TARTU ÜLIKOOL
BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

TRIINU KÕRESSAAR

Polümeraasi ahelreaktsiooni modelleerimine DNA amplifitseerimiseks bakteriaalsetes genoomides

Magistritöö

Juhendaja Prof Maido Remm, PhD

TARTU
2006

Sisukord

LÜHENDID JA MÕISTED....4
SISSEJUHATUS....6
I KIRJANDUSLIK TAUST....8
1. Bakterite molekulaarne diagnostika....8
1.1 Nukleiinhapetel põhinevad tehnoloogiad....8
1.2 Meetodite eelised ja puudused....10
2. Lühi-ülevaade PCR-i tehnoloogiast....11
2.1 PCR-i sensitiivsus ja spetsiifilisus....12
2.2 PCR-i mõjutavad tegurid....12
3. Termostabiilsed DNA polümeraasid....13
3.1 Polümeraasi tertsiaarne struktuur....14
3.2 Taq polümeraasi töö mehhanism....14
3.3 PCR-i artefaktide teke....16
4. DNA dupleksi struktuur ja seda koos hoidvad jõud....17
4.1 Dupleksi struktuur....18
4.2 Vesiniksidemed....18
4.3 Aluste stacking....18
4.4 Steerilised efektid....19
4.5 Fosfaatide tõukumine....19
5. dsDNA termodünaamika mudel (nearest-neighbor mudel)....20
5.1 Kaheahelalise DNA sulamine....20
5.2 Termodünaamika mudelid....21
5.3 N-N termodünaamika mudel....22
5.4 DNA oligonukleotiidsete dupleksite sulamistemperatuuri arvutamine....23
5.4.1 PCR-i praimeri seondumistemperatuur....24
5.4.2 PCR-i praimeri sulamistemperatuur....25
6. DNA dupleksi sekundaarstruktuurid ja nende termodünaamilised parameetrid....26
7. Soolade mõju DNA dupleksi stabiilsusele, soolakorrektsiooni arvutamine....28
7.1 Erinevad katioonid, mõju dupleksi stabiilsusele....29
7.2 Primaarstruktuurist sõltumatu soolakorrektisoon....29
7.3 Primaarstruktuurist sõltuv soolakontsentratsioon....30
7.4 Bivalentsete katioonide teisendus monovalentseteks....31
II PRAKTILINE OSA....32
TÖÖ ISELOOMUSTUS JA EESMÄRGID....32
ALGANDMED JA METOODIKA....33
1. Andmete päritolu ja struktuur....33
2. Uuritavad bakteritüved....33
3. Liigispetsiifiliste kordusjärjestuste leidmine....34
4. Mitmene joondus ning praimeri disain....35
5. Sõltumatute tunnuste leidmine....36
6. Praimerite valimine eksperimentaalseteks katsetusteks....43

7. Eksperimentaalsete katsete kirjeldus.....44
8. Kasutatud statistilise analüüsi meetod.....47
8.1 GLM mudel.....47
8.2 Kasutatud logistilise regressiooni protseduur statistikapaketis SAS: logistic.....48
8.3 Tunnuste olulisuse määramine, kollineaarsuse probleem. 50
8.4 Andmete teisendamine.....51
8.4 Lõpliku mudeli koostamine.....52
TULEMUSED.....53
1. Eksperimentaalselt uuritavate tunnuste kirjeldus.....53
1.1. Uuritava küsimuse defineerimine.....53
1.2. PCR-i mõjutavad praimerite omaduste uurimine.....53
1.3. Teised PCR-i mõjutavad tunnused.....57
2. Liigispetsiifiliste kordusjärjestuste otsimine.....58
2.1. Liigispetsiifiliste kordusjärjestuste otsimise eesmärk.....58
2.2. Kordusjärjestuste leidmine.....59
3. Praimerite valik eksperimentaalseteks katseteks.....61
3.1. Kandidaatpraimerite disain. 61
3.2. Uuritavate tunnuste väärtuste arvutamine.....63
3.3. Katses kasutatavate praimerite valik. 63
4. PCRi tulemuste intensiivsuste tasemeid mõjutavate tunnuste leidmine.....64
4.1. Gruppide olulisemad esindajad.....64
4.2. Praimerite seostumiskohtade arvu mõju PCRi intensiivsuse tasemetele 66
4.3. Teiste tunnuste mõju PCR-i tulemsusetele.....71
1.5. PCRi praimerite intensiivsuse tasemeid ennustav mudel.....73
ARUTELU.....76
RESÜMEE.....76
RESUME.....76
TÄNUAVALDUSED.....76
BIBLIOGRAAFIA.....77
LISAD.....82

LÜHENDID JA MÕISTED

Aluse *stacking (base stacking)*	-lämmastikaluse interakteerumine järgmise lämmastikalusega piki DNA ahelat
bp *(**b**ase **p**air)*	-aluspaar
cDNA *(**c**omplentary **DNA**)*	-mRNA-lt pöördtranskriptaasi abil sünteesitud mRNA-ga komplementaarne DNA
DNA *(**D**esoxyribo**N**ucleic **A**cid)*	-desoksüribonukleotiidhape
EST *(**E**xpressed **S**equence **T**ag)*	-lühike DNA järjestus, mis on saadud cDNA 3' või 5' otsast
mismatch	-mitte-paardunud nukleotiidid; Watson-Crick paardumist mitte omav nukleotiidide paar
NAAT (Nucleic Acids Amplification Test)	nukleiinhapete paljundamisel põhinev test
NCBI (National Center for Biotechnology Information)	Rahvusvaheline biotehnoloogilise informatsiooni keskus
NN (Nearest-Neighbor model)	- mudel termodünaamiliste parameetrite arvutamiseks
PCR (Polymerase Chain Reaction)	-polümeraasi ahelreaktsioon
Pfu polümeraas	-termostabiilne polümeraas bakterist *Pyrococcus furiosus*
PNA (*Peptide Nucleic Acid*)	kunstlik DNA analoog

RT-PCR (*Reverse Transcription-Polymerase Chain Reaction*)	pöördtranskriptsiooni polümeraasi ahelreaktsioon
ssDNA ***(single-stranded DNA)***	-üheahelaline DNA
stacking	-vt aluste *stacking*
T_m	-sulamistemperatuur
T_a	-seondumistemperatuur
Taq polümeraas	-termostabiilne polümeraas bakterist *Thermus aquaticus*
Tfu polümeraas	-termostabiilne polümeraas bakterist *Thermococcus fumicolans*
WC (**W**atson-**C**rick)	- Watson-Crick vesiniksidemed

SISSEJUHATUS

Spetsiifiliste bakteritüvede identifitseerimine on vajalik meditsiinis (kohtu- ja kliinilisesmeditsiinis), veterinaarias, toiduainetööstuses, keskkonnakaitses jm. Järjest vähem on kasutusel bakterite kultiveerimine ja laialdasemalt bakterite detekteerimine nukleiinhapete amplifitseerimise testidega (NAAT, *nucleic acid amplification test*) ning mikrokiipidega. Kõrvuti levinud nukleiinhapete amplifitseerimise meetodiga PCR leiavad kliinilistes laborites kasutust ka teised NAAT lähenemisel baseeruvad metoodikad - LCR *(Ligase Chain Reaction),* RT-PCR *(Reverse Transcription-Polymerase Chain Reaction),* SDA *(Strand-Displacement Amplification),* TMA *(Transcription-Mediated Amplification)* jt (60). Bakterite tuvastamine kultiveerimise läbi leiab tänapäeval veel erinevates praktilistes eluvaldkondades palju kasutamist (toiduainete tööstus, veterinaaria). Viimase suurteks puudusteks on ebapiisav spetsiifilisus, kuna on erinevaid lähedasi tüvesid pole lihtne (võimalik) silmaga eristada, limiteeritus detekteeritavate bakterite liikide arvu suhtes, kuna vajaliku keskkonnatingimuste (söötmed, temperatuur jne) loomine bakterite kasvatamiseks pole triviaalne ning proovi kiire ning vajalikele tingimustele vastav toimetamine doonorilt laborisse, samuti on osade bakterite kultiveerimine pikaajaline protsess.

Kliinilises diagnostikas on seoses PCR-i põhise patogeenide detekteerimisega seotud peamiselt kaks probleemi - kasutatavate praimerite spetsiifilisus ning sensitiivsus. Kliinilised proovid sisaldavad väga erinevates konsentratsioonides erinevate liikide (tüvede) genoomset DNAd. Et oleks võimalik proovist detekteerida selles ka väga väikeses koguses esinevat DNAd, on oluline, et disainitud praimeripaar oleks sensitiivne - võimaldaks produkti akumuleerumist efektiivselt. Üks bakteri liik võib esineda erinevate alamliikidena (tüvedena), mida omakorda eristatakse teatud iseloomulike antigeenide komplekti alusel serotüüpideks *(serovar, serotype)* või siis suurematekshulkadeks - serogruppideks *(serogroup)*. Kahe serogrupi (ka kahe tüve)

genoomne DNA järjestus erineb primaarstruktuuri tasandil vähe; sellest tulenevalt disainides praimerid ühele grupile on suur tõenäosus, et disainitud praimeripaar on võimeline paljundama ka teiselt tüvelt produkti. Viimane aga ei pruugi olla patogeen. Ilmne on, et kvaliteetsete PCR-i praimerite disainimiseks on vajalik kogu bakteri genoomi primaarjärjestus. Tänaseks on sekveneeritud suhteliselt palju erinevate prokarüootsete liikide genoomne DNA (315 lõpetatud genoomset DNA järjestust *National Center for Biotechnology Information* ehk NCBI andmebaasis; märts 2006). Siiski pole käesoleval hetkel sekveneeritud paljude bakteriaalsete liikide erinevate alamliikide (tüvede), rääkimata erinevatest serotüüpidest, genoomse DNA järjestus.

Selleks, et oleks võimalik paljundada produkti vaid soovitult tüvelt tuleb praimeripaar disainida äärmise tähelepanelikkusega - oluline on, et disainitud praimeripaar annaks produkti ühe bakteri genoomse DNA (patogeenne genoom ehk sitmärk-järjestus) pealt ning ei annaks produkti teise, sarnase primaarjärjestusega genoomse DNA pealt (mitte-patogeen). Oluline on mõista, millised asjaolud on tarvilikud ja samas ka piisavad kvaliteetsete PCR-i praimerite disainis. Kuidas tuvastada patogeeni kliinilisest proovist täpselt ja tundlikult? Käesolevas töös üritame kasutada bakterite liigispetsiifilisi kordusi suurema tundlikkusega PCRi praimerite disainimiseks. Nendele kordustele disainitud praimereid testitakse töö käigus eksperimentaalselt uurimaks, millised PCRi praimeri omadused mõjutavad PCR-i erinevaid intensiivsuse tasemeid.

Käesoleva töö esimeses osas püütakse anda ülevaadet PCR-i praimerite disainimise protsessi taga olevast biokeemiast ning vastavatest mudelitest ning teises osas püütakse neid rakendada bakteriaalsete patogeenide tuvastamiseks PCR-i tehnoloogia abil.

I KIRJANDUSLIK TAUST

1. Bakterite molekulaarne diagnostika

Molekulaarsed meetodid bakterite identifitseerimiseks on viimastel aastatel arenenud jõudsalt (67-69,71). Kuigi patogeensete bakterite identifitseerimiseks kliinilisest proovist kasutatakse ka tänapäeval veel mitte-nukleiinhapetel põhinevaid tehnoloogiaid (kultiveerimine, seroloogilised meetodid), siis enamasti on trend nukeiinhapetel põhinevate meetodite kasutamise suunas. Võrreldes nukleiinhapetel ning mitte-nukleiinhapetel põhinevaid tehnoloogiaid omavahel, on esimesed kiiremad, täpsemad, tundlikumad ja vähem töömahukad. Levinumad meetodid on otseselt nukleiinhapete amplifitseerimisel baseeruvad tehnoloogiad (tavapärane PCR, RT-PCR), signaali amplifitseerimisel põhinevad tehnoloogiad ning nanotehnoloogiad. Mainitud tehnoloogiad eeldavad identifitseeritava organismi genoomse DNA primaarjärjestuse (vähemalt osalist) olemasolu.

1.1 Nukleiinhapetel põhinevad tehnoloogiad

PCR-il otseselt baseeruvad meetodid. Molekulaarses diagnostikas kasutatakse bakterite identifitseerimiseks kliinilistest proovidest laialdaselt erinevaid konventsionaalse PCR-i modifitseeritud variante ning omakorda modifitseeritud variantide modifikatsioone. *Hot-start* PCR (61) ja *touch-down* PCR (62) on välja töötatud selleks, et leevendada probleemi PCRi ebaspetsiifikaga. Esimeses denatureeritakse kaheahelaline proovis olev DNA enne amplifikatsiooni saavutatds seeläbi kõrgema produkti akumuleerumise taseme. Teise korral kasutatakse ära PCR eksponentsiaalset produkti kogunemise iseloomu ning langetatakse iga PCR-i tsükli korral seondumistemperatuuri teatud ^{o}C võrra.

PCR-il baseeruvate tehnoloogiate erinevust võib vaadata ka kui PCR-i erinevalt disainitud praimeritega (Sunrise®, Amplifluor®, *Scorpions, 'light upon extension'* (LUX) *fluorogenic, DzyNA* praimerid). Reaalaja PCR *(real-time* PCR e Q-PCR*)* võimaldab hinnata alates

reaktsiooni algusest sihtmärk-järjestuse ehk amplifitseeritava produkti kogust. Reaalaja RT-PCR (*real-time reverse transcriptase* PCR e RRT-PCR e QRT-PCR) võimaldab reaalaja PCR-i rakendamist kasutades algmaterjalina RNA-d (transkribeeritakse DNA-ks). Erinevates reaalaja PCR-i tehnoloogiates kasutatakse erilisi fluorestsents märgisega praimereid, mis olles DNA sünteesimisel praimeriks, kiirgavad valgust (67).

PCR-il mitte-baseeruvad meetodid. Sellised on signaali amplifitseerimisel ja proovi amplifitseerimisel põhinevad meetodid, isotermilised ja termilised meetodid. Pürosekveneerimine *(pyrosequencing)* on reaalaja bioluminomeetriline meetod, mis mõõdab pürofosfaadi (DNA polümeriseerumise kõrvalprodukt) tekkimist ehk DNA polümeriseerumise tulemusena eralduva valguse intensiivsust (67). Mikrokiipide tehnoloogia on tänapäeval üks laialdasemalt levinud rakendus erinevates valdkondades. Meetod on termiline ja põhineb signaali amplifitseerimisel ning sellest on arendatud erinevaid bioloogilisi tööriistu. Näiteks PNA mikrokiibid on mikrokiibid, kus proovina kasutatakse sünteetilist DNA-d, so peptiid-nukleiinhappeid (PNA, *peptide nucleic acid*). Võrreldes DNA-ga on PNA-s negatiivne fosfaat-selgroog asendatud laenguta glütsiinist koosneva selgrooga. Mitte-laetud selgroo tõttu on PNA-DNA dupleksid termiliselt stabiilsemad ning pole niivõrd sõltuvad soola kontsentratsioonist ja pH tasemest ja on resistentsemad nukleaasidele ja proteaasidele võimaldades seega laiemas katsetingimuste vahemikus uuritavat järjestust detekteerida (66). Signaali amplifitseerimisel põhineva meetodi näiteks võib tuua ka *branched-DNA* (*b*DNA), kus spetsiifilise sihtmärk-nukleiinhappega seonduvad aheldunud (nn puukujulised) ja märgistatud (proovide harud) proovid tekitavad signaali läbi alkaliin-fosfataasi (71). NASBA *(Nucleic Acid Sequence-Based Amplification)* on isotermiline nukleiinhappe järjestuse paljundamisel põhinev meetod, kus paljundatakse RNA järjestus, kas DNA või RNA järjestuselt kasutades selleks kolme erinevat ensüümi (pöördtranskriptaas, RNaas H, T7 RNA polümeraas) (82). HDA (*Helicase Dependent Amplification*) on isotermiline *in vitro* DNA amplifitseerimine, kus kasutatakse kahe-ahelalise DNA denatureerimiseks helikaase ning DNA paljundamiseks DNA polümeraasi (65,72).

Üks uuemaid ja revolutsioonilisemaid tehnoloogiaid on nanotehnoloogia, mis põhineb erinevatel nanoobjektidel - nanopartiklid, nanopoorid, nanotuubid, nanokanalid, nanomahutid (69). Näiteks BCA *(bio-barcode assay)* - valgu või nukleiinhappe molekulid on kinnitatud magnetiliste kullast nanopartiklite külge. Tehnoloogia võimaldab detekteerida nii valku kui nukleiinhapet. Kasutades prooviks nukleiinhappeid seotakse nanopartikli külge tuhandeid spetsiifilisi oligonukleotiide, detekteeritav järjestus 'haaratakse' kahe nanopartikli poolt ning seejärel partikkel-nukleiinhape-partikkel eemaldatakse magnetiliselt. Kuna partikkel on kaetud tuhandete spetsiifiliste oligonukleotiididega, siis on automaatselt signaali amplifitseerimine tagatud vähemalt tuhande kordselt. Tehnoloogia on äärmiselt sensitiivne - detekteerimise sensitiivsuse limiit on ligikaudu 500 zeptomolaarne (10^{-21} molaarne) sihtmärk-DNA kontsentratsioon (68).

