimiseks on kasutatud DoPrimer (Kämpke *et al.*, 2001), Oligo (Rychlik *et al.,* 1989) jt. Programme võib jagada selle järgi, kui automatiseeritud nad on. See tähendab, kui palju peab kasutaja sekkuma nende töösse. Näiteks kasutaja vahelesegamine on minimaliseeritud PRIDE programmis (Haas *et al.*, 1998). Osad programmid on mõeldud kindlale sihtgrupile. Näiteks PRIMEGENES (Xu *et al.*, 2002) on kasutatud terve genoomi analüüsimiseks mikrokiipidel.

Tabel 4. Praimereid ennustavad programmid ja nende veebi aadressid.

Nimi	Aadress
DoPrimer	http://doprimer.interactiva.de
MEDUSA	http://www.cgr.ki.se/cgr/MEDUSA
OLIGO	http://www.oligo.net/
PRIDE	http://www.dkfz-heidelberg.de/tbi/services/Pride/prideform
Primer3	http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
Primer Design	http://www.embl-heidelberg.de/%7Etoldo/JaMBW/5/2/index.html
PRIMER MASTER	ftp://ftp.ebi.ac.uk/pub/software/dos/primer-master/
Primer Selection	http://alces.med.umn.edu/websub.html
PRIMO	http://bioweb.pasteur.fr/seqanal/interfaces/primo.html
Web-Primer	http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer

II Praktiline osa

4. Töö eesmärgid

Käesoleva töö peamiseks eesmärgiks oli hinnata PCR praimeri disaini edukust mõjutavaid tegureid. Esiteks on töö käigus püütud disainida praimerid kõigile bakterite geenidele. Leitud praimereid saab kasutada bakteri geenide amplifitseerimiseks, kloneerimiseks ja/või hübridisatsiooni proovide tegemiseks. Teiseks püütakse leida põhjused, miks mõnedele geenidele praimerite ennustamine ebaõnnestus. Kolmandaks püütakse hinnata disainitud praimerite võimalikku edukust PCR'l, analüüsides praimerite alternatiivseid seondumiskohti ja alternatiivsete produktide tekkimise võimalust.

5. Materjal ja metodika

5.1 Andmete päritolu ja struktuur

Andmed bakterite genoomide kohta pärinevad NCBI Refseq (veebruar 2005) andmebaasist (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Töös kasutati bakterite genoomide järjestusi ja geenide asukohtade kirjeldusi. Neist esimesed olid fasta formaadis ühtsete järjestustena ja kirjeldused eraldi failides, mille näide on toodud joonisel 10.

Käesolevas töös kasutati 211 bakteri genoomi järjestust. Bakteri genoomid võivad sisaldada mitut geneetilist elementi (kromosoomi, plasmiidi), kuid enamus töös kasutatud genoomidest koosnes ühest kromosoomist. Plasmiide ja mitut kromosoomi sisaldavaid genoome oli 68. Kõige rohkem plasmiide (21) kuulus *Borrelia burgdorferi* genoomi.

Escherichia coli K12, complete genome - 0..4639221
4311 proteins

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
190..255	+	21	16127995	thrL	b0001	-	-	thr operon leader peptide
337..2799	+	820	16127996	thrA	b0002	E	COG0460	aspartokinase I
2801..3733	+	310	16127997	thrB	b0003	E	COG0083	homoserine kinase
3734..5020	+	428	16127998	thrC	b0004	E	COG0498	threonine synthase
5234..5530	+	98	16127999	-	b0005	-	-	hypothetical protein
5683..6459	-	258	16128000	yaaA	b0006	S	COG3022	hypothetical protein
6529..7959	-	476	16128001	yaaJ	b0007	E	COG1115	inner membrane transport protein

Joonis 10. *Escherichia coli* K12 geenide kirjelduste faili NC_000913.ptt 10 esimest rida. Fail sisaldab geenide algus- ja lõppkoordinaate (Location), geenide asukohti ahelal (Strand), geenide identifikaatoreid (PID), geenide nimesid (Gene, Synonym, Code, COG) ja geenide produkte (Product).

5.2 Praimerite ennustamiseks kirjutatud skriptid ja valmis programmid

Käesoleva töö võib põhimõtteliselt jagada kolmeks osaks (joonis 11). Esimeses osas disainitaks geenidele PCR praimerid, teises püütakse leida põhjused, miks praimerite ennustamine ebaõnnestus ja kolmndas analüüsida õnnestunud praimerite alternatiivseid seondumiskohti ja alternatiivsete produktide tekkimise võimalusi.

Täielikult sekveneeritud genoomides kirjeldatud geenidele praimerite ennustamise automatiseerimiseks kasutasime programmi nimega PROGENE.pl (joonis 11 Praimerite disain). PROGENE.pl peamisteks ülesanneteks on ennustamiseks vajalike algandmete pärimine kohalikust andmebaasist, praimerite konstrueerimine geenidele ja teiste programmide töö koordineerimine. Neid programme oli kokku kolm. Esimene programm LISTING.pl tekitab genoomi nimest ja NCBI poolt antud genoomi unikaalsest identifikaatorist nimekirja, mille alusel hakkab toimuma andmete lugemine ja praimerite disain. Praimerite disainimise programmina kasutasime PRIMER3 (primer3_core). Enamus PRIMER3 parameetrite väärtuseid ei muudetud, vaid kasutati vaikimisi antud väärtusi. Muudetud väärtused on toodud tabelis 6. Kolmas programm TABLE.pl tegi ennustamise tulemustest tabeli.