1.2 Meetodite eelised ja puudused

Erinevates uuringutes vajatakse erinevaid molekulaarseid tööriistu, seetõttu vaadatakse konkreetse kasutatava tehnoloogia valikul erinevaid omadusi: kiirus, läbilaskevõime, täpsus, tundlikkus *(detection limit),* kvantifitseerimise võimalus, hind, portatiivsus (ehk tehnoloogiat kandva aparaadi, mida on võimalik raskusteta ja kiiresti ühest kohast teise viia, olemasolu (70)) ning valitakse sobivaim.

PCR-i peamised eelised teiste nukleiinhapetel baseeruvate tehnoloogiate ees on meetodi kiirus, lihtsus, väike töömahukus, suhteliselt madal hind. Reaalaja PCR-i eeliseks on produkti tekkimise jälgimine reaalajas ning kiirus (võrreldes tavapärase PCR-iga puudub eraldi seisev produkti detekteerimise etapp). Bakterite detekteerimine mikrokiipide abil on suure jõudlusega, kuid saavutamaks usaldusväärseid ja korratavaid tulemusi, nõuab paljude faktorite optimiseerimist, nt algmaterjali amplifitseerimine (peamiselt PCR-i abil), proovi spetsiifilisus, tulemuste tõlgendamise metoodika (64). PNA mikrokiipide tehnoloogia puuduseks on eelkõige molekulide kõrge hind; mainida võib ka vastavate dupleksite termodünaamilise stabiilsuse (sulamistemperatuuri) ennustamiseks vajaliku arvutusmetoodika puudulikkust, kuigi

viimane tuleneb peamiselt tehnoloogia vähesest rakendamisest praktikas, mis omakorda tuleneb tehnoloogia kallidusest (66).

Üldiselt keerulise aparatuuri ning spetsiifiliste molekulide kõrge hind on peamiseks takistuseks uute tehnoloogiate kasutusele võtmises. Kliinilise diagnostika ja üldiselt biotehnoloogia tulevikku vaadates on rõhk uute tehnoloogiate arendamisel ja vastavate tehnoloogiate laborite poolt kasutusse valimisel kiirusel ja täpsusel (läbilaskevõimel) minimaalse proovi koguse ja testi/eksperimendi hinna juures.

2. Lühi-ülevaade PCR-i tehnoloogiast

Polümeraasi ahelreaktsioon (PCR) võimaldab paljundada DNA järjestuselt spetsiifilist regiooni kasutades kahte vastasahelatele seonduvat oligonukleotiidi (praimerit). Reaktsioon põhineb DNA ahela tsüklilisel sünteesil, mis kujutab endast temperatuuri toimel DNA ahelate denatureerimist, praimerite seondumist komplementaarsetele ahelatele madaldatud temperatuuril ja praimeri ekstensiooni kasutatava termostabiilse DNA polümeraasi kõige efektiivsemal töötemperatuuril (Taq polümeraasi korral 72 ^{o}C). Tekkinud produkti detekteerimine toimub (pikkuse järgi) agaroos-geelelektroforeesil (11).

Teoreetiliselt - produkti kogus kahekordistub iga tsükli sammu jooksul ehk valemina väljendatult $N=N_0 2^n$, kus N on amplifitseeritud produktide arv, N_0 on algne paljundatavate järjestuste arv, n on PCR-i reaktsioonis kasutatavate tsüklite arv. Praktikas on amplifitseerimise efektiivsus (E) väiksem kui 100% kõikudes vahemikus [0;100[%ni (ehk 0-st 1ni) ehk $N=N\left(1+E\right)^n$ (12).

Kuigi tavakohasest PCR-ist on välja arendatud erinevaid PCR-il baseeruvaid meetodeid (*Real Time*-PCR (1), RAPD (8) jt) on esimene jätkuvalt laialdaselt kasutusel. Konventsionaalne PCR (*conventional* PCR, edasi PCR) leiab tänapäeval kasutust väga erinevates eluvaldkondades –

näiteks põllumajanduses geneetiliselt modifitseeritud organismide tuvastamiseks (2), kohtuanalüüsis keelatud organismide (3) ja ainete (4) tuvastamiseks ning isikute identifitseerimiseks (nt massihauad (5)), kliinilistes uuringutes (6,7). Põhjused, miks PCR-i laialdaselt teiste (tundlikumate ja informatiivsemate, nt RT-PCR; suurema jõudlusega, nt mikrokiibid) molekulaarsete tehnoloogiate kõrval kasutatakse on meetodi odavus, universaalsus, kiirus ning lihtne tööpõhimõte.

2.1 PCR-i sensitiivsus ja spetsiifilisus

PCR-i tulemuste omadusi iseloomustavad kaks asjaolu – oodatava produkti detekteeritavuse määr *(amplification efficiency)* ning usaldusväärsus, et detekteeritud produkt on oodatav. PCR-i tulemused sõltuvad paljudest asjaoludest. Olulisemad nende seas on PCR-i praimerite spetsiifilisus ning sensitiivsus.

PCR-i praimeripaari sensitiivsus näitab, kui efektiivselt on antud praimerid võimelised oodatud produkti paljundama (millisel määral on oodatatud produkt (silmaga) detekteeritav). Bakterite molekulaardiagnostikas väljendab praimeripaari tundlikkus praimerite võimet antud (varieeruva) kliinilise proovi konsentratsiooni juures uuritavat bakteritüve visuaalselt detekteerida.

Praimeripaari spetsiifilisus väljendab praimeripaari võimet paljundada sihtmärkjärjestuselt ainult oodatud produkti (ehk võimet vältida produkti paljundamist valest kohast sihtmärk-järjestuselt). Molekulaardiagnostikas peegeldab praimerite täpsus seda, et antud kliinilisest proovist detekteeritakse (sihtmärk-genoomi olemasolu korral proovis) seda ja ainult seda bakteritüve, millele konkreetne praimeripaar on disainitud või ei detekteerita ühtegi produkti (sihtmärk-genoomi puudumise tõttu).

2.2 PCR-i mõjutavad tegurid

Praimeri(paari) spetsiifilisus ja sensitiivsus on sõltuvad rohketest üksteist teatud määral

katvatest ja üksteisest sõltuvatest teguritest. PCR-i mõjutavad tegurid on laias mõttes jaotatavad järgmisteks gruppideks: praimerite erinevad interaktsioonid, produktide interaktsioonid, praimerite seondumiste arvud mittesihtmärk regioonidesse, kasutatavad katsetingimused.

Lisaks tuleb keeruliste (eukarüootide genoomid) sihtmärk-järjestuste korral eraldi arvestada seal leiduvate kordusjärjestustega. Kordusjärjestustele seonduvad praimerid võivad genereerida alternatiivseid produkte ning madalal tasemel või mitte-oodatud pikkusega (või üldse mitte) detekteeritavat produkti. Seepärast on tarvilik enne PCR-i praimerite disainimise etappi sihtmärkjärjestus maskeerida kordusjärjestusi sisaldavatelt kohtadelt (59, 10).

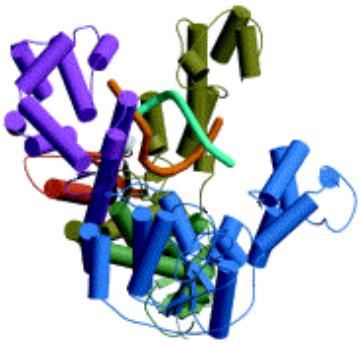
3. Termostabiilsed DNA polümeraasid

Kvaliteetsete PCR-i praimerite modelleerimiseks, on kasulik mõista PCR-i reaktsioonis kasutatavate DNA polümeraaside tööpõhimõtet. Tagamaks PCR-i reaktsiooni piisavat sensitiivsust, on tarvilik, et kasutatav polümeraas ei seonduks vale(de)sse regiooni(desse) kinnitunud praimeri 3' otsa külge. Reaktsiooni spetsiifilisuse määrab täiendavalt eelmisele asjaolu, kas mittesihtmärk-järjestus/praimer dupleksiga seondunud polümeraas on võimaline praimeri 3' otsa uusi nukleotiide liitma.

Alates 1980ndast aastast, mil isoleeriti ja kirjeldati esimene termostabiilne DNA polümeraas (*Taq* polümeraas *Thermus aquaticusest*), on kasutatavate termostabiilsete DNA polümeraaside arv kasvanud kiiresti (28). Erinevad termostabiilsed polümeraasid omavad erinevaid biokeemilisi aktiivsusi (3'->5' ekso- ja 5'->3' nukleaasne, ahela asendamise aktiivsus jt), erinevat reaktiivsuse ning täpsuse taset (14). Peale laialdaselt levinud *Taq* DNA polümeraasi kasutatakse ka *Pyrococcus furiosus* (*Pfu* polümeraas), *Thermococcus fumicolans* (*Tfu* polümeraas) jt (15, 20, 28, 83).

3.1 Polümeraasi tertsiaarne struktuur

Polümeraasi struktuuri mõistmine on vajalik selleks, et aru saada, kuidas lisatakse ahelasse uus nukleotiid ja toimub uue ahela pikendamine ning, mis tingimustes võib toimuda mitte Watson-Crick geomeetriat järgiva nukleotiidi lisamine sünteesitavasse ahelasse. Polümeraasi ja uue ahela sünteesimist vaadatakse läbi parema käe konfiguratsiooni (joonis 1) – sõrmed (*fingers,* va pöial) interakteeruvad sissetuleva nukleotiidiga ja sihtmärk-järjestusega, peopesas *(palm)* on polümeraasi aktiivsait, mis seondub sissetuleva nukleotiidiga ja pöial *(thumb),* mis seondub kaheahelalise DNA-ga (16).

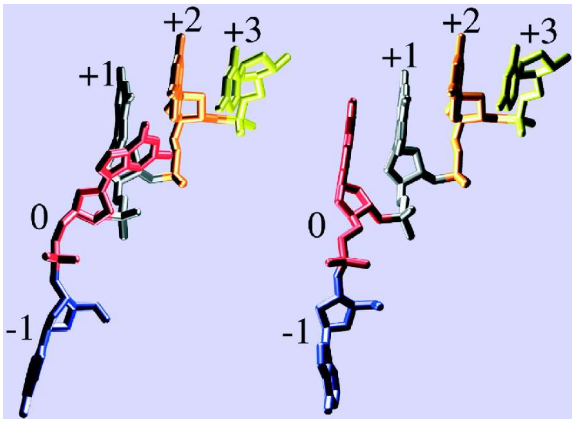


Joonis 1. *Taq* polümeraasi struktuursed domäänid. Näha on inaktiivne *proofreading* eksonukleaasne domään (roheline), peopesa sisaldav polümeraasne domään (tumeoranž), sõrmed (lillad), pöial (khaki) (28, joonis 1).

3.2 *Taq* polümeraasi töö mehhanism

Töö mehhanism on pakutud välja kristall-struktuuri põhjal. Eristatakse 2 erinevat valgu-DNA kompleksi seisundit – avatud (*open*) ja suletud *(close).* Nukleosiid-5'-trifosfaat seondub avatud seisundis olevale ensüüm-DNA kompleksile. Avatud struktuurist kinnisesse struktuuri üleminekul sõrmede domään sulgub 40-60^0 ja sihtmärkjärjestuse alus, mis paardub sissetuleva nukleotiidiga, paindub vähemalt 90 kraadi võrra struktuuri sisse *(base stacking*, vt allpool) (joonis 2), asetatades sissetuleva nukleotiidi optimaalsesse joondusesse, selleks, et teda järgnevalt siduda DNA praimeri külge. Kirjeldatud konformatsiooniline muutus toimub enne keemilisi protsesse ja arvatakse, et on kiirust-piiravaks sammuks üldises reaktsiooni kineetilises mehhanismis (algne dNTP seondumine polümeraas-praimer-sihtmärk-järjestuskompleksile toimub sihtmärk-järjestuse primaarsest struktuurist sõltumatult) (17).

Teine kiirust piirav konformatsiooniline muutus leiab aset peale keemilist reaktsiooni ensüüm-DNA kompleksi suletud seisundist avatud seisundisse minekul, kus toimub pürofosfaadi vabastamine ja DNA translokatsioon (26).



Joonis 2. 5 esimest sihtmärk-järjestuse DNA nukleotiidi avatud (vasak) ja suletud (parem) aktiivsaidi seisundis. Nukleotiidaluse '0' (järgmisena dupleksisse inkorporeeritav alus; +1 nukleotiidalus on sihtmärk-järjestus/praimer dupleksis sihtmärk-järjestuse 5' otsa viimane nukleotiid) geomeetria muutub avatud ja suletud konformatsioonide korral *(base stacking)*. (17, joonis 5)

Arvatakse, et polümeraas on seotud ca 10 aluspaariga dupleksist (18) ja ca 4 ss alusega sihtmärkjärjestusest (19). Uue nukleotiidi inkorporeerimiseks ahelasse on pakutud välja atomaarse tasemeni kirjeldavaid mudeleid (17). Näiteks globaalne ja lokaalne mudel, millest esimese järgi muudetakse rohkema kui ühe (praimeri 3' otsats alates üheahelalise sihtmärkjärjestuse järjestusest; n=2 v n=4 olenevalt polümeraasist) nukleotiidi ss-geomeetria ds-geomeetriaks ja peale sissetuleva nukleotiidi inkorporeerimist ahelasse, n-1 ds nukleotiidi saavad oma algse ss-geomeetria tagasi ning teise mudeli järgi muudetakse täpselt ühe vastava nukleotiidi ss-geomeetria ds-geomeetriaks. (17)

Polümeraasi selektiivsuse uue nukleotiidi lisamisel ahelasse määrab ära tema tertsiaarne struktuur (aktiivsait) – kõrge täpsusega polümeraasid omavad jäigemat nukleotiidi siduvat taskut *(tight nucleotide binding pocket)*, mis ei võimalda suuri geomeetrilisi muutusi olles suuruse ja kuju poolest fikseeritud (24, 25). Kuna valgu ja seonduva nukleotiidi vahel pole palju järjestuse-spetsiifilisi kontakte, siis polümeraasi selektiivsus tuleneb eelkõige selle

struktuursest omapärast. *Mismatch* aluspaarid, mis on tekkinud kahe puriini või kahe pürimidiini vahele, omavad Watson-Crick aluspaaridest erinevat struktuuri *(wobble base pairs)* ning nende inkorporeerimine sünteesitavasse ahelasse pole polümeraasi struktuuri tõttu steeriliselt võimalik. Puriin-pürimidiin *mismatchide* korral on see tõene vaid juhul *(wobble* aluspaaride moodustumine*),* kui paardunud nukleotiidide kuju on enamjaolt määratud (mitte Watson-Crick konformatsioonis) vesiniksidemete interaktsioonidega aluste vahel; sõltuvalt järjestuse kontekstist, erinevate vabade nukleotiidide olemasolust ja rauaioonidest pole polümeraasil võimalik aluseid sundida Watson-Crick konformatsiooni (polümeraasi struktuur soosib Watson-Crick konformatsioonis olevate nukleotiidide inkorporeerimist sünteesitavasse ahelasse). Vastasel korral toimub *mismatchi* lisamine ahelasse (26).

On näidatud, et 5'->3' eksonukleaasset aktiivsust mitte omavad polümeraasid on võimelised pikendama *mismatch* lõpuga praimer-sihtmärkjärjestus dupleksit, viimast küll oluliselt aeglasemalt kui 3' Watson-Crick sidet omavat dupleksit (*mismatchist* tulenevad aeglasemad konformatsioonilised muutused ja keemiliste sidemete moodustamine) (27). On näidatud, et *Taq* polümeraasi poolt katalüüsitud reakstsioonil praimeri 3' otsa erinevate nukleotiidide poolt moodustunud *mismatchid* mõjutavad moodustuva produkti kogust erinevalt. Näiteks A:G, G:A, C:C 3' *mismatchid* vähendavad PCR-i produktide kogust 100 kordselt, A:A 20 kordselt, samas teised 3' *mismatchid* ei mõjuta PCR-i produkti kogust oluliselt, kõige väiksemat mõju produkti kogusele avaldavad nukleotiidaluse T *mismatchid* G-, C-, T-ga. Siinkohal tuleb märkida, et erinevate *mismatchide* moodustumine ei sõltu vaid kasutatavast polümeraasist (olulised on ka nukleotiidide struktuur ja geomeetria, moodustuvad keemilised sidemed jne) (22).

3.3 PCR-i artefaktide teke

Kimäärsed produktid tekivad, kui toimub praimeri ebatäielik pikendamine PCR-i tsüklis (liiga lühikeses) ahela sünteesimise faasis ning, kui siht-märkjärjestus on kompleksne (nt genoomne DNA), siis järgnevas tsüklis ebatäielikult pikendatud praimer (liiga aeglase temperatuuri

jahutamise tõttu sulamistemperatuurini) seondub sihtmärk-järjestuses teise (kas pikendatud praimeri ulatuses identsesse või mitteidentsesse) kohta (va kohta, kust praimer peaks andma oodatud produkti), kus ta pikendatakse täielikult. Kimäärsete produktide teket vähendab pikema-ajaliste elongatsioonitsüklite kasutamine ning võimalikult minimaalse arvu PCR-i tsüklite kasutamine detekteeritava produkti saavutamiseks (23). Kimäärsete produktide teket ja ka deletsioonide teket PCR-i produktides seletatakse replikatsiooni (polümeraasi) libisemise *(replication slippage, copy-choice or primer-template misalignment)* mehanismiga. Replikatsiooni libisemine toimub DNA järjestuse alades, kus esinevad sekundaarstruktuurid ning kordusjärjestused. On näidatud, et vähemalt 4-s korduses $(CA)_n$ või 8-s korduses $(A)_n$ (~8bp mõlemat) traktide olemasolu korral sihtmärk-järjestuses (kasutades *Taq* DNA polümeraasi) detekteeritakse oodatud PCRi produkti pikkuse kõrval lühemaid (ka pikemaid, kuigi oluliselt vähem) produkte (29). Polümeraasi ahela-vahetamise *(strand displacement)* aktiivsust omavad polümeraasid suudavad vältida libisemist (*Taq, Pfu* polümeraas omab PCR-i rekatsiooni tingimustes (väga) madalat ahela-vahetamise aktiivsust, *Tfu* omab kõrget ahela-vahetamise aktiivsust). On näidataud, et juuksenõelstruktuuri deleteerumine toimub PCR-i reaktsiooni jooksul väga efektiivselt, va juhul, kui kasutatakse väga kõrge ahela-vahetamise aktiivsusega polümeraasi. On näidatud, et polümeraasi *(Escherichia coli DNA* polümeraas I ja *Taq, Pfu* polümeraasid*)* kontsentratsiooni vähenemisel (juuksenõelstruktuuri juurde jõudes polümeraas peatub ning dissotseerub, kui polümeraasi konts. on kõrge, siis suurem tõenäosus, et polümeraas reassotseerub) ning Mg^{2+} ioonide kontsentratsiooni suurenemisel (magneesiumioonide stabiliseeriv efekt juuksenõelstruktuurile) PCR-i segus kasvab PCR-i artefaktide teke (15).

4. DNA dupleksi struktuur ja seda koos hoidvad jõud

4.1 Dupleksi struktuur

Kaheahelaline DNA tekib omavahel Watson-Crick (edasi WC) vesiniksidemeid moodustavate A ja T ning G ja C nukleotiidide paardumisel. G-C nukleotiidaluste vahel tekib kolm ning A-T vahel kaks vesiniksidet. DNA kaksikahelas on nukleotiidalused suunatud heeliksi keskele ning suhkru- ja fosfaatjäägid jäävad struktuuris väljapoole moodustades nn heeliksi selgroo (37).

Kaheahelalisele DNA-le mõjuvad nii stabiliseerivad kui destabiliseerivad (*electrostatic repulsion*) jõud (35).

4.2 Vesiniksidemed

Kahte ahelat aitavad DNA heeliksis koos hoida vastasahelate komplementaarsete nukleotiide vahele tekkinud WC vesiniksidemed *(base-pairing)*. On näidatud, et kuigi stabiilse dsDNA sünteesiks polümeraasi poolt pole vesiniksidemeid tarvis (vesiniksidemete energeetiline väärtus on suhteliselt madal), siis aluste paardumise selektiivsuse tagamiseks on WC vesiniksidemete moodustumine tähtis (35).

4.3 Aluste *stacking*

Teine stabiliseeriv jõud on interaktsioonid samal ahelal kõrvuti asetsevate (lämmastik)aluste vahel *(base-stacking)*. See on mittekovalentne mittetriviaalne interaktsioon, mille energeetiline väärtus sõltub nii lahuse koostisest, keskkonna temperatuurist, DNA primaarstruktuurist *(stackingus* olevate nukleotiidide ja neid ümbritsevate nukleotiidide tüübist antud ahelas*)*. Tänu kaksikheeliksi struktuurist tulenevale keerdumisele, on s*tacking* võimalik ka vastasahelates olevate diagonaalis olevate nukleotiidide vahel (34). *Stacking* on ühe elektrostaatilise ja kahe mitte-elektrostaatilise komponendi summa; mitte-elektrostaatilisteks jõududeks on van der Waalsi (dispersiooni efekt) ja hüdrofoobne jõud (31). Seda, milline jõud omab aluse *stackingus*

dominatset efekti ja, milline on vastavate jõudude biokeemia, alles uuritakse (32). Aluste *stacking* on tugevaim puriinide vahel (nt A-A, G-G) ja nõrgim pürimidiin-pürimidiin vahel (nt T-T); puriin-pürimidiin *stacking* (nt A-T) on vahepealse tasemega (puriin-pürimidiin ≥ pürimidiin-puriin). Aluste *stacking* on tugevam dielektrilisemates lahustes (nt vesi) ning väiksem madalamates dielektrikutes (nt kloroform) (33). Aluste *stacking* on rohkem eelistatud DNA ahela 5' otsas kui 3' otsas (35). On näidatud, et aluste vahelised Watson-Crick vesiniksidemed ei suuda vesikeskkonnas üksi tagada dsDNA stabiilsust; stabiilsuse tagamiseks on tarvilik ka (lämmastik)aluste *stacking.* Viimane mõjub mitte ainult kahe nukleotiidi vahel, mis on vastasahelates olevate nukleotiidaluste vaheliste vesiniksidemete kaudu paardunud, vaid ka linge moodustavate lateraalsete nukleotiidaluste vahel (ssDNA-s); sellisel juhul (paardumata alades, lingudes) mõjub aluse *stacking* dsDNA-le pigem destabiliseerivalt, kuid üldine DNA stabiilsus on tänu *stackingule* suurendatud (30). Yakovchuk jt on näidanud, et DNA stabiilsus on saavutatud peamiselt aluste *stacking*-interaktsioonide tõttu; G-C paardumine (st *base pairing* ehk vesiniksidemete moodustumine) ei aita kaasa DNA dupleksi stabiliseerimisele ning A-T paardumine on alati destabiliseeriv (32).

4.4 Steerilised efektid

Nukleotiidide omavaheliste ruumiliste efektide uurimine on komplitseeritud, kuna heeliksit stabiliseeriva summaarse jõu lahutamine erinevateks komponentideks (vesiniksidemed, aluste *stacking,* steerilised efektid) pole selgepiiriline. Steeriliste efektide olemust ning määra on uuritud tuginedes erinevate *mismatchide* energeetilistele väärtustele; sõltuvalt *mismatche* moodustavate nukleotiidide geomeetriast, peaks nukleotiidide struktuurist tulenevaid ruumilisi kitsendusi eeldades erinevate *mismatchide* energeetiline väärtus olema erinev. Viimane on ka tõestatud (36). Samuti on näidatud, et dsDNA otsmised *mismatchid* on energeetiliselt vähem kulukad, kuna alused saavad lihtsamalt saavutada endale sobiva konformatsiooni säilitades seejuures *stackingu.* Steerilised efektid on paremini jälgitavad koos dsDNA sünteesiga polümeraasi poolt, kuna polümeraas tagab oma struktuuriga steerilisest efektist tuleneva

nukleotiidide selektiivsuse (35). Täpsemalt vt 1.3.

4.5 Fosfaatide tõukumine

Destabiliseerivaks jõuks kaksikheeliksile on elektrostaatiline efekt, mis seisneb fosfaatide tõukumises piki DNA ahelat (5'->3' fosfodiesterside kannab negatiivset laengut) ja ahelate vahel (34).

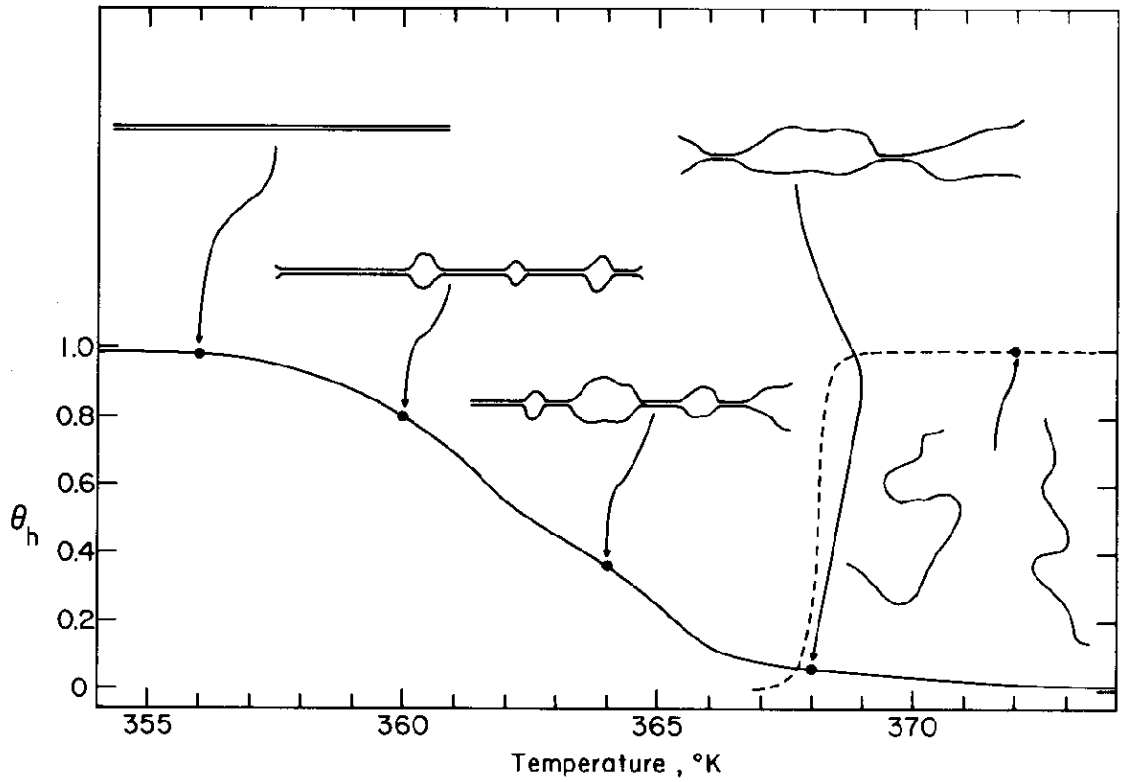
5. dsDNA termodünaamika mudel (*nearest-neighbor* mudel)

Paljude tänapäeval kasutatavate nukleiinhapetel baseeruvate molekulaarsete tehnoloogiate tarvis on DNA struktuuri (sekundaarstruktuuride moodustumise, ssDNA ahelate seondumise dsDNAks) ennustamine nii teaduslikult kui ka majanduslikult olulise tähtsusega. On teada, et antud lahuses olevate erinevate molekulide kontsentratsioonide ning lahuse temperatuuri juures, DNA dupleksi stabiilsus sõltub tema nukleotiidsest koostisest.

5.1 Kaheahelalise DNA sulamine

Toatemperatuuril on DNA kaheahelaline heeliks, kus ahelaid hoitakse koos WC sidemete ja aluste *stackingu* abil. Temperatuuri tõustes muudavad kaheahelalise DNA heeliksi konformatsiooni kaks protsessi – vastasahelate paardunud aluste vahelised vesiniksidemed võivad katkeda ja moodustada üheahelalisi DNA mulle ning piki ahelat võivad nukleotiidid lakata *stackingut* omamast. Kriitilisel temperatuuril, mil kaks ahelat on täielikult eraldi, on siiski võimalik, et üheahelalises DNA-s on märkimisväärselt *stackingut* (so sekundaarstruktuuri); temperatuuri edasisel tõstmisel see kaob ning ahelad omandavad juhusliku spiraalikujulise struktuuri (40). Üldiselt eeldatakse lühikese DNA dupleksi sulamisel kahe seisundi kiiret üleminekut üksteiseks *(two-state melting theory)* - kaks ahelat on, kas teineteisest täielikult eraldunud või on üksteisega täielikult paardunud. Pikemate ahelate sulamine arvatakse toimuvat järk-järgult – tekivad üheahelalised DNA 'sulamismullid' *(melting*

bubbles) ehk denaturatsiooni tsoonid, mida temperatuuri mõjul tekib järjest juurde ning olemasolevad suurenevad kuni ahelad on täielikult eemaldunud teineteisest (joonis 3). Minimaalne üheahelalise DNA pikkus sulamismullis on 20 bp, so ca 2 heeliksi keerdu. Seega lühikesed DNA oligod ei saagi sulada mitte-kaheetapiliselt (43,49).



Joonis 3. Pikemate DNA dupleksite lagunemine temperatuuri mõjul. Θ_h tähistab kaheahelaliste heeliksite protsenti (13).

5.2 Termodünaamika mudelid

Laialdasemalt kasutatav DNA dupleksi stabiilsust kirjeldav mudel eeldab, et DNA dupleksi stabiilsus on määratud kahe kõrvuti asetseva nukleotiidiga; enim praktikas kasutatav antud eeldusel põhinev DNA (RNA, RNA-DNA) mudel on lähima-naabri ehk *nearest-neighbor* (edasi N-N) termodünaamika mudel. Selles mudelis omistatakse kõigile võimalikele (kümme erinevat N-N interkatsiooni WC sidemeid loovate nukleotiidide vahel; AA/TT, AT/TA, TA/AT, CA/GT, GT/CA, CT/GA, GA/CT, CG/GC, GC/CG, GG/CC) dimeeridele vabaenergia

(entroopia, entalpia) väärtused ning üldistatult, dupleksi kogu vabaenergia väärtus on üksikute dimeeride energiate summa; aluste paardumise ja *stackingu* energiad on võetud kokku üheks suuruseks (38). Kaheahelalise DNA transitisooni temperatuuri (T_m) on uuritud ka Poland-Scheraga (PS) ja Peyrard-Bishop (PB) mudelite abil, millest esimene arvestab temperatuuri tõstmisel dsDNA-sse tekkiva(te) mulli(de) transitsioonidega ning teine võtab eraldi suurustena arvesse aluste *stackingut* ja paardumist (40). Lisaks eelpool kirjeldatud eeldusele baseeruvatele mudelitele on dsDNA stabiilsuse kirjeldamiseks välja töötatud erinevaid N-N eeldusele mitte põhinevaid mudeleid. Viimased on dsDNA stabiilsust kirjeldava suuruse jaotanud väga detailseteks biokeemilisteks ning biofüüsikalisteks komponentideks. Praktilisest aspektist vaadatuna N-N eeldusest sõltumatud mudelid laialdast kasutatust veel ei oma (puuduvad osad termodünaamiliste parameetrite väärtused, olemasolevaid on vähe testitud, mudeli käitumist erinevates reaktsioonitingimustes pole piisavalt uuritud jpmt) (41).

5.3 N-N termodünaamika mudel

Kaheahelalise DNA stabiilsust väljendatakse Gibbsi vabaenergiaga, mida defineeritakse järgneva valemiga:

$$\Delta G = \Delta H - T\Delta S \quad \text{(i)},$$

kus ΔG on Gibbsi vabaenergia muut, ΔH on entalpia muut, T on temperatuur kelvinites, ΔS on entroopia muut.

Dupleksi vabaenergia arvutamisel võetakse arvesse dupleksi moodustumise initsiatsiooni energia muutu kasutades selleks erinevaid väärtusi terminaalsete GC ja AT paaride (vastavalt $\Delta G^o\left(initw/termG\cdot C\right)$ ja $\Delta G^o\left(initw/termA\cdot T\right)$) jaoks, dupleksi sümmeetria korrektsiooni energia muutu $\Delta G^o\left(süm\right)$ ning vabaenergia muutusid kümne võimaliku WC lähima-naabri paari jaoks $\sum n_i \Delta G^o\left(i\right)$. Valem DNA dupleksi vabaenergia arvutamiseks on järgnev:

$$\Delta G^o\left(totaalne\right) = \sum n_i \Delta G^o\left(i\right) + \Delta G^o\left(initw/termG\cdot C\right) + \\ + \Delta G^o\left(initw/termA\cdot T\right) + \Delta G^o\left(süm\right) \quad \text{(ii)}$$

kus sümbol i tähistab lähima-naabri paari esinemise positsiooni ning n_i antud positsioonis oleva N-N paari esinemiste arvu dupleksis (36).

Termodünaamika mudelil baseeruvaid DNA stabiilsust kirjeldavaid termodünaamiliste parameetrite (entroopia, entalpia, vabaenergia väärtused erinevate N-N paaride seisundite jaoks) väärtusi on publitseerinud erinevad teaduslaborid; kuna erinevate laborite poolt tehtud mõõtmised on teostatud erinevates katsetingimustes, siis vastavate parameetrite korrektsuse otsene võrdlemine pole õige (39). Peamised erinevused parameetrite väärtuste mõõtmise katsetingimustes on soolakontsentratsioon ning mõõtmistemperatuur (enamasti 1M NaCl ja 37 ^{o}C). Enim kasutatud DNA-DNA termodünaamiliste parameetrite väärtused on publitseeritud töögruppide SantaLucia jt (36), Breslauer jt (38), Sugimoto jt poolt (42). On näidatud väga head korrelatsiooni N-N mudeli ning eksperimnetaalsete kontrollkatsete vahel, mis tähendab, et N-N mudel annab piisavalt täpse tulemuse kasutamaks seda nukleiinhapetel baseeruvates tehnoloogiates. Vastavatele termodünaamilistele parameetritele on omandatud täpsetele ligilähedased väärtused, nii et termodünaamilise stabiilsuse veel täpsemaks määramiseks tuleks muuta baasmudelit (44).