Praimerite konstrueerimiseks lahutasime geeni alguskoordinaadist 30 nukleotiidi ning geeni lõppkoordinaadile lisasime 30 nukleotiidi praimerite seondumiskohtadeks. 30 nukleotiidi on küllaldki lühike ala ja ei pruugi sobida praimeri disainiks. Seetõttu korrati ennustust neile geenidele, millele jäid praimerid leidmata ja sealjuures suurendasime praimerite otsimisala veel 30 nukleotiidi võrra. Tsüklit korrati niikaua kuni kõikidele geenidele olid praimeripaarid leitud või 10 korda ei suudetud samale arvule geenidele praimeripaare teha.

Praimerite disaini ebaõnnestumise analüüsi jaoks (joonis 11 B) kasutasime PRIMER3 omadust näidata praimerite mittesobivuse põhjuseid. Õnnestunud disaini puhul otsitakse üles kõik PCR praimerite seostumiskohad genoomis ja ennustatakse alternatiivsete produktide teket (joonis 11 A). Praimerite alternatiivsete seostumiskohtade otsimiseks kasutasime FASTAGREP'i. Teostasime nii täpse otsingu kui ka ligikaudse, lubades 1-2 valepaardumist. Mõlemal juhul jälgisime seondumiskohtade arvu >10. Eukarüootide puhul on tõdetud, et praimerite seostumisel üle 10 korra, langeb PCR kvaliteet sedavõrd, et genoomilt produkti ei saada. Praimeri täpse seostumise korral arvutasime produktide tekkimise võimalust. Selleks lahutasime DNA vastasahela praimeri koordinaadist kodeeriva ahela praimeri koordinaadi ning produkti pikkus ei tohtinud ületada neljakordset vastava järjestuse keskmist geeni pikkust.

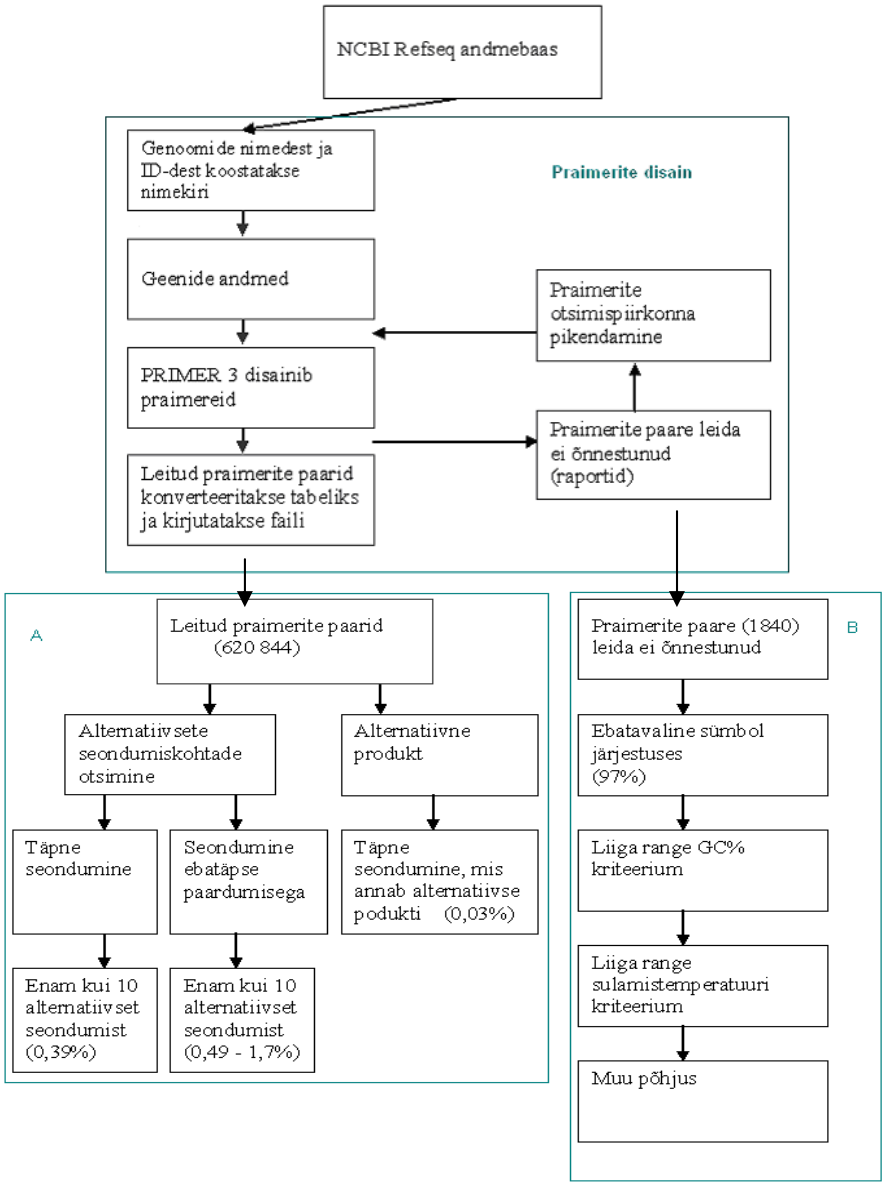
Tabel 6. Parameetrite väärtused, mida kasutas PRIMER3 praimerite ennustamisel.

Parameeter	Minimaalne	optimaalne	maksimaalne
praimeri pikkus	16 nt	21 nt	30 nt
Produkti pikkus	100 nt	600 nt	20000 nt
sulamistemperatuur	50^{O} C	55^{O} C	60^{O} C
maksimaalne temperatuuri erinevus praimeripaaride vahel 4^{O} C			
Soola kontsentratsioon 20mM			
maksimaalne lubatud mononukleotiidide kordus 4			

5.3 Programmide tööaeg

Praimerite automaatseks ennustamiseks kasutatud programmi PROGENE.pl tööaeg oli 4 tundi ja 11 minutit. Tuhande geeni läbiprotsessimiseks kulus umbkaudu 24sekundit.