Kuigi praktilises töös on N-N mudeli rakendamine andnud häid tulemusi, on näidatud, et vaid kahe kõrvuti asetseva nukleotiidi mõju arvestamine dsDNA stabiilsuse kirjeldamiseks pole piisav. Dupleksid identsete lähimate naabrite paaridega ja identsete ots-nukleotiididega peaksid N-N eelduse põhjal olema termodünaamiliselt sama stabiilsed. Viimane on eksperimentaalselt valsifitseeritud viidates sellega mitte-lähima-naabri efektile (nt dupleksid AACUAGUU ja ACUUAAGU erinevad G väärtuste poolest 1 kcal/mol võrra) (41). Samuti on näidatud, et sõltuvalt mitte-kanooniliste nukleotiidide positsioonist (*mismatchid* dupleksi keskosas omavad suuremat destabiliseerivat efekti), on oligodupleksi stabiilsus erinev; N-N mudel ei arvesta *mismatch* nukleotiidide erinevat positsiooni dupleksis (84).

5.4 DNA oligonukleotiidsete dupleksite sulamistemperatuuri arvutamine

Sulamistemperatuur (T_m, *melting temperature*) on füüsikaline suurus, mis iseloomustab DNA dupleksi stabiilsust teatud keskkonnas. Sulamistemperatuur on temperatuur, mille juures pooled nukleotiidide ahelad (praimerid ning nende sihtmärgid) on kaheahelalised heeliksid ja pooled üheahelalistes juhuslikes spiraalides *(random coil)*. Sulamistemperatuuri ennustamine on oluline, kuna see määrab PCR-is praimeri seondumistemperatuuri (T_a, *annealing temperature*) sihtmärk-järjestusega. Liiga madala seondumistemperatuuri kasutamine soodustab ka mitte-täielike praimer/sihtmärk-järjestuste dupleksite ehk valeseondumiste tekkimist. Vastupidisel, so liiga kõrge seondumistemperatuuri kasutamise korral on ka õiged seondumised nõrgad ning ebastabiilsed ja PCR-i produkti akumuleerub vähe ja sellest tulenevalt pole PCR piisavalt tundlik (46).

5.4.1 PCR-i praimeri seondumistemperatuur

Optimaalse seondumistemperatuuri $\left(T_a\right)$ ja praimerite sulamistemperatuuride $\left(T_m\right)$ suhet on tänaseks veel vähe (õigesti) modelleeritud. Töögrupp SantaLucia jt on ennustanud PCR-i optimaalset seondumistemperatuuri, kuid tulemust pole publitseeritud, vaid on realiseeritud kommertsionaalses tarkvaras (21). Praktikas kasutatakse õigele seondumistemperatuurile piisavalt ligilähedase temperatuuri leidmiseks PCR-i tingimuste optimeerimist läbi temperatuuri gradiendi. Seejuures lähtutakse välja arvutatud praimerite sulamistemperatuurist, et saavutada maksimaalset PCR-i produkti akumuleerumist. Empiiriliselt on optimaalset seondumistemperatuuri valemit püüdnud formuleerida Rychlik jt, kasutades selleks nii praimeri kui ka produkti sulamistemperatuuri:

$$T_a^{OPT} = 0.3\,T_m^{primer} + 0.7\,T_m^{produkt} - 14.9 \qquad \text{(iii)}$$

kus T_m^{primer} on labiilsema praimeri sulamistemperatuur, $T_m^{produkt}$ on produkti sulamistemperatuur.

Valemi järgi arvutatud ja empiiriliselt saadud seondumistemperatuur erinevad üksteisest ≤0.7ºC. Valemi kasutamisel tuleb arvestada asjaoluga, et valemi välja töötamiseks ja praimeri sulamistemperatuuri arvutamiseks on kasutatud N-N parameetrite väärtusi, mis on avaldatud Breslauer jt poolt. Viimased aga on mõõdetud katsetingimustes, mis pole PCR-i katsetingimustele piisavalt sarnased ning sellest tulenevalt erinevus ennustatud praimeri seondumistemperatuur ja tegeliku seondumistemperatuuri vahel on oluline (46).

5.4.2 PCR-i praimeri sulamistemperatuur

Sõltuvalt lühemate ja pikemate DNA dupleksite erinavatest temperatuurile vastavatest transitsiooni mudelitest, kasutatakse vastavate dupleksite sulamistemperatuuri arvutamiseks erinevaid lähenemisi. Pikemate DNA dupleksite (nt produktide sulamistemperatuuri) arvutamiseks kasutatakse erinevaid Baldino jt poolt arendatud sulamistemperatuuri valemi modifikatsioone (47):

$$T_m^{Baldino} = 81.5 + 16.6 * \log\left(K^+\right) + 0.41 * GC_{sisalduse\ \%} - 675 / pikkus_{produkt} \quad \text{(iv)}$$

kus $\log\left(K^+\right)$ on kümnendlogaritm monovalentsete (soola-) metalliioonide kontsentratsioonist.

Laborites leiab lühemate DNA dupleksite arvutamiseks veel kasutust üldistatud valem, mis vaatamata oma triviaalsusele annab suhteliselt häid tulemusi. Valem johtub oligos sisalduvate GC ja AT nukleotiidide sisalduse protsendist omistades G ja C nukleotiididele sulamistemperatuuri arvutamisel suurema kaalu (48):

$$T_m = 2\ ^oC\left(A+T\right) + 4\ ^oC\left(C+G\right) \quad \text{(v)}$$

Detailsem valem, mis eeldab kaheetapilist oligodupleksi sulamist baseerub N-N mudelil:

$$T_m = \frac{\sum\left(\Delta H_d^o\right) + \Delta H_i}{\sum\left(\Delta S_d^o\right) + \Delta S_i + \Delta S_{self} + R * \ln\left(C_T / x\right)} \quad \text{(vi)}$$

kus ΔH_d^o ja ΔS_d^o on vastavalt entalpia ja entroopia muut, $\Delta H_i, \Delta S_i$ vastavalt dupleksi moodustumise initsiatsiooni entalpia ning entroopia muudud, ΔS_{self} on entroopne penalti

järjestuse iseendaga komplementaarsuse eest, R on universaalse gaasi konstant (1.987cal/K*mol), C_T on molaarne üheahelalise oligo kontsentratsioon, x võrdub neljaga mitte-komplementaarsete ahelate korral ja ühega iseendaga komplementaarsete ahelate korral. Kuigi valem vi on põhjalikum kui v, on ka esimene (vi) tegelikke biofüüsikalisi-keemilisi protsesse üldistav, kuna tegelikkuses kaheahelaline DNA ei järgi rangelt kaheetapilist mudelit, samuti on viimased uuringud näidanud, et valemi (vii) eeldus, et lühikeste kaheahelaliste DNA oligote soojusmahtuvuse muut (*heat capacity change*) ahelate sulades on $C_p \approx 0$, pole korrektne (50). Kaheetapiliselt arvatakse sulavat enamus 4-20 bp (6-50 bp) pikkuseid duplekseid, kuid valem (vi) sobib ka mitte-kaheetapiliselt sulavate lühikeste dupleksite piisavalt täpse sulamistemperatuuri arvutamiseks (43,57).

Sulamistemperatuur ja N-N parameetrite väärtused sõltuvad lahuses olevatest soola-metalliioonidest. Kuna N-N parameetrite väärtuste mõõtmised on tehtud teatud soolakontsentratsiooni sisaldavas lahuses, siis sulamistemperatuuri valemi kasutamisel teistsuguse soolakontsentratsiooni juures, peab sulamistemperatuuri arvutamise valemisse lisama soola kontsentratsiooni parandavat liikme. Viimasest on juttu punktis 7.1.

6. DNA dupleksi sekundaarstruktuurid ja nende termodünaamilised parameetrid

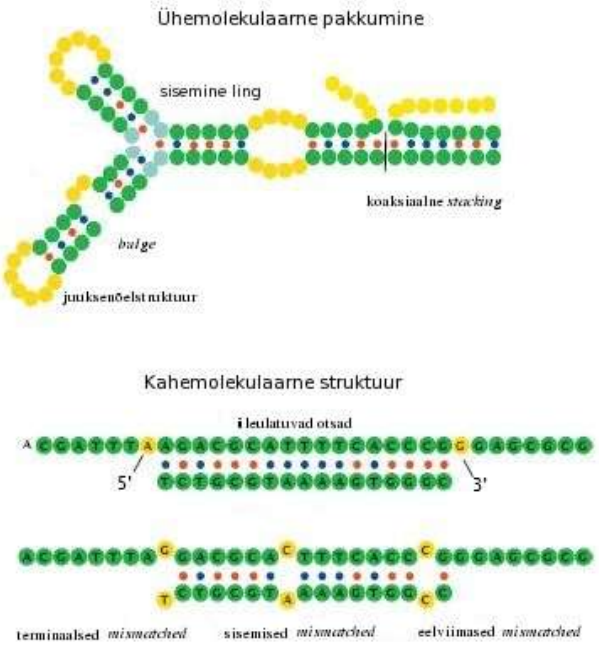
Kaks üheahelalist DNA molekuli moodustavad täieliku paardumise vaid olukorras, kus molekulid on mitte pikema ahela ulatuses teise mitte lühema ahela vastavas piirkonnas komplementaarsed. Kui ahelad sisaldavad mitte-komplementaarseid nukleotiide, võib tekkida energeetiliselt stabiilne dupleks, mis on vähem stabiilne kui dupleks täieliku paardumise korral sisaldades erinevaid sekundaarstruktuure (joonis 4). Selliste dupleksite moodustumist on võimalik ennustada arvutades vastava sekundaarstruktuuri stabiilsuse ehk vabaenergia väärtuse.

Töögrupi SantaLucia jt poolt on avaldatud siiamaani suurim erinevate termodünaamiliste parameetrite ja nende väärtuste kollektsioon. Viimane sisaldab N-N mudelil baseeruvaid termodünaamilisi väärtusi kaheahelaliste nukleotiidhapete erinevate sekundaarstruktuuride jaoks. Täielik avaldatud parameetrite list puudub veel dupleksis terminaalsete mitte-paardunud nukleotiidide jaoks.

Üksikud **sisemised mitte-paardunud nukleotiidid** *(single internal mismatches)* on ühest aluspaarist koosnevad mitte-komplementaarsed sisemised nukleotiidid, mis paiknevad kõrvuti paardunud nukleotiididega. Sisemised *mismatchid* võivad olla nii dupleksit stabiliseerivad kui ka mitte stabiliseerivad. Nukleotiidipaarid stabiilsuse järgi langevas järjekorras:
G-C > A-T > G-G > G-T ≥ G-A > T-T ≥ A-A > T-C ≥ A-C ≥ C-C. 'G' nukleotiid on kõige 'valimatum' nukleotiid, kuna moodustab kõige tugevama aluspaari ning kõige tugevamad *mismatch* paarid, samas 'C' nukleotiid on kõige diskrimineerivam moodustades tugevaima paari ja kolm kõige nõrgemat *mismatchi* (44).

Erinevat mõju dupleksi koguenergiale avaldavad **otsmised üleulatuvad nukleotiidid** *(terminal dangling ends)* ning **terminaalsed mitte-paardunud nukleotiidid** *(terminal mismatches)*. Terminaalsed mitte-paardunud nukleotiidid on dupleksit alati stabiliseerivad.

Terminaalse mitte-paardunud nukeleotiidipaari (nt $\begin{matrix} GA \\ CA \end{matrix}$) energeetilise väärtuse arvutamiseks on võimalik dimeer kaheks otsmiseks üleulatuvat nukleotiidi omavaks nukleotiidi-paariks jaotada ($\begin{matrix} GA \\ \mathrm{C} \end{matrix}$ ja $\begin{matrix} \mathrm{G} \\ CA \end{matrix}$) ning kasutada vastava sekundaarstruktuuri stabiilsuse arvutamiseks üleulatuvate nukleotiidide parameetreid. On näidatud, et terminaalne mitte-paardunud nukelotiidipaar omab võrdväärset (kuigi vähem stabiilne) energeetilist väärtust kahe vastava üleulatuva nukleotiidi energeetilise väärtusega (45). Erinevat mõju dupleksi stabiilsusele annavad ka eelviimased või eel-eelviimased *(penultimate, pen-penultimate)* otsmised *mismatchid* võrreldes otsmiste või sisemiste *mismatchidega* (44).



Joonis 4. Erinevad võimalikud dupleksi sekundaarstruktuurid. Näidatud on ühe üheahelalise DNA molekuli (üleval) pakkumisel ja kahe üheahelalise DNA ahela seondumisel moodustatud võimalikud struktuurid. Vastavad struktuurid on näidatud kollasega.

Koaksiaalne (kahe dupleksi) ***stacking*** *(coaxial stacking)* tekib kõrvuti asetsevate üheahelaliste DNA molekulide vahel, kui kaks oligonukleotiidi asetsevad (seonduvad) sihtmärkjärjestusel kõrvuti või kui oligonukleotiid seondub juuksenõelstruktuuri moodustanud sihtmärk-DNA dupleksi osa kõrvale. Fenomeni võib vaadelda kui ahela katkestust *(nick) (*44).

Lingu-taolised struktuurid on **juuksenõelstruktuur** *(hairpin loops),* **sisemised lingud** (*internal loops*), vähamalt ühe nukleotiidi pikkused **insertsioonid-deletsioonid** (*bulge)*, **mitmeharulised lingud** (*multibranched loops*). Erinevate lingu-taoliste struktuuride stabiilsuse ehk vabaenergia väärtuse arvutamiseks kasutatakse sõltuvalt struktuuri omapärast erinevaid modifikatsioone eelpool toodud DNA dupleksi vabaenergia väärtuse arvutamise valemist ii (44).

7. Soolade mõju DNA dupleksi stabiilsusele, soolakorrektsiooni arvutamine

Lahuses stabiliseerivad peale veeioonide kahe- ja üheahelalist DNAd seal olevad metalliioonid. Kuna metalliioonid seonduvad rohkemal määral kaheahelalisele kui üheaehlalisele DNAle, siis dupleksi formeerumine on ka seetõttu energeetiliselt soodustatud (57). Positiivsed metalliioonid stabiliseerivad negatiivset laengut omavat biheeliksit (vähendavad negatiivseid jõude).

7.1 Erinevad katioonid, mõju dupleksi stabiilsusele

Käsitletakse eraldi monovalentseid ning bivalentseid katioone, kuna need stabiliseerivad biheeliksit erineval tasemel seondudes erineva affiinsusega, erinevatesse kohtadesse (DNA heeliksi suur ja väike vagu) ja erinevate nukleotiidide osadega kaheahelalisel DNA-l (nukleiinhapete fosfaatgrupid, nukelosiidalustega - N7 puriinidel ja O6 guaniinil). Kui erinevad monovalentsed katioonid (K^+, Na^+) käituvad biheeliksi stabiliseerimisel võrdväärselt, siis kahevalentsete katioonide ioonide korral on erinevate ioonide mõju dupleksile erinev (54-56). Võrdsete soola kontsentratsioonide juures samade dupleksite korral sulamistemperatuuri väärtus T_m on vastavalt pingereale: $MgCl_2 > CaCl_2 > MnCl_2 \gg LiCl > NaCl \approx KCl \approx CsCl$. On näidatud, et K^+ ja Na^+ ioonid seonduvad DNAle järjestuse ja struktuuri spetsiifiliselt (54) ning nende efekt dupleksi stabiilsusele on sõltuv fosfaatide arvust dupleksis ning dupleksi nukleotiidsestjärjestusest (53).

Kuna N-N mudeli termodünaamiliste parameetrite väärtused on mõõdetud kindla kontsentratsiooniga (enamasti 1M NaCl) monovalentsete katioonide juuresolekul, siis on vajalik algoritm soola kontsentratsiooni parandamiseks vastavalt kasutatavale soolale ja kontsentratsioonile. Välja on pakutud meetodeid ennustamaks sulamistemperatuuri muutust vastusena soola kontsentratsiooni muutusele. Eristada või kahte üldist lähenemist soolakorrektisooni arvutamiseks – järjestuse primaarstruktuurist (nukleotiidsest koostisest) sõltuv ja sõltumatu (57).