Kuid töö kõige aeganõudvam osa jäi kvaliteedi hindamise etappi ja seisnes praimerite seondumiskohtade otsingus. FASTAGREP'il kulus selleks ligikaudu nädal. Lubades 1-2 valepaardumist suurenes tööaeg veelgi. Kahe valepaardumise korral kulus FASTAGREP'il kuni 2 nädalat seondumiskohtade otsinguks.



Joonis 11. Bakalaureusetöö praktilise osa üldskeem. Skeemil on kirjeldatud töö järjekorda. Esmalt toimub praimerite disain, ning järgneb saadud tulemuste analüüs, mis on jagatud kahte ossa. A. Kirjeldatakse leitud praimerite paaride PCR kõlblikkuse analüüsi. B. Otsitakse praimerite mitteleidmise põhjusi.

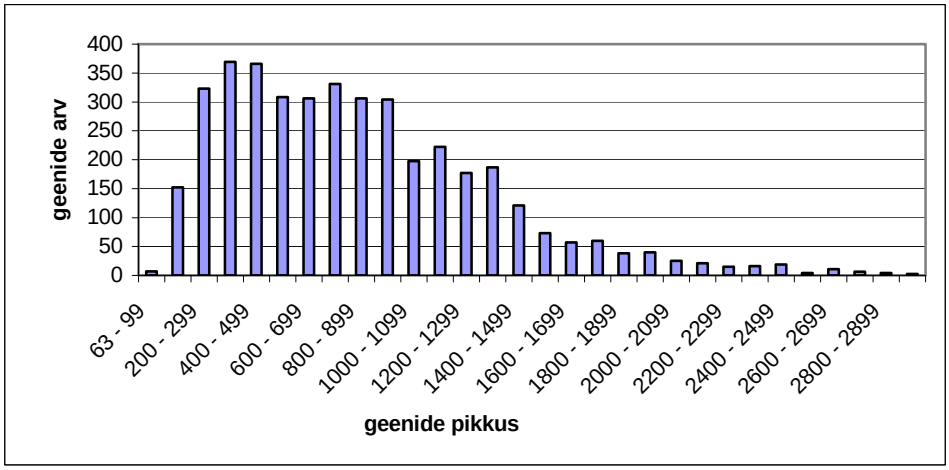
Praimerite automaatseks disainimiseks kasutataks PROGENE.pl programmi (joonisel „Praimerite disain“), mille töö algab praimerite ennustamisel osalevate geenide nimest ja ID’st nimekirja koostamisega (LISTING.pl). Järgmisena loetakse genoomijärjestuste andmed, disainitakse praimerid (primer3_core) ja kirjutatakse leitud praimerite fail ja leidmata jäänud praimerite fail. Kui kõiki praimeripaare ei leitud, pikendatakse otsimisala ning minnakse tsüklisse, millest väljutakse, kui praimerid leitud või kümnel korral ei leitud samale arvule geenidele praimeripaare.

6. Tulemused

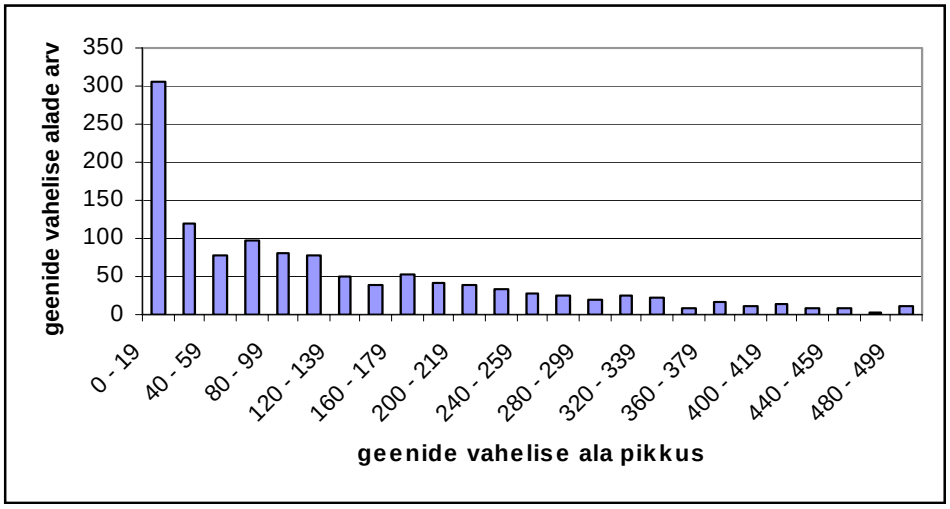
6.1 Bakteriaalsete geenide kirjeldus

Geenide arv genoomi kohta varieerus vahemikus 484 kuni 8317. Väiksema geenide arvuga genoom kuulus *Mycoplasma genitalium*ile ja suurem *Bradyrhizobium japonicum*ile. Enamus genoomides sisalduvate geenide arv jäi 1500- 2500 vahemikku ja keskmine geenide arv genoomi kohta oli 2956.

Kui võrrelda geenide ja geenide vaheliste alade suurusi, siis enamik geene oli pikemad kui nende vahelised alad. Suurem osa geenide pikkuseid jäid vahemikku 200- 1000 aluspaari. Joonisel 12 on toodud vaadeldud genoomidest tüüpiline geenide pikkuste jaotus. Geenide vaheliste alade pikkused jäävad alla 20 aluspaari nagu on näha joonisel 13, mis on samuti tüüpiline näide. Samas selgus, et enamikes genoomides esines kõige rohkem geenide 4 aluspaariline ülekattumine. Kuid keskmine geeni pikkus on väiksem kui keskmine geenide vahelise ala suurus. Vastavad arvud on 93 ja 14445 aluspaari. Selle põhjuseks on geenide vaheliste alade pikkuste varieerumine suuremates vahemikkudes kui geenide pikkuste puhul.



Joonis 12. *Bacillus subtilis*e geenide arv geenide pikkuse lõikes vahemikus 0-2999 aluspaari.



Joonis 13. *Streptococcus mutans*i geenide vaheliste alade pikkuste jaotus vahemikus 0-499 aluspaari

6.2 PCR praimerite ennustamine

Käesolevas töös kasutati 211 bakteri genoomile kuuluvat 350 DNA järjestust. Genoomide kogupikkus oli 675 990 189 aluspaari ja geene, millele praimereid ennustati, oli 622 684. Sellest 620 844 geenile suudeti disainida praimeripaarid. Järelikult 99,7% geenidele leiti ja 1840 geenile, mis on 0,3%, ei leitud praimerid. Kui vaadata tulemusi genoomide lõikes, selgus, et 171 genoomi kõikidele geenidele leiti praimeripaarid, mis on 81% töös kasutatud genoomidest. Keskmiselt 8,7 geenile igas genoomis ei suudetud antud tingimustel praimereid teha.

6.3 Ebaõnnestumiste analüüs

PCR praimerite ennustamine ebaõnnestus 40’s genoomis, milles ei suudetud kõikidele geenidele praimeripaare leida. Peamiseks põhjuseks olid ebatavalised sümbolid (N, R, M, S, W, K, Y, B, V, D, H) genoomide nukleotiidsetes järjestustes, mida PRIMER3 ei suutnud ära tunda, kui nad sattusid programmi sisendjärjestusse. Praimereid ei disainitud juhul, kui ebatavalised sümbolid asusid nii praimeri alas kui ka produkti piirkonnas (välja arvatud ’N’ produktis). Ebatavalisi sümboleid esines 20’s genoomis. Neist enamus asus *Escherichia coli* genoomi O157H7_EDL933 tüves. Selles genoomis jäi ka kõige rohkem geene (1125) ilma praimeriteta.

Suurt rolli ennustamise ebaõnnestumises mängisid ka praimerite parameetrid. Nendel põhjustel jäid praimeripaarid leidmata 48 geeni jaoks 20’s genoomis. Peamiselt jäid rangeks

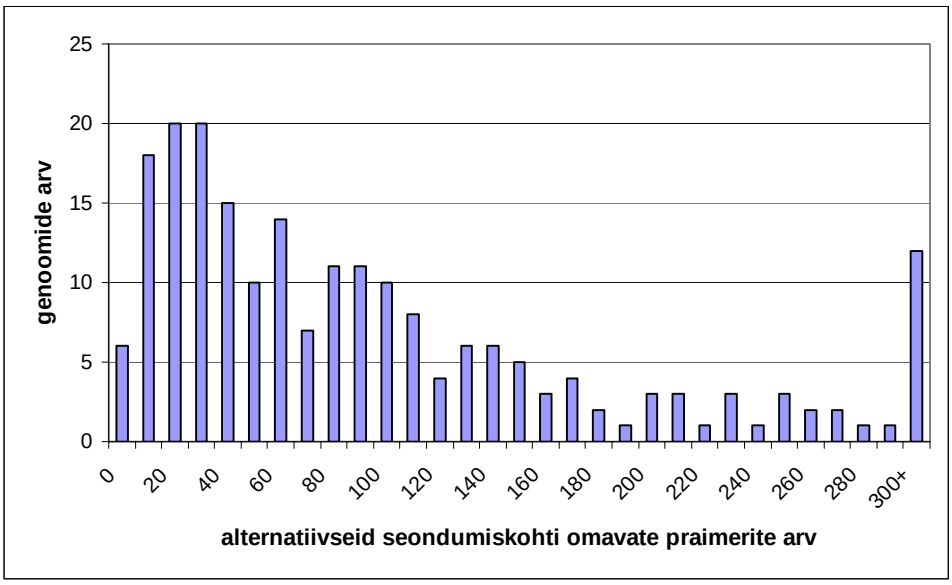
GC sisalduse ja sulamistemperatuuri kriteeriumid olles kas suuremad või väiksemad lubatud väärtusest. Kõrged sulamistemperatuuri erinevused kahe praimeri (+ ja – ahela) vahe tekitasid samuti probleeme. Olulised olid ka lubatud praimerite mononukleotiidide arv, komplementaarsus iseenda ja vastaspraimeriga ning produkti pikkus. Praimerite parameetritest tulenevalt jäi kõige rohkem praimeripaare (22) leidmata *Wigglesworthia brevipalpis* genoomis. Selles genoomis on ka keskmine GC sisaldus väga madal 22,48%.

6.4 Õnnestunud praimerite analüüs

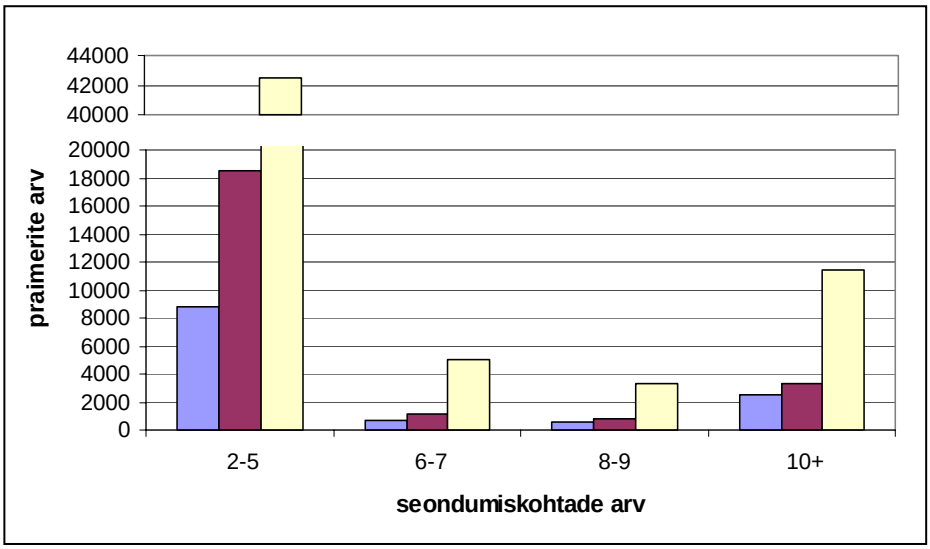
6.4.1 Alternatiivsed seondumiskohad genoomis