7.2 Primaarstruktuurist sõltumatu soolakorrektisoon

Dupleksi sulamistemperatuuri muutus vastusena soolakontsentratsiooni muutusele on sõltuv dupleksi pikkusest ja sõltumatu järjestuse kompositsioonist. Vanim (tuntuim) sellist idealoogiat järgiv soolakorrektsiooni valem on avaldatud Schildkraut ja Lifsoni poolt (58).

$$T_m(2) = T_m(1) + 16.6 * \log \frac{\left[Na^+\right]_2}{\left[Na^+\right]_1} \quad \text{(vii)}$$

kus $T_m(1)$ on valemi (vi) järgi arvutatud sulamistemperatuur, $\left[Na^+\right]_1$ ja $\left[Na^+\right]_2$ on vastavalt monovalentsete ioonide kontsentratsioon mida kasutati termodünaamiliste parameetrite väärtuste mõõtmisel ning monovalentsete katioonide kontsentratsioon, mida soovitakse kasutada reaalses eksperimendis.

Vaatamata sellele, et valem (vii) on välja töötatud kasutades väga kitsast soolakontsentratsiooni vahemikku ning valemi näol tehtud üldistusele ei leidu kinnitust, on valem veel ka tänapäeval rutiinselt ja laialdaselt kasutusel soolakorrektisooni valemina (57).

Soolakontsentratsioonist on sõltuvad kaheahelalise DNA entroopia ja vabaenergia suurused; entalpia muut arvatakse olevat soolakontsentratsioonist sõltumatu (tõestatud Na^+ ioonide kontsentratsiooni 0.05 kuni 1.1 M juures) (36,44). SantaLucia jt poolt pakutud sulamistemperatuuri valem, mis arvestab soolakorrektisooniga (57):

$$\frac{1}{T_m(2)} = \frac{1}{T_m(1)} + \frac{0.368 * N}{\Delta H^o} \ln \frac{\left[Na^+\right]_2}{\left[Na^+\right]_1} \quad \text{(viii)}$$

Autorite järgi töötavad samad valemid naatriumi, kaaliumi ja ammooniumi ioonide korral. Valem (viii) on leidnud praktikas palju kasutust ning näib olevat andnud korrektseid tulemusi (44).

7.3 Primaarstruktuurist sõltuv soolakontsentratsioon

Dupleksi soolakontsentratsioonist sõltuv stabiilsus on seotud biheeliksi GC nukleotiidide sisalduse määraga. Ulatuslikke uuringuid soola efektist DNA termilisele stabiilsusele on läbi viinud ning erinevate avaldatud soolakorrektsiooni valemite täpsust on hinnanud töögrupp

Owczarzy jt. Viimaste poolt avaldatud detailne valem, mis arvestab kasutatavat soola kontsentratsiooni on järgmine (57):

$$\frac{1}{T_m(2)} = \frac{1}{T_m(1)} + \left(4\ .29 * f(GC) - 3\ .95\right) * 10^{-5} \ln \frac{\left[Na^+\right]_2}{\left[Na^+\right]_1} + 9\ .40 * 10^{-6} \left(\ln^2\left[Na^+\right]_2 - \ln^2\left[Na^+\right]_1\right)$$

(ix)

kus $f(GC)$ tähistab GC nukleotiidide osakaalu dupleksis.

Antud valem on DNA kontsentratsioonist ning dupleksi pikkusest sõltumatu.

Uuringud näitavad valemi (ix) sulamistemperatuuri ennustamise täpsuseks ±1.6, (viii) ja (vii) täpsuseks vastavalt ±2.6 ning ±5.5 (57).

7.4 Bivalentsete katioonide teisendus monovalentseteks

Termodünaamilised valemid, mis arvestavad lahuse soolakontsentratsiooniga on välja töötatud baseerudes monovalentsetele $\left(\text{täpsemalt } Na^+\right)$ ioonidele. Vähe on teostatud vastavaid mõõtmisi kasutades bivalentseid katioone, kuigi praktikas on vaja sageli kasutada valemeid bivalentsete katioonide jaoks.

Ahsen jt. pakkusid välja järgmise valemi divalentsete katioonide kontsentratsiooni teisendamiseks monovalentsete katioonide kontsentratsiooniks (51):

$$\left[Na^+_{eq}\right] = \left[Monovalentsed\ katioonid\right] + 120 * \left(\sqrt{\left(\left[Mg^{2+}\right] - \left[dNTPs\right]\right)}\right) \quad \text{(x)}$$

kus $\left[Na^+_{eq}\right]$ tähistab naatriumioonide ja tema ekvivalentide kontsentratsiooni. Valem arvestab bivalentsete katioonide seondumisega trinukleotiidele ning näitab, et bivalentsete ja monovalentsete katioonide kontsentratsioonide erineva mõju vahel pole lineaarset sõltuvust. Lineaarset sõltuvust vastavate katioonide mõjude vahel usuvad olevat teised uurimisgrupid, kes on leidnud 80-100 kordse (52) või 140 kordse (53) erinevuse Na^+ ja Mg^{2+} ioonide mõju vahel.

II PRAKTILINE OSA

TÖÖ ISELOOMUSTUS JA EESMÄRGID

PCR-i metoodika on jätkuvalt kasutusel erinevates eluvaldkondades (meditsiin, toiduainetööstus, teadustöö jt) ja seda kindlasti ka tulevikus; seetõttu on oluline, et antud metoodika oleks usaldusväärne. Uuritud on erinevaid PCR-i mõjutavaid üksikasjaolusid (sulamistemperatuur, dupleksite vabaenergia, polümeraasi spetsiifika, katioonide mõju jt), kuid vähe või peaaegu polegi läbi viidud süsteemseid uurimusi PCR-i tulemusi määravate tunnuste välja selgitamiseks. Ilmne on, et üksikute spetsiifiliste PCR-i mõjutavate asjaolude uurimine on tarvilik PCR-i õnnestumise ennustamiseks, kuid mitte piisav; ilma süsteemsete uuringuteta ei saa hinnata reaalselt toimuvate PCR-i katsete edukust. Oleme läbi viinud uurimuse PCR-i tulemusi mõjutavate tunnuste välja selgitamiseks bakteriaalsetes genoomides.

Bakteriaalsete patogeenide identifitseerimine toimub paljudes laborites tavalise PCR-i *(conventional* PCR*)* tehnoloogia abil. Rutiintöös kasutatavate praimeripaaride peamine puudus on ebapiisav sensitiivsuse tase; samuti on oluline spetsiifilisuse probleem. Töö eesmärgiks on luua mudel, mis ennustab iga kandidaat-praimeripaari tundlikkuse taset.

Püüame leida praimeripaaride arvutatavaid tunnuseid, mis kõige tugevamini ennustavad PCR-i tundlikkust (st on korrelatsioonis oodatud bändi intensiivsusega). Nimetatud tunnused moodustavad summaarse mudeli, mille abil on võimalik hiljem ennustada praimeripaari tunnuste põhjal PCR-i sensitiivsuse astme (bändi intensiivsust) tõenäosust. Eriline rõhk oli kordusjärjestustele disainitud praimerite käitumise uurimisel.

Töö tulemusena valmis PCRi kirjeldav mudel (valem), mis võimaldab ennustada praimerite, produkti järjestuse kaudu PCRi edukust ja tundlikkust antud tehnoloogilses keskkonnas.

ALGANDMED JA METOODIKA

1. Andmete päritolu ja struktuur

Tööd alustati sekveneeritud ja vabalt kättesaadavate bakteriaalsete genoomijärjestuste hankimisega (TABEL 1). Genoomijärjestused laaditi alla NCBI (*National Center for Biotechnology Information)* ftp-lehelt (*ftp.ncbi.nih.gov*). Ajal, mil tööd alustati (sept, 2004) oli mainitud lehelt võimalik alla laadida 377 täisgenoomi. Lisaks kopeeriti kahe genoomi järjestused (*Neisseria meningitidis serogrupp C FAM18* ning *Neisseria lactamica*) *The Wellcome Trust Sanger Institute'i* ftp-lehelt (*ftp://ftp.sanger.ac.uk/pub/pathogens/*). Genoomide järjestused olid FASTA-formaadis failis.

Kuna kliinilises proovis leidub ka inimese genoomset DNA-d, siis oli tööks vajalik ka inimgenoomi järjestus. Viimane oli kohalikku serverisse alla laaditud ENSEMBLI ftp-leheküljelt (*ftp://ftp.ensembl.org/pub/*, NCBI inimese genoomi versioon 35). Andmed on talletatud FASTA-formaadis failis.

TABEL 1. Töös kasutatavad genoomiandmed

Genoomijärjestuse allikas	***Organism***	***Genoomide arv***
NCBI	bakter	377
Sanger	bakter	2
NCBI	inimene	1
	KOKKU	380

2. Uuritavad bakteritüved

Lähema uurimise alla võetakse 5 inimese nakkushaigusega seotud patogeeni. Bakterite valimisel oli oluline täispika genoomijärjestuse olemasolu ning kättesaadavus. Huvi pakkuvateks bakteritüvede kandidaatideks olid sellised patogeenid, kelle tuvastamiseks

kvaliteedilt rahuldavad PCR-i praimerid seni puudusid. Välja valituteks osutusid (sulgudes genoomi pikkus) - *Neisseria gonorrhoeae FA 1090* (2,153,922 bp), *Helicobacter pylori 26695* (1,667,867 bp), *Treponema pallidum subsp. pallidum str. Nichols (1,138,011* bp*), Chlamydia trachomatis D/UW-3/CX* (1,042,519 bp), *Mycoplasma genitalium G37* (580,074 bp).

3. Liigispetsiifiliste kordusjärjestuste leidmine

Et tagada bakteriaalsete genoomide identifitseerimise suuremat tundlikkust PCR-i läbi, otsustasime disainida praimerid bakteriaalsetele kordusjärjestustele. PCR-i täpsuse probleemi lahendab nõue, et kordusjärjestus peab olema antud bakteriaalse genoomi spetsiifiline.

Üldine põhimõte kordusjärjestuste leidmiseks on järgmine:

1. uuritava genoomi järjestus *tükeldatakse* n bp pikkusteks m bp võrra ülekattes (nn samm) olevateks lõikudeks.

2. eelmises etapis saadud n bp pikkuseid lõike kasutatakse joondusprogrammi BLAST (63) päringuna (andmebaasina kasutatakse uuritava organismi genoomset järjestust), et leida kõik sarnased järjestused genoomist.

3. järjestused, mis ületavad eelnevalt paika pandud *cutoffi* (läve) joondatakse kõigi teiste olemasolevate bakteriaalsete genoomsete järjestuste vastu kasutades selleks teatud konstandi võrra madalamat *cutoffi* väärtust kui organismist endast vastete leidmiseks.

4. eemaldataksejärjestused, mille joondamisel saadakse *cutoffist* kõrgem skoor.

5.1 kui lõigupikkus n on võrdne soovitud kordusjärjestuse pikkusega S või, kui kõik järjestused on eemaldatud lõpeb programmi töö

5.2 kui punkt 5.1 on väär, siis tehake leitud kordusjärjestustele (st iga leitud kordusjärjestuse koopiatele omavahel) mitmene joondus, et teha kindlaks, kas korduse mõlemal otsal on piisavalt järjestikuseid (ca keskmine praimeri pikkus + konstant) konserveerunud nukleotiide, et disainida sinna praimer.

5.2.1 kui mõlemal järjestuse poolel on piisavalt ruumi, siis lisatakse järjestuse mõlemale

otsale *flanking* regioonid pikkusega s

5.2.2 kui vaid ühel järjestuse otsal on ruumi, siis lisatakse vastavale poolele *flanking* regioonid pikkusega 2 x s.

5.2.3 kui kummalgi otsal pole ruumi, siis järjestus eemaldatakse

5.2.3.1 kontrollitakse, kas on veel järjestusi - kui ei ole, lõpetab programm töö

6. minnakse tagasi punkti 2.

Näiteks kasutades sammu 100 ja algset lõigu pikkust 100 leiame antud genoomist üles kõik vähemalt 200 bp pikkused kordusjärjestused.

Saadud kordusjärjestustest eemaldatakse nn korduvad kordusjärjestused, st näiteks kui lõik 1 on kuskil mujal genoomis asuva lõigu 2 (kaks) koopia, siis otsides eelpool kirjeldatud metoodikaga kordusi, saame kaks erinevat, kuid sisuliselt sama kordusjärjestust: kordusjärjestus 1 - lõik1 koopiaga lõik 2 ja kordusjärjestus 2 - lõik 2 koopiaga lõik 1; alles jäetakse sellisel juhul kordusjärjestus 1, teine eemaldatakse.

4. Mitmene joondus ning praimeri disain

Kuna leitud kordusjärjestuste koopiad pole identsed üksteisega, siis nad võivad sisaldada mittepaardunud *(mismatch)* või puuduvaid/lisa nukleotiide *(indel).* Tarvilik on enne praimeri disaini vastavad varieeruvad nukleotiidid ära maskeerida, et vältida praimerite disaini mitte-ühemõttelistele nukleotiididele. Antud kordusjärjestuse koopiad joondatakse omavahel mitmest joondust võimaldava programmiga CLUSTALW (74), mis asetab *mismatch* nukleotiidide kohale tärni (*) ja puuduvate nukleotiide kohale kriipsu (-). Seejärel asetatakse mittekanooniliste ja lisatud nukleotiide positsioonidele sümbol, mis ei kuulu hulka {A,T,G,C}, puuduva nukleotiidi korral märgitakse puuduvast nukleotiidist järgnev (positsiooni +1) nukleotiid.

Leitud kordusjärjestustele disainitakse praimerid programmi PRIMER3 versiooni 1.0 (73) modifitseeritud variandiga. PRIMER3 on laialdaselt kasutusel olev PCR-i praimerite

disainimisprogramm, mis võimaldab kasutajal erinevaid järjestusel ja katsetingimustel põhinevaid parameetreid muuta (maksimaalne-optimaalne-minimaalne väärtus, nt produkti pikkuse vahemik, järjestikuste GC nukleotiidide arv, praimeri pikkus, sulamistemperatuur, praimeri GC protsent, monovalentsete katioonide kontsentratsioon, oligote kontsentratsioon, praimer dimeeride mitte moodustumist tagavad parameetrid ehk kui palju võivad praimerid omavahel ja iseendaga paarduda). Kuigi PRIMER3 kasutab praimerite sulamistemperatuuri arvutamiseks lähima-naabri mudelil põhinevat valemit (46), on viimane ebapiisava täpsusega sulamistemperatuuri arvutamisel (ei arvesta nukleotiidist sõltuvat dupleksi initsiatsiooni entroopiat, initsiatsiooni entalpia võrdub 0-ga, kasutab soolakorrektsiooni arvutamiseks valemit ix) ja Breslauer jt (38) poolt avaldatud termodünaamiliste parameetrite tabelit, mille väärtuste mõõtmine on toimunud suhteliselt palju aega tagasi (enne 1986, mõõtmiste teostamiseks täpne tehnoloogia puudus) ning PCR-i reaktsioonitingimustest kaugetes reaktsioonitingimustes (temperatuuril 25 ºC). Realiseerisime PRIMER3 programmis PCR-i praimerite sulamistemperatuuri arvutamiseks töögrupi Santa-Lucia töögrupi poolt avaldatud lähima-naabri mudelil põhineva valemi vi (36), mis järgib summaarsete termodünaamiliste parameetrite arvutamiseks valemi iii põhimõtet. Soolakorrektsiooni arvutamiseks realiseeriti kaks valemit - Santa-Lucia jt (36) ning Owczarzy jt (57) poolt avaldatud valemid vastavalt viii ja ix. Kasutatavad termodünaamiliste parameetrite väärtused on avaldatud Santa-Lucia töögrupi poolt (36). Divalentsete katioonide teisendamiseks monovalentseteks katioonideks kasutati valemit x (51).

5. Sõltumatute tunnuste leidmine

Edasi vaadeldakse iga disainitud praimeripaari kui vaatlust või objekti, mida iseloomustavad erinevad sõltumatud tunnused (Y) ja eksperimentaalselt määratav sõltuv tunnus X (PCR-i bändi intensiivsus).

Üldiselt võib vaadeldavaid sõltumatuid tunnuseid jagada järgnevateks alapunktideks ja gruppideks (tunnuste iseloomu, arvutamiseks kasutatud programmi ja kasutatud genoomi alusel):

A. praimeripaaridele arvutatavad tunnused, mis võib jagada järgnevateks gruppideks:

GRUPP 1. Nukleotiidsed tunnused. Praimeri 3' otsa ja praimeri 3' otsa kõrval oleva amplikoni nukleotiidne koostis.

1.1 (1.-4.) 1 ja 2 nukleotiidi mõlema praimeri 3' otsast

1.2 (5.-12.) 1 ja 2 nukleotiidi praimeri 3' otsa kõrval olevast amplikonist (iga erineva amplikoni kohta)

1.3 (13.-20.) praimeri 3' otsa 3 nukleotiidi ja praimeri 3' otsaga külgneva amplikoni 2 nukleotiidi kombinatsioon (kui kordusjärjestuse koopiatest moodustunud amplikonidel on erinevad nukleotiidid, siis antakse vastavad nukleotiidid iga erineva amplikoni kohta).

GRUPP 2. Kasutatud DNA kontsentratsioonid.

2.1 (21.) inimese DNA kontsentratsioon (µg/µl)

2.2 (22.) kasutatud bakteri DNA kontsentratsioon (µg/µl)

2.3 (23.) inimese ja kasutatud bakteri DNA molaarsete kontsentratsioonide suhe

GRUPP 3. Praimerite vabaenergia väärtused (24. - 35.).