PCR edukust mõjutavad praimerite seostumine alternatiivsetesse kohtadesse. Eelnevalt disainitud praimerite juures uurisime nii praimerite seondumiskohtade arvu kui ka alternatiivseid seondumiskohti omavate praimerite arvu genoomis. Seostumise analüüs näitas, et enamus praimeritel on ainult üks seondumiskoht genoomis. Mitut seostumiskohta omavate praimerite jaotus on toodud joonisel 15. Selgus, et suuremal osal nendest on 2 – 5 kohta genoomis, kuhu võiks seostuda. Praimereid, mille seondumiskohtade arv >10, oli keskmiselt 23 igas genoomis, täpse seondumise korral. Kahe valepaardumise korral oli see suurus 106.

Alternatiivseid seondumiskohti omavate praimerite arv genoomi kohta varieerus 1 kuni 666 (*Escherichia coli* O157H7 EDL933). Kõige rohkem alternatiivsete seostumiskohtadega praimereid oli *Shigella flexneri* 2a genoomis. Samas selgus, et kuue genoomi kõikidel praimeritel puudusid alternatiivsed seostumiskohad (joonis 14).



Joonis 14. Alternatiivseid seostumiskohti omavate praimerite jaotus genoomide lõikes täpse seondumise korral.

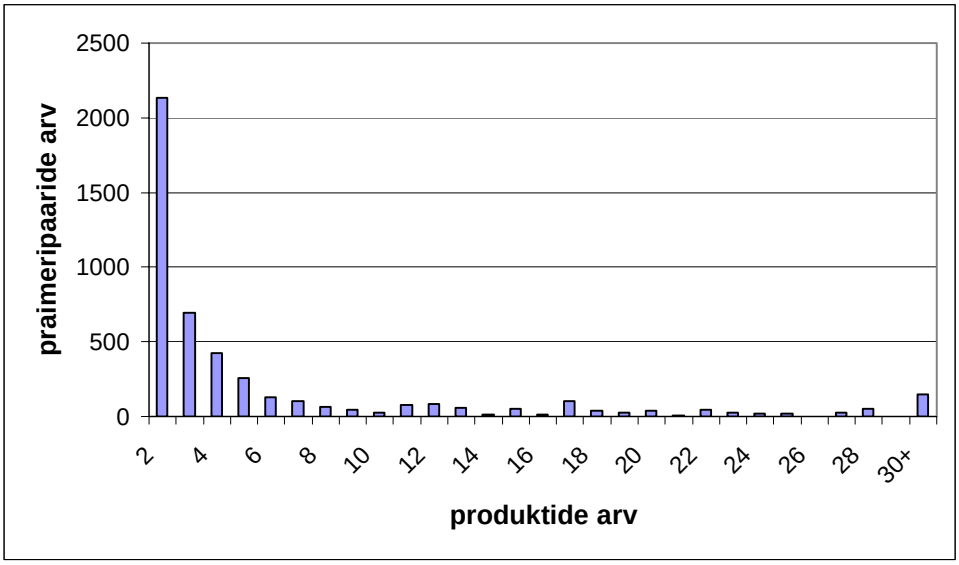


Joonis 15. Praimerite seondumiskohtade arvu jaotus. Suuremal osal praimeitel on 1 seondumiskoht genoomis (andmeid pole näidatud) Joonise koostamisel on kasutatud 108 genoomi andmeid. Sinine = täpne seondumine, punane = lubatud 1 valepaardumine, kollane = lubatud 2 valepaardumist.

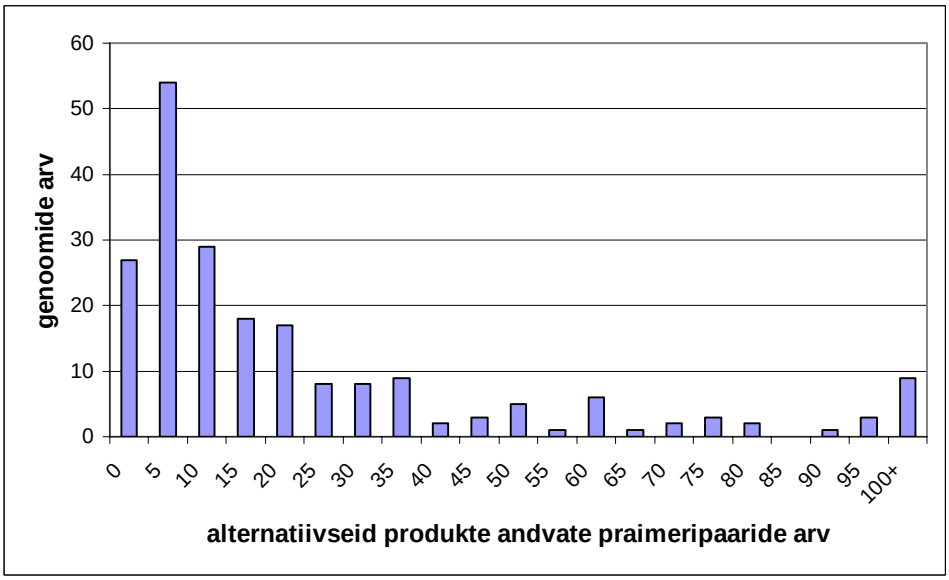
6.4.2 Alternatiivsete produktide teke

Alternatiivsete seostumiste tõttu võivad soovitud produktide kõrvale tekkida ka alternatiivsed produktid. Praimeripaaride arvu ja produktide arvu võrdlus näitab, et enamik praimeripaare annavad 1 produkti. Praimeripaare, mis andsid alternatiivseid produkte oli kokku 181 ja alternatiivseid produkte tekkis kokku 4704. Kõige rohkem esines praimeripaare (2132), millel tekkis 1 alternatiivne produkt (joonis 16). Suurim produktide arv (580) ühe

praimeripaari kohta oli *Leptospira interrogans* serovar Lai genoomis Joonisel 17 on toodud alternatiivseid produkte andvate praimeripaaride arvu jaotus genoomide arvu lõikes. Selgus, et uuritud tingimustel puudusid alternatiivsed produktid 27 genoomi praimeritel. Kõige rohkem produkte genoomi kohta leidus *Shigella flexneri* tüvedes 2a ja 2a_2457T, vastavalt 237 ja 208 võimalikku lisaprodukti. Enamus genoome sisaldas praimeripaare, mis andsid 1 – 5 alternatiivset produkti.



Joonis 16.Võimalike produktide arvu võrdlus praimeripaaride arvuga.



Joonis 17. Alternatiivseid produkte andvate praimeripaaride jaotus genoomide lõikes täpse seondumise korral.

Arutelu

Kasutades eelnevalt kirjeldatud meetodit leidsime 99,7% bakterite geenidele praimeripaarid. Genoomide analüüs näitas, et kõikidele geenidele suudeti praimeid ennustada (171 genoomis). Seitsme enamesindatud (bakteri liikidega paremini kaetud) hõimkonna osas oli õnnestunud bakterite jaotuvus hõimkondade lõikes ühtlane. Keskmine GC sisaldus nendes genoomides oli vahemikus 23,9% - 69,3%. Praimereid õnnestus edukalt disainida nii suurematele kui ka väiksematele genoomidele.

Disainitud praimerite kvaliteedi hindamine.

Praimerite kvaliteeti mõjutavad genoomide kordused. Kui eukarüootides on väga palju lühikesi kordusi, siis bakterite genoom sisaldab oluliselt vähem korduvaid järjestusi. Kordusjärjestustuste mõju leitud praimeripaaride kvaliteedile hindasime paimerite seondumiskohtade arvu ja alternatiivsete produktide tekkimise võimalusega. Kuna korduvale DNA'le disainitud praimerid võivad seonduda alternatiivsetesse kohtadesse ja tekitada alternatiivseid produkte. Kuna bakterites on suhteliselt vähe kordusjärjestusi, siis alternatiivsetesse kohtadesse seonduvate praimeite arv oli ka väike. Disainitud 620 844 praimeripaarist 29608 praimerit omasid sekundaarseid seondumiskohti ja 0,03% praimeripaaridest võimaldasid alternatiivsete produktide teket täpse seostumise korral. Lubades praimerite seondumisel DNA'le 1 - 2 valepaardumist kasvab seondumikohtades arv keskmiselt 2 – 4,6 korda. Mõne genoomis on korduste arv tunduvalt suurem kui teistes. Seega on praimeritel nendes genoomides rohkem seondumiskohti. Suurema korduste arvuga genoomides oleks vajalik kasutada kordusjärjestuste maskeerimist, mis vähendaks praimerite seondumiskohtade arvu.

Analüüsides praimerite seondumisi oli näha seoseid genoomi suuruse ja praimerite seostumiskohtade ning praimerite seostumiskohtade ja tekkivate lisaproduktide vahel. Mida suurem on genoom, seda rohkem on praimeritel seondumiskohti ja mida rohkem on seostumiskohti, seda enam võib tekkida ka lisaprodukte. Samas on suuremates genoomides üldiselt rohkem kordusi. Seoseid GC sisalduse ja alternatiivsete seondumiskohtade vahel ei leidnud.

Praimeripaaride unikaalsuse tagamiseks on mitu võimalust. Esiteks võib kasutada praimerite alternatiivsete seondumiskohtade otsimist ning need maskeerida. Teiseks võib korduvad alad enne praimerite disaini maskeerida. Maskeerimine on suhteliselt kiire tegevus. Kuid niiviisi maskeerides võivad kaduda ka "head" praimerite seondumiskohad (Haas *et al*,. 1998). Samas ei välista maskeerimine 100% alternatiivsete produktide tekkimist. Selles töös

on seondumiskohtade leidmiseks kasutatud FASTAGREP'i. FASTAGREP on suhteliselt aeganõudev, eriti kui otsida lubada ebatäpseid paardumisi.

Praimerite disaini ebaõnnestumise analüüs.

Praimerite disaini ebaõnnestumiste analüüs näitas, et praimereid ei suudetud leida 0,3% geenidele 40's genoomis. Ligikaudu ¼ analüüsitud genoome, millele kõiki praimeripaare ei suudetud disainida, kuuluvad *γ-Proteobacteria* hõimkonda. Sellesse fülooloogilisse gruppi kuulub ka *Escherichia coli* tüvi O157H7_EDL933, mille 21% geenidele ei õnnestunud praimereid ennustada. Samuti oli selles genoomis disainitud praimeritele palju alternatiivseid seondumiskohti, mis võib olla tingitud sellest, et genoom sisaldab sadu liini-spetsiifilisi DNA segmente (saarekesi), mis on jaotunud üle kogu genoomi. Teistest rohkem praimerid ei suudetud disainida *Mycobacterium tuberculosis* CDC1551, *Haemophilus influenzae*, *Borrelia burgdorferi*, *Bacillus cereus* ATCC 10987, *Wigglesworthia brevipalpis*, *Treponema pallidum*, *Chlamydophila pneumoniae* AR39, genoomides. Sõltuvalt oganismist jäi 1,8% - 5,8% geene praimeiteta.