Praimeri t 3' otsa alamjärjestuse s pikkusega $|s| \in \{6,8,10,12,14,16\}$ nukleotiidi deltaG maksimaalsed ja minimaalsed väärtused.

Väärtused arvutatakse kohaliku programmi deltagtest.c muudetud versiooniga (deltagtest1.c). Antud akna parameeter antakse minimaalse (väiksemat deltaG väärtust omav praimer) ja maksimaalse (suuremat deltaG väärtust omav praimer) väärtusena. Programmile deltagtest1.c antakse ette oligote järjestustega fail *(query),* pikkus *(length),* millele tahetakse vabaenergia väärtust arvutada, väljund suunatakse standardväljundisse. Programm kasutab vabaenergia väärtuse leidmiseks valemit ii ja Santa-Lucia jt poolt avaldatud termodünaamiliste väärtuste tabelit (36).

GRUPP 4. Praimerite GC nukleotiidide sisaldus (36.-47.).

Praimeri t 3' otsa alamjärjestuse s pikkusega $|s| \in \{8,10,12,14,16\}$ nukleotiidi ja täispika $|t|$ praimeri GC nukleotiidi väärtused maksimaalsete ja minimaalsete väärtustena.

GRUPP 5. Produktide omadused (48.-53.).

5.1 (48.) produkti GC nukleotiidide sisaldus (skaalas [0-1])

5.2 (49.) stabiilseima sekundaarstruktuuri vabaenergia väärtus. Väärtuse arvutamiseks kasutatakse nukleiinhapete sekundaarstruktuure ennustavat ja energeetilist stabiilsust arvutavat programmi MFOLD (75). MFOLD arvutab DNA järjestuse vabaenergia ette antud temperatuuril ja Mg^{2+} kontsentratsiooni juures. Temperatuurina kasutatakse kahe praimeri sulamistemperatuuri aritmeetilisest keskmisest teatud konstandi võrra madalamat temperatuuri ning Mg^{2+} kontsentratsiooniks divalentsete katioonide kontsentratsiooni.

5.3 (50.) maksimaalne lokaalne GC nukleotiidide protsent. Leitakse maksimaalne GC nukleotiidide sisaldus kasutades selleks akna pikkusi alates $min(|t1|,|t2|)$ kuni $|prod|$, kus $|t1|,|t2|$ on vastavalt vasaku ja parema praimeri pikkus ning $|prod|$ on produkti pikkus.

5.4 (51.) maksimaalne lokaalne deltaG. Analoogiliselt eelmise punktiga leitakse iga akna maksimaalne keskmine deltaG väärtus ühe nukleotiidi kohta (antud akna deltaG väärtus jagatakse akna pikkusega) ning seejärel korrutatakse saadud deltaG väärtus antud akna pikkusega, st leitakse kõrgeimat keskmist deltaG väärtust omava akna vabaenergia väärtus.

GRUPP 6. Kahe praimeri pikkuste vahe ja amplikoni pikkus (52.-53.).

GRUPP 7. Kordusjärjestuste arv, millele praimeripaar on disainitud (54.).

GRUPP 8. Praimerite nukleotiid-nukleotiid valeseostumiste arvud bakteris (55.-64.). Väärtuste leidmiseks kasutatakse programme GINDEXER ja GTESTER praimeridisaini paketist GENOMEMASKER (alamrakendusest GenomeTester) (59). GTESTER on programm, mis võimaldab leida praimerite

seostumiste ja produktide arvu ja asukohad antud genoomsest järjestusest. GTESTER vajab tööks programmi GINDEXER poolt loodud indeksfaili, milles on sorteeritult kirjas kõigi antud sõna (järjestuse) s pikkusega $|s| \in \{8,10,12,14,16\}$ sõnade esinemiste positsioonid antud genoomis (genoomne järjestus antakse GINDEXER-ile ette FASTA-formaadis failina). GTESTER kasutab kahendotsimise algoritmi, et leida indeksfailist antud järjestuse (praimeri) 3' otsa vastava pikkusega (vastavalt, millise pikkuse järgi on moodustatud indeksfail, nt otsib praimeri 3' otsa kõik 16 nukleotiidi pikkused esinemised antud genoomis) järjestuse esinemiste kohad genoomis. Leitud esinemiste positsioonide järgi arvutab võimalike tekkivate produktide arvu, pikkuse ja positsiooni (*http://bioinfo.ebc.ee/genometester/*).

8.1 (55.-57.) maksimaalne seondumiste arv (maksimaalne kahest praimerist praimeripaaris) arvutatuna aknas $|s| \in \{8,10,12,14,16\}$ alustades praimeri 3' otsa esimesest nukleotiidist.

8.2 (58.-64.) maksimaalne seondumiste arv arvutatuna aknas $|s| \in \{8,10,12,14,16\}$ alustades praimeri t 3' otsa $t[i+1],t[i+2]$ jne nukleotiidist kuni $t[k]$ nukleotiidini, kus i=0, $k=|t|-|s|$, $|s|$ on antud alamjärjestuse ehk akna pikkus ja $|t|$ on praimeri pikkus.

<u>GRUPP 9.</u> Praimerite nukleotiid-nukleotiid seostumiste arvud inimeses (65.-74.). Sama, mis eelmine, kuid arvutatuna inimese genoomis.

<u>GRUPP 10.</u> Praimerite deltaG põhised valeseostumiste arvud bakteris (75.-110.). Kasutatud vabaenergia arvutamise valem ii ja termodünaamilised parameetrid publitseeritud Santa-Lucia jt poolt (36).

Arvutatakse programmiga FASTAGREP (85).

10.1 (75.-79.) maxG; maksimaalne deltaG põhiste seondumiste arv (st rohkemaid seondumisi omava praimeri praimeripaarist seondumiste arv). Arvutatakse juhuslike järjestuste pikkusega $|s| \in \{8,10,12,14,16\}$ nukleotiidi keskmine deltaG väärtus (ühekordne arvutamine) ning leitakse kõigi vastavast deltaG väärtusest stabiilsemate seondumiste arvud kasutades praimeri alamjärjestuse pikkusi $|s| \in \{8,10,12,14,16\}$ nukleotiidi. Mittekanoonilised nukleotiidid pole

lubatud; FASTAGREP võtmega -dgregion3

10.2 (80.-84.) maxGp; sama, mis eelmine, kuid leitakse praimeri praimeripaarist maksimaalne seondumiste arv arvutatuna MAX(praimeri 3' otsa 2st nukleotiidist loetud seondumiste arv, praimeri 3' otsa 3st, praimeri 3' otsa 4st jne kuni $|t|-|s|$ nukleotiidist loetud seondumiste arv), kus $|t|$ on praimeri pikkus ja $|s|$ vastava alamjärjestuse pikkus. FASTAGREP võtmega -dgregion3.

10.3 (85.-88.) max_dg; arvutatakse akendes $|s| \in \{10,12,14,16\}$ nukleotiidi; leitakse antud akna keskmisele deltaG väärtusele vastava antud praimeri alamjärjestuse pikkus (programmiga deltagtest.c); FASTAGREP leiab antud akna keskmisest deltaG väärtusest stabiilsemad seondumised, kui on täidetud tingimus, et leitud alamjärjestuse pikkuse (ei pruugi olla sama pikk kui antud aken) kattes on praimer täielikult seondunud. FASTAGREP võtmega -dg.

10.4 (89.-92.) max_dgp; sama, mis eelmine, kuid arvutatuna MAX(praimeri 3' otsa 2st nukleotiidist loetud seondumiste arv, praimeri 3' otsa 3st, praimeri 3' otsa 4st jne kuni $|t|-|s|$ nukleotiidist loetud seondumiste arv), kus $|t|$ on praimeri pikkus ja $|s|$ vastava alamjärjestuse pikkus. FASTAGREP võtmega -dg.

10.5 (93.-96.) max_dgreg; arvutatakse akendes $|s| \in \{10,12,14,16\}$ nukleotiidi; leitakse antud akna keskmisele deltaG väärtusele vastava antud praimeri alamjärjestuse pikkus, võetakse praimerist vastav alamjärjestus (kas on oluline vaadata antud vabaenergia väärtusest stabiilsemate ja järjestuse pikkuselt maksimaalselt sama pikkade alamärjestuste seondumisi; võrreldes täispika praimeri otsimisega väiksem ajakulu) ning leitakse kõik vastava alamjärjestuse seondumised, mis on stabiilsemad kui antud keskmine deltaG väärtus; mittekanoonilised nukleotiidid on lubatud. FASTAGREP võtmega -dgregions;

10.6 (97.-100.) max_dgregp; sama, mis eelmine, kuid arvutatuna MAX(praimeri 3' otsa 2st nukleotiidist loetud seondumiste arv, praimeri 3' otsa 3st, praimeri 3' otsa 4st jne kuni $|t|-|s|$ nukleotiidist loetud seondumiste arv), kus $|t|$ on praimeri pikkus ja $|s|$ vastava alamjärjestuse pikkus. FASTAGREP võtmega -dgregions;

10.7 (101.-104.) max_dg_tot; mõlema praimeri jaoks arvutatakse 10.3 ja 10.4 kõigi antud

pikkusega läbitud akende summa ning leitakse praimeri praimeripaarist maksimaalne seondumiste arv.

10.8 (105.-108.) sum_dg_tot; sama, mis eelmine, kui viimase sammuna leitud maksimumi asemel leitakse summa.

10.9 (109.-110.) max_dgregions ja min_dgregions; arvutatakse täispikkade praimerite vabaenergia väärtused (programmiga SantaLucia.cpp, mis kasutab valemit i ning vajab sisendiks lisaks praimeri järjestusele monovalentsete katioonide ja oligote kontsentratsiooni ning programmi PRIMER3 poolt arvutatud sulamistemperatuuri). Leitud vabaenergiaväärtust ja antud praimeri täispikka järjestust kasutades leitakse programmiga FASTAGREP võtmega -dgregions praimeri seondumised antud bakteris.

GRUPP 11. Praimerite deltaG põhised seostumiste arvud inimeses (111.-128.).

11.1 (111.-114.) sama, mis 10.3

11.2 (115.-118.) sama, mis 10.4

11.3 (119.-122.) sama, mis 10.7

11.4 (123.-126.) sama, mis 10.8

11.5 (127.-128.) sama, mis 10.9

GRUPP 12. Võimalike valeproduktide arvud bakteris (129.-133.). Arvutatud programmiga GTESTER kasutatdes aknaid pikkusega $|s| \in \{8,10,12,14,16\}$

GRUPP 13. Võimalike valeproduktide arvud inimeses (134.-138.). Arvutatud programmiga GTESTER kasutatdes aknaid pikkusega $|s| \in \{8,10,12,14,16\}$

GRUPP 14. '*specificity-determining subsequences*' ehk SDSS parameetrid arvutatud bakteris (139.-140.). Arvutatud töögrupi Miura jt poolt avaldatud parameetrina (76). Iga praimeri jaoks leitakse SDSS pikkus $|s|$ ning leitakse vastava praimeri SDSS pikkusega alamjärjestuse s esinemiste arv (millest on lahutatud õiged seondumised) antud genoomis. Parameeter on antud maksimaalse ja minimaalse väärtusena.

SDSS parameetri arvutamise põhimõte: antud suurus f, mis on sihtmärk-järjestusega seondunud

praimeri osa ehk fraktsioon. f väärtus kõigub vahemikus [0;1]; täielikult paardunud dupleks omab f väärtust ühe lähedal ning, mida ebastabiilsem dupleks on, seda kaugemal on väärtus ühest. Leitakse praimeri pikkus, mille juures f väärtus ületab eelnevalt määratletud f väärtuse *(threshold)*:

$$f = \frac{Cp_0 * K_{as}}{1 + Cp_0 * K_{as}} \qquad \text{(xi)}$$

kus $K_{as} = e^{(\Delta G / RT)}$, Cp_0 on oligote algne kontsentratsioon, K_{as} on assotsiatsiooni konstant, G on vabaenergia muut, R on gaasikonstant ja T praimerite sulamistemperatuur Kelvinites. Saadud *threshold*-ile vastava praimeri pikkuse järgi otsitakse praimerite seondumisi antud genoomis. f-väärtuse ja vastava praimeri pikkuse leidmise programm sdss.pl kasutab programmi fraction_sdss.c leidmaks f väärtuse antud praimeri pikkusele. fraction_sdss.c kasutab vabaenergia väärtuse leidmiseks valemit i (divalentsete katioonide teisendamiseks monovalentseteks Miura jt järgi Nakano jt poolt avaldatud suurust, (53)) ja parameetrite väärtusi, mis on avaldatud Santa-Lucia töögrupi poolt (36). Vastava praimeri pikkuse esinemised leitakse programmiga sdss_blast.pl, mis kasutab seondumiskohtade otsimiseks programmi BLAST (63).

GRUPP 15. '*specificity-determining subsequences*' ehk SDSS parameetrid arvutatud inimeses (141.-142.). Vt eelmine.

GRUPP 16. PCR-i temperatuur (143.-145.). Kasutatud seondumistemperauuri absoluutväärtus, seondumistemperauuri erinevus kõrgema sulamistemperatuuriga praimerist, seondumistemperatuuri erinevus madalama sulamistemperatuuriga praimerist.

B. üldised katseid iseloomustavad tunnused

GRUPP 17. (146.-157.) Sellesse rühma kuuluvad eksperimentaalsete katsete tulemusena tekkinud erinevate tunnuste väärtused. PCR-i masina bloki veeru number, PCR-i masina bloki rea number, PCR-i masina bloki number, geelelektroforeesi ja PCR-i tegemise nädalapäev (E,T,K jne), geeli ja PCR-i sooritanud laborandi nimi, aeg päevades alates katsete tegemise

algusest, toatemperatuur katsesegu kokkusegamise ajal, kellaaeg (astronoomilise tunni täpsusega).

6. Praimerite valimine eksperimentaalseteks katsetusteks

Punktis 4 kirjeldatud metoodikaga disainitakse ühele kordusele praimereid nö varuga, et oleks võimalik eksperimentaalseteks katsetusteks valida praimereid võimalikult erinevate tunnuste väärtustega andmestikust. Viimane on tarvilik selleks, et saaks leida eksperimentaalsete katsete järgselt statistiliselt usaldusväärselt seoseid eelpool kirjeldatud tunnuste ja PCR-i õnnestumise, tundlikkuse ning täpsuse vahel. Kandidaatpraimerite seast andmete võimalikult suurt varieeruvust kirjeldavate praimerite (maksimaalselt informatiivsete) ehk esindava valimi *(representative sample)* välja valimine on võimalik statistiliste meetoditega. Ülesannet, kus tahetakse leida võimalikult erinevaid objekte (so praimeripaare) on võimalik lahendada klasterdamise meetoditega. Klasterdamine on objektide jagamine rühmadesse ehk klastritesse, nii et ühes klastris asuvad objektid on üksteisele rohkem sarnased kui klastrite vahelised objektid. Tahetakse leida nii palju erinevaid klastreid kui palju on planeeritud praimeripaare eksperimentaalseteks katsetusteks; tahetakse leida igast klastrist kõige rohkem teistest klastritest erineva liikme (nn klastri tsentri). Kuna arvutatud tunnused on üksteist osaliselt dubleerivad (sarnaselt käituvad tunnused ehk võrdväärset infot sisaldavad tunnused, nt kui ühe tunnuse väärtus kasvab mingi suhteosa võrra, siis teeb seda ka teise tunnuse väärtus; nähtust nimetatakse kollineaarsuseks, kui tegu on rohkema kui kahe tunnusega, siis multi-kollineaarsuseks), siis praimeripaaride klasterdamine kõigi välja arvutatud tunnuste järgi pole üheselt interpreteeritav. Olukord lahendatakse vaadeldes esmalt klasterdavate objektidena tunnuseid, st leiame üksteist dubleerivate tunnuste rühmad ning valime nende seast välja võimalikult informatiivsed ja üksteist välistavad tunnused, st vähendame tunnuste arvu; seejärel püüame praimeripaare grupeerida klasterdamise tulemusena leitud vähemate

tunnustega.

Klasteranalüüsiks kasutatakse statistikapaketti SAS (SAS Institute Inc. 2006. Version 9. *http://support.sas.com/documentation/onlinedoc/sas9doc.html*) protseduuri FASTCLUS, mis võimaldab suuri andmestikke kiiresti rühmitada. Protseduur realiseerib mittehierarhilist klasterdamismeetodit k-keskmist (*k-means*), millele tuleb ette anda moodustavate klastrite arv. Sobivaim (sisaldab optimaalset võimalikult informatiivsete tunnuste komplekti) klastrite arv leitakse nn katse-eksituse meetodil proovides läbi erinevaid moodustavate klastrite arve ning analüüsides vastavaid klastreid moodustavate komplektide sisulist tähendust. Leitud optimaalsetest klastritest võetakse klastri tsentrile kõige lähemad sisuliselt mõttekad tunnused.

Kuna arvutatud tunnuste väärtused omavad väga laia (nt 1 kuni 10,000) ning erinevat (nt pidevad tunnused vahemikus 1-10,000 ning -5.00 kuni +1.00) muutumispiirkonna vahemikku, samuti ei pruugi erinevate tunnuste väärtused olla normaaljaotusega (mis on enamuste statistiliste meetodite eelduseks), siis mõttekate klastrite tekitamiseks on tarvilik andmete eelnev standardiseerimine. Viimaseks kasutatakse Johnsoni neljaparameetrilist jaotuste perekonda *(Johnson's system of distributions)* (77), mis surub tunnuste väärtuste muutumispiirkonnad kokku ning teisendab andmed normaalkujule. Johnsoni teisenduse kasutamiseks oleme realiseerinud vastava algoritmi määrates Johnsoni parameetrite väärtused 4-protsentiili meetodil (78). Kasutaja annab programmile ette normaliseeritud normaaljaotuse N (0;1) jaotusfunktsiooni fikseeritud väärtuse z (väärtus, mille järgi andmed normaaljaotuse sarnaseks teisendatakse; väärtused vahemikus]0;1[), faili veerud -f (eraldatud komaga ja/või järjestikuste veergude kasutamise korral sidekriipsuga), mida soovitakse teisendada ning vastav andmefail.

7. Eksperimentaalsete katsete kirjeldus

Eksperimentaalsed katsed viiakse läbi võimalikult standardiseeritud tingimustes, et vältida PCR-i katsesüsteemist tulenevat varieeruvust. Viimane on vajalik selleks, et hiljem oleks võimalik efektiivselt teostada katsetulemuste ning praimeripaaridele arvutatud faktorite vaheline usaldusväärne ning mõttekas statistiline analüüs. Sellest tulenevalt kasutatakse katseseeriates ka sama kunstlikku kliinilist proovi (antud genoomi kunstliku kliinilise proovi lahjendamine toimub üks kord kõikide katseseeriate tarvis), tagamaks kõikide katseseeriate jooksul testitava proovi samasuse ning vältimaks seeläbi tundlikkuse tõusu tingituna analüüsitava genoomi erinevast kontsentratsiooni tasemest proovis.

Kliinilised proovid Katsetes kasutatakse inimese ja uuritavate bakterite genoomide kunstlikke kliinilisi proove. Vastavate proovide valmistamisel kasutatakse, kas ainult uuritava ATCC (*American Type Culture Collection)* tüve DNA-d või testitava bakteritüve ATCC DNA proovi segatuna evolutsiooniliselt lähedase liigi või inimese genoomse DNA-ga. Kirjeldatud segusid kasutatakse vastavalt disainitud praimeripaaridele, kas sensitiivsuse või spetsiifilisuse uurimiseks.

Katsetes kasutatavaks inimese genoomseks DNA-ks on HEK293 (T-rex) rakuliinist puhastatud DNA kontsentratsiooniga $1x10^{-3}$ µg/µl ($1x10^{-5}$ µg/µl), millesse lisatakse teatud lahjendusastmega ATCC tüve DNA-d (näiteks *Mycoplasma genitaliumi* puhul $1x10^{-8}$ µg/µl ja $1x10^{-9}$ µg/µl). ATCC tüve lahjendusastmed saadakse eksperimentaalselt katsetades, kuna erinevad patogeenid koos inimese DNAga võivad käituda erinevalt (sõltuvalt genoomi suurusest, korduvatest genoomi osadest, GC nukleotiidide sisaldusest jpm). Kindlustamaks tulemuste usaldusväärsust testitakse kõiki erinevate ATCC tüve lahjendusastmetega proove vähemalt kolme PCR-i kordusreaktsiooniga. Reaktsiooni korral, kus vastavalt disainitud praimeritele lisatakse segusse vaid ühe ATCC tüve DNAd (praimeripaarid, mille spetsiifilisust testitakse uuritavas organismis endas), võetakse (vajadusel) kunstliku kliinilise proovi

kontsentratsioon 10 korda suurem. See on tarvilik, kuna tuleb arvestada asjaoluga, et testitavat DNAd seondub nii PCRi tuubide kui ka pipetiotsikute seinte külge. Reaktsiooni korral, kus vastavalt disainitud praimeritele lisatakse segusse kahe evolutsiooniliselt lähedase organismi DNAd, võetakse kunstliku kliinilise proovi kontsentratsioonid võrdseteks ($1x10^{-6}$ µg/µl).

Reaktsioonitingimused

Fikseeritud on PCR reaktsiooni maht (50 µl), mis sisaldab 1,25 U Hot-start polümeraasi; 0,2 µM mõlemat praimerit; 200 µM igat järgnevat desoksüribonukleotiidi dATP, dGTP, dCTP, dTPP; 1x PCR puhvrit; 2,5 mM $MgCl_2$ lahus; 10 µl kunstlikku kliinilist proovi.

PCR-i proovid segatakse kokku toatemperatuuril. PCRi masinas on fikseeritud kõik peale seondumistemperatuuri (sõltuv praimeri sulamistemperatuurist ning valitakse vastavalt sellele): esimene denaturatsioon 95 ^{0}C 15 min, järgnevad 40 tsüklit – 95 ^{0}C 1 min, hübridisatsioonitemperatuur vastavalt praimeripaaris olevate praimerite madalamat sulamistemperatuuri omava praimeri sulamistemperatuurist paar kraadi madalam temperatuur 30 sek, süntees 72 ^{0}C 40 sek, peale 40 tsüklit lõpusüntees 72 ^{0}C 5 min.

Katseid teostavad kaks inimest - teadur ja laborant. Võimaluse korral viib katseid läbi ainult laborant viimaks inimfaktori viga võimalikult väikeseks.

Katsete jaoks on kasutada kolme plokiga PCRi masin. Iga katse juures pannakse kirja, millist plokki masinast ja, millisesse positsiooni proovid asetatakse. Viimane on vajalik, kuna erinevad PCRi masinaosad soojenevad erineva kiirusega ja erineva tasemeni.

Produkti detekteerimine

PCR produkti olemasolu testitakse agaroos-geelelektroforeesil. Vastavalt bändi tugevusele tehakse märge 4 ühikulises skaalas (vastavalt bändi/märke tugevusele): 0 (produkt puudub); + (nõrk); ++ (keskmine); +++ (tugev). Jäädvustatakse ka PCR-i produkti pikkus. Samuti määratakse oodatavale produktile lisaks tekkinud ebaspetsiifiliste produktide tugevus (sama skaala, mis tundlikkuse korral) ning pikkus.

Katseprotokoll

Vastav vorm on toodud Lisas 3 'Katseprotokolli vorm'. Protokolli vormistus on koostatud PCR-

i eksperimentaalsete katsete läbiviijate poolt. Protokollis on väljad praimeripaari nime, praimerite numbrite ja sulamistemperatuuride, praimeripaari seondumistemperatuuri, katsesegusse lisatavate genoomsete DNAde nimede ning nende kontsentratsioonide, reaktsioonisegu konsistentsi, erinevate PCR-i tsüklite kestuste, katsete läbiviijate nimede, PCR-i segu kokkusegamise temperatuuri, katse läbiviimise kuupäeva, kellaaja jmt jaoks.

8. Kasutatud statistilise analüüsi meetod

Praimeripaaridele arvutatud tunnuste, üldiseid katseid iseloomustavate tunnuste ja PCR-i edukuse, tundlikkuse ning täpsuse vaheliste seoste leidmiseks kasutatakse üldistatud lineaarse mudeli *(Generalized Linear Model*, GLM) ühte rakendust - logistilist regressiooni.

8.1 GLM mudel

GLM mudel sisaldab kolme komponenti:

1. juhuslik komponent, mis on funktsioontunnus ehk sõltuv tunnus Y ja tema tõenäosusjaotus *(probability distribution*; kas normaal-, binomiaal- (binaarse tunnuse korral), Poissoni (loendatavate tunnusväärtuste korral), gamma-, negatiivne binomiaaljaotus*)*
2. Argument- ehk sõltumatud tunnused X, mis võivad olla nii pidevad kui kategoorilised
3. Linkfunktsioon g(μ), mis seob omavahel komponendi 1 ja 2:

$$g(\mu)=\beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots \quad \textbf{(xii)}$$

kus on β_0 vabaliige *(intercept)*, β_1 β_2 ... on argumenttunnuste kordajad, mida soovitakse leida. Linkfunktsioon valitakse vastavalt uuritava tunnuse Y ja sõltumatute tunnuste X tõenäosusjaotusele (79). Antud olukorras võib eeldada tunnuste binomiaaljaotust X,Y~B(n,p). Binomiaaljaotust võib ette kujutada järgnevalt: meil on tunnus x (nt täispika praimeri täielike seondumiste arv), mille väärtus kujuneb juhuslike tegurite mõjude summana $x=\xi_i + \xi_{i+1} + \xi_{i+2} + \ldots + \xi_n$, kus liidetavad ξ_i on statistiliselt sõltumatud binaarsed (0 või 1 väärtusega) juhuslikud

suurused (igas genoomi positsioonis, kas praimer on täielikult seondunud või pole täielikult seondunud), kusjuures väärtus 1 esineb tõenäosusega p, mis on igal katsel (st igas genoomi positsioonis) sama ja ei muutu katseseeria (genoomi läbimise jooksul) käigus; n tähistab maksimaalset tunnuse väärtust (nt maksimaalselt saab praimer esineda igas vaadeldavas genoomi positsioonis) (80). Sellest johtuvalt kasutatakse logistilises regressioonis *logit* linkfunktsiooni, mis eeldab uuritava tunnuse binomiaaljaotust; avaldub kujul $g(\mu)=\log[\mu/(1-\mu)]$, kus μ tähistab uuritava tunnuse keskväärtust.

Logistilise regressiooni mudeli eeldusteks on adekvaatselt valitud Y tõenäosusjaotus ning sõltumatute tunnuste vahelise tugeva kollineaarsuse puudumine.

8.2 Kasutatud logistilise regressiooni protseduur statistikapaketis SAS: LOGISTIC

Binaarsete, ordinaalsete (järjestatavate) ja loendatavate funktsioontunnuste Y ning vastavate argumenttunnuste X vaheliste seoste leidmiseks on statistikapaketis SAS välja arendatud logistilise regressiooni protseduur LOGISTIC (81).

Olgu x sõltumatute tunnuste vektor, α vabaliige ja β argumenttunnuste vektor (mudeli parameetrid ehk arumenttunnuste kordajad) ning uuritava tunnuse väärtus 1 (uuritaval tunnusel kaks võimalikku väärtust), siis uuritava tunnuse tõenäosus on $\pi=Pr(Y=1|x)$ ja protseduuri LOGISTIC mudeli seosefunktsioon avaldub kujul:

$$\text{logit}(\pi)\equiv\log(\pi/(1-\pi))=\alpha-\beta x \quad \text{(xiii)}$$

Kui uuritavaks tunnuseks on rohkema kui kahe tasemega (väärtusega) tunnus (nt bändi intensiivsused), kasutatakse protseduuri LOGISTIC poolt automaatselt selle asemel kumulatiivset *logit* funktsiooni *(cumulative logit model),* mis on pööratud *logit* funktsioon ($F(x)=1/(1+\exp(-x))$). Toodud seosefunktsioonide asemel võimaldab LOGISTIC kasutada ka teisi linkfunktsioone (*probit, log-log* jt). Mudeli parameetrite hindamiseks kasutatakse protseduuri LOGISTIC poolt vaikimisi rakendatavat suurima tõepärameetodit, milleks on realiseeritud kaks iteratiivset algoritmi (Fisher-scoring ja Newton-Raphson meetod, millest esimest kasutatakse vaikimisi)

(81).

Tunnuste valimise meetod. LOGISTIC pakub tunnuste valimiseks erinevaid meetodeid (*backward, stepwise, forward, score)*. Kasutatakse *STEPWISE* optsiooni, kus tunnuste lisamine ja eemaldamine mudelisse toimub sammhaaval: lisamine ja eemaldamine toimub tunnuse mõju olulisuse järgi (χ^2).

Nullhüpoteesi testimiseks kasutatakse Wald statistikut ehk Wald χ^2 väärtust. Wald statistik on suurima tõepärameetodi versioon t-testile, mis kujutab endast ennustatud tunnuse väärtuse suhet ennustatud tunnuse standardveasse: b_1/s_{b1} (79).

Mudeli kooskõla määramiseks ja erinevate **mudelite omavaheliseks võrdlemiseks** kasutatakse c-statistikut ja hälbimust (D, *deviance*), täpsemalt hälbimust jagatuna vabadusastmete arvuga (D/DF). Mudeli kooskõla näitab, kas mudel on lähedane põhimõtteliselt kõige paremale mudelile, st ei erine statistiliselt oluliselt parimast mudelist.

c-statistic ehk *concordance index* iseloomustab mudeli spetsiifilisust ja sensitiivsust. Mudeli spetsiifilisus on näiteks, kui ennustatakse PCR-i intensiivsuse kolmandat taset, siis kõik kolmanda intensiivsuse tasemega ennustatud objektid (PCR-i praimerid) ka reaalselt annavad kolmanda PCR-i intensiivsuse taseme (st ei ennusta reaalselt nt 1. intensiivsuse tasemega objektile kolmandat intensiivsuse taset); järgides eelnevat näidet - mudeli sensitiivsus näitab, kas mudel ennustab kõigile reaalselt kolmanda intensiivsuse tasemega objektidele kolmanda intensiivsuse taseme. c-statistiku väärtus 1 tähistab ideaalset mudelit, 0.5 vastab juhuslikele sündmustele. c-statistiku väärtus ei sõltu uuritava tunnuse väärtuse sagedustest.

c-statistiku arvutamise põhimõte: ennustatakse teatud arvu vaatluste jaoks uuritava tunnuse antud väärtuse tekkimise tõenäosused; vastavalt tunnuse väärtusele (nt 0 ja 1) jagatakse ennustatud tunnuste väärtused gruppideks (nt 2 gruppi) ning hakatakse võrdlema tunnuste väärtuste tekkimise tõenäosuste kõikvõimalikke paare gruppide vahel (esimese grupi esimese vaatluse uuritava tunnuse tekkimise tõenäosus versus teise grupi esimese, teise jne vaatluse tekkimise tõenäosus; esimese grupi teise vaatluse uuritava tunnuse tekkimise tõenäosus versus

teise grupi esimese, teise jne vaatluse tekkimise tõenäosus jne). Loetakse kokku vastuolus n_d (nt 1 tekkimise tõenäosus on 0.3 ja 0 tekkimise tõenäosus on 0.2), mittevastuolus n_c (1 tõenäosus on 0.3, 0-il 0.8), võrdsed $t-n_c-n_d$ (ehk seotud e *tied*, 1 tõenäosus 0.3, 0-il 0.3) paaride arvud. Vastav valem:

$$c=\left(n_c+0\ .5\ \left(t-n_c-n_d\right)\right)/t \quad \text{(xiv)}$$

kus t on kõikide vaadeldavate paaride arv.

Hälbimus (D) mõõdab mudeli erinevust küllastunud mudelist (reaalsust perfektselt kirjeldavast mudelist), st näitab kui palju uuritava tunnuse hajuvusest jääb mudeli poolt kirjeldamata.

$$D=-2\ \left(L_M-L_S\right)$$

kus L_M küllastunud mudeli (tunnuste arv võrdne vaatluste arvuga) suurim logaritmiline tõepärafunktsioon, L_S on uuritava mudeli suurim logaritmiline tõepärafunktsioon.

D näitab kui palju uuritava tunnuse hajuvusest jääb mudeli poolt kirjeldamata. Mida suurem on mudeli hälbimus, seda halvemini on ta reaalsete andmetega kooskõlas. D/DF iseloomustab antud mudeli hajuvust kasutatud tunnuste hajuvuse korrektsuse suhtes, nimetatud suhe peab olema võimalikult väärtuse 1 lähedal, suured kõikumised (>+/-0.5) näitavad andmete üle- või alahajuvust (andmestiku tasakaalutust).