Kõikidele geenidele ei õnnestunud praimeripaare leida kõige rohkem 4 - 6Mb suurusega genoomides ning genoomide keskmine GC sisaldus varieerus 22,2% - 71,1% vahel. Kuid keskmine GC sisaldus ei määra veel ära genoomi paimeri otsimisala GC sisaldust. Näiteks *Streptomyces coelicolor* genoomis, mille keskmine GC sisaldus on 71,1%, jäi ainult 1 geen praimeriteta.

Lähemal analüüsil selgus, et peamiseks praimerite mitteleidmise põhjuseks oli genoomses alas esinevad ebatavalised sümbolid, mis kuuluvad küll IUB/IUPAC koodi, kuid ei ole A, T, G, C. Lahenduseks oleks lisa skripti kirjutamine, mis asendaks ebatavalised sümbolid A, T, G või C tähega ning jätaks meelde asenduste asukohad, et nendele aladele ei oleks võimalik praimereid disainida. Teine lahendus on kasutada PRIMER3 võimalust muuta ebatavalised sümbolid 'N' tähtedeks. Samas võib praimerite valimisel lubada ka 1 – 2 'N' praimerite seondumiskohas, kuna 1 - 2 valepaardumisega seondub praimer samuti DNA'le.

PCR praimerite automaatiseeritud ennustamisel pole kõikidele geenidele võimalik praimeripaare leida. Põhjuseks on bakterite nukleotiidsete järjestuste küllaltki suur varieeruvus. Kuid kasutades praimerite ebaõnnestumiste põhjuste ja alternatiivsetest seondumiskohtade analüüsist saadud andmeid ning neid arvesse võttes on võimalik parandada paimerite automaatse ennustuse edukust tunduvalt.

Kokkuvõte

Käesolevas töös anti ülevaade bakterite üldistest omadustest, sekveneerimise hetke olukorrast ja genoomide andmebaasidest ning PCR disainist. PCR praimerite disainimiseks, PCR praimerite disaini edukuse hindamiseks ja ebaõnnestunud praimeridisainide analüüsiks koostati kolm perli programmi.

Praimereid ennustati 211 bakterigenoomile, millest 171 genoomile õnnestus praimereid disainida kõikidele geenidele. Ülejäänud 40 genoomi puhul ei suudetud praimereid leida 1840 geenile.

Praimerite ennusatmise ebaõnnestumise analüüsil selgus, et peamiseks praimerite mitteleidmise põhjuseks oli genoomsetes järjestustes asuvad ebatavalised sümbolid (IUB/IUPAC välja arvatud ATGC). Eelnevat arvesse võttes õnnestub untsu läinud PCR praimerite koguhulka vähendada 97%.

Kvaliteetseid praimereid, mis ei omanud mitut seostumiskohta genoomse DNA peal, oli vähemalt 88,6% . Sõltuvalt valepaardumise lubamisest seostusid 2,4 – 11,4% praimeritest alternatiivsesse kohta. Praimeripaare, mis andsid alternatiivseid produkte, oli kokku 181, täpse seondumise korral.

Automaatse praimeridisainis katmata geenide arvu saab lihtsate vahenditega vähendada 97%'ni ja alternatiivsete seondumiste ja produkti arvestamine aitab suurendada disainitud praimerite edukust PCR's.

Summary

Automated PCR primer design for bacterial genes – genome based approach

Helle Uibokand

In this study we described general properties of bacteria, current status of sequencing and databases of genome and PCR design. Three programs has been made; one for design PCR primers, second to assess success of PCR primer design and third to analyse primer quality in prokaryotic genomes.

PCR primers were successfully designed for 99,7% of genes from 211 bacterial genomes. For 171 genomes we design primers for all genes. And for 40 genomes we didn't find PCR primers for 1840 genes.

Analysis of primer design fault reasons show, that main reason of faults was IUB/IUPAC codes for ambiguous bases. In consideration of foregoing we can eliminate 97% of faults in primer design.

Primers with a high quality was at least 88,6%. 2,4 – 11,4% of primer can bind to secondary binding site when be strict or allow 2 mismatch respectively. In case of exact match we found 181 primer pairs what give an alternative product.

Fault rate of automated PCR primer design can be decreased up to 97%. In consideration of alternative binding sites and alternative product help to rise PCR primer quality and PCR success.

Kasutatud kirjandus

Achaz G., Coissac E., Netter P., Rocha E.P.C. Association Between Inverted Repeats and the Structural Evolution of Bacterial Genomes. Genetics 2003, 164(4):1279-1289.

Achaz G., Rocha E.P.C, Netter P., Coissac E. Origin and fate of repeats in bacteria. Nucleic Acids Res 2002, 30(13):2987-2994.

Bansal A.K., Meyer T.E. Evolutionary Analysis by Whole-Genome Comparisons. Journal of Bacteriologi 2002, 184(8):2260-2273.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. GenBank: update. Nucleic Acids Res 2004, 32: D23-D26.

Bentley S.D., Parkhill J. Comparative Genomic Structure of Prokaryotes. Annu.Rev.Genet. 2004, 38:771-791.

Bermingham N., Luettich K. Polymerase chain reaction and its applications. Current Diagnostic Pathology 2003, 9:159-164.

Breslauer J.K., Frank R., Blocker H., Markey L.A. Predicting DNA duplex stability from the base sequence. Proc.Natl.Acad.Sci. USA 1986, 83:3746-3750.

Celestino P.B.S., de Carvalho L.R., de Freitas L.M., Dorella F.A., Martins N.F., Pachcco L.G.C., Miyoshi A., Azevedo V. Update of microbial genome programs for bacteria and archaea. Genetics and Molecular Research 2004, 3(3):421-431.

Coenye T., Vandamme P. Simple sequence and compositional bias in the bipartite Ralstonia solanacearum GMI1000 genome. BMC Genomics 2003, 4:10.

Cole S.T., Saint- Girons S. Bacterial genomes – all shapes and sizes. In: Organisation of the prokaryotic genome (Edited by: Charlebois R.L.) Washington DC, American Society for Microbiology 1999, 35-62.

Ermolaeva M.D., White O., Salzberg S.L. Prediction of operons in microbial genomes. Nucleic Acids Res 2001, 29(5):1216-1221.

Fraser C.M., Eisen J.A., Salzberg S.L. Microbial genome sequencing. Nature 2000, 406(17):799-803.

Garcia-Vallvé S., Romen A., Palan J. Horizontal Gene Transfer in Bacterial and Archaeal Complete Genomes. Genome Res 2000, 10:1719-1729.

Haas S., Vingron M., Poustka A., Wiemann S. Primer design for large scale sequencing. Nucleic Acids Res 1998, 26(12):3006-3012.

Heidelberg J.F., Eisen J. A., Nelson W.C., Clayton R.A., Gwinn M.L. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature 2000, 406:477-483

Jensen L.J., Friis C., Ussery D.W. Three views of microbial genomes. Res Microbiol 1999, 150:773-777.

Jordan J.K., Makarova K.S., Spouge J.L., Wolf Y.I., Koonin E.V. Lineage-Specific Gene Escpansions in Bacterial and Archaeal Genomes. Genome Res 2001, 11:555-565.

Kemmer D., Fraser A. Whose genome is next? Genome Biol 2002, 3:4037.1-4037.3.

Kim Y., Flynn T.R., Bonoff R.B, Wong D.T.W, Todd R. The Gene:The Polymerase Chain Reaction and Its Clinical Application. J Oral Mascillotac Surg 2002, 60:808-815.

Kondrashov F.A., Rogozin J.B., Wolf Y.I., Koonin E.V. Selection in the evolution of gene duplications. Genome Biology 2002, 3(2):research 0008.1-0008.9.

Konstantinidis K.T., Tiedje J.M. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 2004, 101(9):3160-3165.

Kunin V., Ouzounis C.A The Balance of Driving Forces During Genone Evolution in Prokaryotes. Genome Res 2003, 13:1589-1594.

Kämpke T., Kieninger M., Mecklenburg M. Efficient primer design algorithms. Bioinformatics 2001, 17(3):214-225.

Li P., Kupfer K.C., Davies C.J., Burbee D., Evans G.A., Garner H.R. PRIMO: A Primer Design Program That Applies Base Quality Statistics for Automated Large-Scale DNA Sequencing. Genomics 1997, 40(3):476-85.

Liò P. Investigating the Relationship Between Genome Structure, Composition, and Ecology in Prokaryotes. Mol. Biol. Evol. 2002, 19(6):789-800.

Lobry J.R. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 1996, 13:660-665.

McLean M.J., Wolfe K.H., Devine K.M. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 1998, 47:691-696.

Mullis K.B. The unusual origin of the polymerase chain reaction. Sci Am 1990, 4:56-65.

Nakabachi A., Yamashita A., Toh., Ishikawa H., Dunbar H.E., Moran N.A., Hattori M. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science 2006, 314(5797):267.

Nelson K.E. The future of microbial genomes. Environ Microbiol 2003, 5:1223-1225.

Podowski R.M., Sonhammer E.L.L. MEDUSA – large scale automatic selection and visual assessment of PCR primer pairs. Bioinformatics 2001, 17(7):656.

Rocha E.P.C. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends Microbiol 2002, 10:393-396.

Rocha E.P.C., Danchin A. Essentiality, not expressiveness, drives gene strand bias in bacteria. Nat Genet 2003, 34:377-378.

Rogozin I.B., Makarova K.S., Natale D.A., Spiridonov A.N., Tatusov R.L., Wolf Y.I., Yin J., Koonin E.V. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res 2002, 30(19):4264-4271.

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000, 132:365-386.

Roy R. Chaudhuri, Arshad M. Khan, and Mark J. Pallen *coli*BASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. Nucleic Acids Res. 2004, 32: D296-D299.

Rychlik W., Rhoads R.E. A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res. 1989, 17(21):8543-51.

Salzberg S.L., Delcher A.L., Kasif S., White O. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 1998, 26(2):544-548.

Suggs S. V., Hireose T., Kawashima M. J. Using Purified Genes. Symposium on Developmental Biology 1981, 23.

Ussey D.W., Binnewies T.T., Gouveia- Oliveira R., Jarmer H., Hallin P.F. Genome update: DNA repeatsin bacterial genome. Microbiology 2004, 150:3519-3521.

Ussery D.W., Hallin P.F. 2004. Genome update: AT content in sequenced prokaryotic genomes. Microbiology 2004, 150:749-752.

van Baren M.J., Heutink P. The PCR Suite. Bioinformatics 2004, 20(4):591-593.

Varotto C., Richly E., Salamini F., Leister D. GST-PRIME: a genome wide primer design software for the generation of gene sequence tags. Nucleic Acids Res 2001, 29(21):4373-4377.

Weinel C., Tümmler B., Hilbert H., Nelson K.E., Kiewitz C. General method of rapid Smith/Birnstiel mapping adds for gap closure in shotgun microbial genome sequencing projects: application to Pseudomonas putida KT2440. Nucleic Acids Res 2001, 29(22):e110.

Wetmur J.G. DNA probes: applications of the principles of nucleic acid hybridization. Crit Rev Biochem Mol Biol 1991, 26:227-259.