8.3 Tunnuste olulisuse määramine, kollineaarsuse probleem.

Kuna kasutatavad sõltumatud tunnused on tugevalt multi-kollineaarsed, siis pole võimalik logistilise regressiooni algoritmil leida ühtset usaldusväärset mudelit (LOGISTIC poolt kasutatav iteratiivne algoritm ei koondu). Vähendamaks kollineaarsuse probleemi leitakse esiteks 'ALGANDMED JA MEETODITE' alampunktis 5 toodud gruppide kõigile tunnustele Wald χ^2 väärtused ning vastavad tõenäosused, sorteeritakse saadud tulemused iga grupi jaoks Wald statistiku tõenäosuse järgi kasvavalt ning Wald statistiku väärtuse järgi kahanevalt (vastavad Wald hii-ruudu väärtused koos p-väärtustega on toodud gruppide kaupa Lisas 2). Seejärel leitakse iga

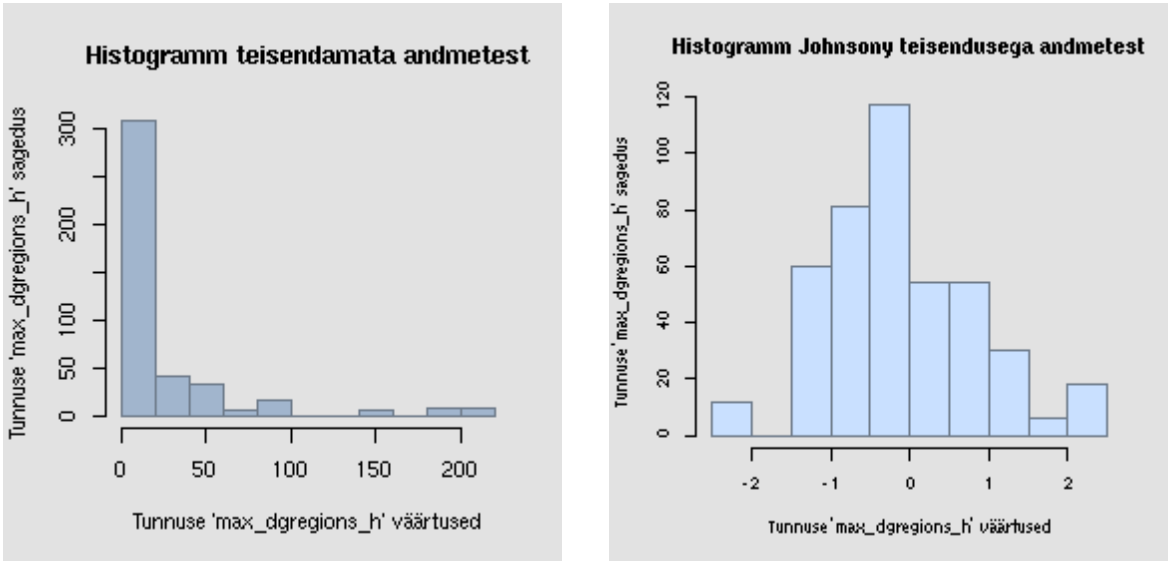
grupi jaoks üks või mitu kõige mõttekamat kõige suurema Wald statistiku väärtusega või kõige lihtsamini arvutatavat tunnust. Leitud nö peamised tunnused fikseeritakse mudelis (LOGISTIC protseduuris) ning lisatakse seejärel mudelisse ülejäänud grupi liikmed. Sel viisil protseduuri jooksutamise tulemusena leitakse mitte-kollineaarsed informatiivsed tunnused igast tunnuste grupist. Leitud tunnused pannakse jällegi ühtsesse mudelisse ning rakendatakse eelpool kirjeldatud gruppidest nö peamiste tunnuste leidmise protseduuri.

8.4 Andmete teisendamine

Andmetele (pidevaid väärtusi omavatele tunnustele) rakendati Johnsoni transformatsiooni (vt antud sektsiooni alampunkt 6), tunnuse ruutu ja kuupi võtmist (x^2, x^3), logaritmimist alusel 10 $\left(\log_{10}\right)$, tunnuse logaritmitud väärtuse $\left(\log_{10}\right)$ teise ja kolmandasse astmesse võtmist.

Transformatsioon andmetele on vajalik selleks, et andmed täidaksid kasutatava mudeli eeldusi (andmed ja mudeli viga normaaljaotusega, dispersiooni homogeensus, väikese mõjuga kõrvalekalded) ning oleks võimalik sõltumatute tunnuste ja uuritava tunnuse vahelise mitte-lineaarse seose avastamine.

Samuti on tarvilik andmete standardiseerimine (Johnsoni teisendus) üksteise suhtes, kuna erinevad tunnuste väärtused on mõõdetud erinevas skaalas (79). Näide Johnsoni teisenduse efektist on toodud JOONISEL 6.



JOONIS 6. Näide Johnsoni teisenduse teisendamata andmetest (vasakul) ja Johnsoni teisendusega (paremal) andmetest. Jooniselt on näha, et andmete muutumispiirkonnad lükatakse kitsamatesse vahemikesse ning andmed saavad normaaljaotuse.

8.4 Lõpliku mudeli koostamine

PCR-i mudeli koostamiseks kasutatakse *backward-stepwise* logistilise regressiooni analüüsi, st esmalt kaasatakse mudelisse kõik tunnused, mis leiti grupiviisilise analüüsi tulemusena ning seejärel hakatakse tunnuseid vastavalt tunnuste olulisusele (nii gruppi lisamine kui eemaldamine toimub juhul, kui $p>0.01$) eemaldama ja lisama. Kasutades parameetrite valimiseks sammuviisilist ehk tunnuste lisamise-eemaldamise meetodit, leitakse üks (paljude analoogsete seast) võimalikult kompaktne (minimaalne arv maksimaalselt informatiivseid tunnuseid, mida on võimalik arvutada optimaalse aja- ja mälukuluga) valem PCR-i intensiivsuse tasemete ennustamiseks. Tunnuste valiku kriteeriumiks kasutatakse Wald statistíkut. Mudeli kooskõla määramiseks ning ühe mudeli eelistamiseks teise ees kasutatakse võrdlusstatistikut c ja hälbimust D.

TULEMUSED

1. Eksperimentaalselt uuritavate tunnuste kirjeldus

1.1. Uuritava küsimuse defineerimine

Töö eesmärgiks on PCR-i tulemuse arvutuslik ennustamine bakteriaalsete liikide eristamiseks PCR-i abil. Uuritakse, millist tüüpi (millisest grupist) tunnused avaldavad kõige enam mõju PCR-i intensiivsuse tasemetele. Otsitakse vastust küsimusele, milline on praimereid kirjeldavate tunnuste minimaalne komplekt, mis on tarvilik ja samas ka piisav selleks, et tulevikus usaldusväärselt ennustada PCR-i tulemusi.

1.2. PCR-i mõjutavad praimerite omaduste uurimine

Siinkohal selgitatakse, millised küsimused iga arvutatud tunnuse grupi juures püstitati ja milliseid bioloogilisi protsesse erinevad tunnused võiksid kirjeldada. Kokkuvõtlikult on grupid lühikirjeldusega ning vastavate tunnuste arvutamiseks vajalike programmidega toodud TABELIS 2. Allpool toodud tunnuste jaotus gruppideks on tinglik ja on mõeldud vaid järgneva statistilise analüüsi lihtsustamiseks. Ühes grupis olevate andmete mõju analüüsiti esmalt omavahel (grupi siseselt) ja seejärel koostati lõplik mudel iga grupi olulistest tunnustest.

GRUPP 1. Praimeri 3' otsaga seotud nukleotiidide mustrid. Tahetakse välja selgitada, kas erinevad(te) nukleotiidid(e) (kombinatsioonid) praimeri 3' otsas või praimeri 3' otsa vahetus läheduses olevad amplikoni erinevad(te) nukleotiidid(e) (kombinatsioonid) või nii praimeri 3' otsa kui praimeri 3' otsa vahetus läheduses oleva amplikoni nukleotiidid koos mõjutavad PCR-i tulemusi. Kas polümeraasi jaoks on olulised praimeri 3' otsas olevad nukleotiidid (nt nukleotiidid, mis tugevalt stabiliseerivad praimeri 3' otsa ja sihtmärk-järjestuse dupleksit) või on olulised esimesed nukleotiidid, mis dupleksisse inkorporeeritakse või on/ei ole olulised mõlemad/kumbki.

GRUPP 2. Kasutatud DNA kontsentratsioonid. Püstitatakse küsimus, kas inimese või antud

bakteri(te) DNA kontsentratsioonid mõjutavad PCR-i tulemusi, sealhulgas ka kas on olulised kasutatavate kontsentratsioonide absoluutväärtused või erinevate genoomsete DNA-de suhted.

GRUPP 3. Praimerite vabaenergia väärtused. Kas praimeri 3' otsa erinevate pikkustega alamjärjestuste stabiilsus on seotud PCR-i edukusega? Kas on oluline, kui stabiilne (max) praimeri 3' ots on või, kui labiilne (min)? Milliste vabaenergia väärtuste vahemike korral PCR viib soovitud tulemuseni?

GRUPP 4. Praimerite GC nukleotiidide sisaldus. Kas piisab, kui vaadata praimeri 3' otsa erineva pikkusega alamjärjestuste või täispika praimeri GC nukleotiidide sisalduse protsenti, st tahame teada, kas piisab triviaalsest GC nukleotiidide sisalduse leidmisest või on tarvis nt termodünaamilist lähenemist, st eelmist gruppi. Milline on optimaalne GC sisalduse vahemik?

GRUPP 5. Produktide omadused. Tahame teada, kui suurt mõju avaldab PCR-i amplikoni omapära: tema ennustatud sekundaarstruktuuri stabiilsus, kas üldine GC nukleotiidide sisalduse määr on seotud PCR-i tulemuste omadustega (nt peale ahela sünteesi on kõrgema GC sisaldusega produktid tugevamalt kokku kleepunud ning PCR-i sensitiivsus on madalam) või on oluline vaadata PCR-i produkti mingi lühema osa GC nukleotiidide sisaldust (vaid teatud ala kõrge GC nukleotiidide sisaldusega põhjustab produktide tugeva kleepumise, kuna pikemad oligod ei sula nn kahe-etapilise mudeli järgi, vaid teatud madalama stabiilsusega alad denatureeruvad esmajärjekorras) või vabaenergia väärtust (samad ideed, kui GC sisaldust vaadates ainult termodünaamika kontekstis).

GRUPP 6. Varia. PCR-i produkti pikkus, praimerite pikkus ja nende pikkuste vahe. Milline on lühim PCR-i produkt, mida on võimalik usaldusväärselt detekteerida (mida pikemad kordusjärjestused, seda vähemate eksemplaridena ehk koopiatena nad genoomis esinevad)? Kas on oluline praimerite pikkuste vahe (või piisab PRIMER3 poolt ennustatud sarnasest sulamistemperatuuri väärtusest kahele praimerile praimeripaaris)?

GRUPP 7. Kordusjärjestuste arv, millele praimeripaar on disainitud. Kas kehtib seos bändi tugevuse ja kordusjärjestuste arvu vahel? Kui kehtib, siis milline (mida rohkem kordusi, seda intensiivsem bänd või leidub kasutatud suurima ja väikseima kordusjärjestuste arvu vahel

teatud optimaalne kordusjärjestuste väärtus, mis viib intensiivsuse haripunktini)?

GRUPP 8, 9. Ennustatud praimerite seostumiskohtade arv vastavalt bakteri ja inimese genoomides. Praimeri seondumiskoht modelleeritud kindla pikkusega nukleotiidse järjestusena (näiteks 12 nukleotiidi praimeri otsast).

PCR-i tundlikkust mõjutab praimerite valede seondumiste arv, kuna rohkete valeseondumiste tõttu langeb katsesegus õigetesse regioonidesse seonduvate praimerite ja sellega seoses ka soovitavat produkti sünteesiva ensüümi kontsentratsioon. Meid huvitab, kui paljusid valeseondumisi võib praimer maksimaalselt omada, et oleks veel võimalik kliinilise proovi (so madala ja varieeruva) kontsentratsiooni juures produkti detekteerida.

Püstitatakse viis peamist küsimust:

1) kas praimeri seondumiste arv bakteris on oluline;

2) kui on oluline, siis kas piisab seondumiste arvu loendamisest otse praimeri 3' otsast või on tarvis/piisav praimeri maksimaalne seondumiste arv loetuna alates 2st, 3st jne praimeri 3' otsa nukleotiidist;

3) kui 1. punkt on oluline, siis milline akna pikkus on kõige olulisem

4) Mitu seondumist võib praimer maksimaalselt/minimaalselt omada, et see ei mõjutaks PCR-i tulemusi negatiivselt?

5) Kas oluline on seondumiste arvu vaatamine bakteris või inimeses või mõlemas/mitte kummaski?

GRUPP 10, 11. Ennustatud praimerite seondumiskohtade arv vastavalt bakteri ja inimese genoomides. Praimerite seostumiskoht modelleeritud deltaG põhiselt (varieeruva pikkusega nukleotiidne järjestus praimeri 3' otsast).

Püstitatakse samad küsimused, mis GRUPP 8-9 korral, kuid vabaenergia kontekstis. Kas on tarvis praimerite seostumiste arvu leidmine kasutades termodünaamilist lähenemist? Küsimus püstitatakse, kuna termodünaamikal põhinevate seondumiste otsimine antud genoomist on oluliselt ressursinõudlikum (nii aja kui ka kettamahu poolest) kui kindla pikkusega nukleotiid-nukleotiid seondumiste otsimine. Kui vabaenergial baseeruvate seondumiste otsimine on

vajalik, siis millises praimeri pikkuse/regiooni ulatuses ning, kas on vajalik *mismatchide/indel*-idega seondumiste leidmine? Viimane ajanõudlikum kui täielikult paarduvate seondumiste leidmine. Millist termodünaamikal põhinevat seondumiste viisi (praimeri 3' otsa 1st, 2st jne nukleotiidist alates, mil määral lubada seondumisse mitte-kanoonilisi nukleotiide) on kõige optimaalsem kasutada (milline seondumiste viis peegeldab kõige enam PCR-i tulemuste omadusi). Kas on tarvilik seondumiste vaatamine inimese genoomis (võrreldes bakteri genoomiga väga ajakulukas) või on piisav/tarvilik vaid bakteri genoomis?

GRUPP 12, 13. Võimalike valeproduktide arvud vastavalt bakteris ja inimeses. Kas ennustatud produktide arvud mõjutavad PCR-i tulemusi? Millise akna pikkusega ennustatud produktide arv mõjutab PCR-i tulemusi (edu, tundlikkust)? Millise akna pikkusega ja, millise produkti pikkusega leitud valeprodukt on veel/juba detekteeritav? Kas antud parameetrid on olulised bakteris-inimeses või mõlemas/mitte-kummaski?

GRUPP 14, 15. SDSS parameetrid vastavalt bakteris ja inimeses. Kas antud parameeter on oluline tavalises PCR-is (avaldatud artiklis (76) kirjeldati SDSS parameetri olulisust PCR-i tehnoloogias, kus kasutatakse vaid ühte unikaalset praimerit - *adaptor-tagged competitive* PCR, ATAC-PCR ning *rapid amplification of cDNA ends,* RACE). Kas on tarvilik arvutada igale praimerile individuaalselt tema seondumiste leidmiseks genoomist vabaenergia väärtusel põhinev alamjärjestuse pikkus? Kas kasutatud f väärtus on õigustatud (mida kõrgem f väärtus, seda tõenäolisem on, et järjestus seondub genoomis alternatiivsesse kohta, st seda lühemat alamjärjestust praimerist vaadatakse; kasutatud f väärtus on vastava artikli autorite poolt soovitatud, so 0.01)? Kas on oluline bakteris, inimeses või mõlemas/kumbaski? Kas antud parameetri arvutamine õigustab ennast (ajanõudlik)?

GRUPP 16. PCR-i temperatuur. Kuna sulamistemperatuuri õige valimine on PCR-i protsessis väga tähtis, siis püüame mõista, kas oluline on PCR-is kasutataud seondumistemperauuri absoluutväärtus või temperatuuride erinevus madalamast-kõrgemast praimerite praimeripaarist arvutatud sulamistemperatuurist.

1.3. Teised PCR-i mõjutavad tunnused

GRUPP 17. Siia gruppi kuuluvad mitte-arvutuslikud ehk järjestustega mitte seotud sõltumatud tunnused (nn laboritunnused). Antud grupp on vajalik, et olla kindel, et PCR-i tulemuste omadused ei ole tingitud eksperimentaalsete katsete käigus tekkinud vigadest või katsetingimustest. Samuti tahame uurida, kui suurt mõju avaldavad erinevad praimeridisainist mitte-sõltuvad asjaolud PCR-i tulemuste omadustele. Näiteks, kas leidub seos nädalapäeva ja PCR-i tulemuste vahel, katsete algusest möödunud aja (päevades) ja PCR-i tulemuste vahel, proovi asendi plokis ning PCR-i tulemuste vahel, toatemperatuuri katsesegu kokkusegamise ajal ja PCR-i tulemuste vahel jmt.

TABEL 2. Praimeripaaride sõltumatud tunnused. Helehalliga PCR-i edukust ja tundlikkust mõjutavad tunnused; tumehalliga nn laboritunnused. Sulgudes antud perli programmi (*.pl) siseselt käivitatud programmide nimed.

Grupi nr	Kirjeldus	arvutamiseks kasutatud programm
1	Nukleotiidsed tunnused	primers_param.pl
2	Kasutatud DNA kontsentratsioonid	perli alamprogramm: sub molar_conc
3	Praimerite vabaenergia väärtused	primers_param.pl (deltagtest1)
4	Praimerite GC nukleotiidide sisaldus	primers_param.pl
5	Produktide omadused	prod_param.pl (MFOLD, deltagtest1)
6	Produkti pikkus, praimerite pikkuste erinevus	primer3_to_GtBlin.pl, primers_param.pl
7	Sihtmärkjärjestuste arv genoomis	finalRepeats.pl
8-9	Praimerite nukleotiid-nukleotiid seostumiste arvud vastavalt bakteris ja inimeses	count_anneals.pl (GTESTER)
10-11	Praimerite deltaG põhised seostumiste arvud vastavalt bakteris ja inimeses	deltaG.pl (deltagtest, SantaLucia)
12-13	Võimalike valeproduktide arvud vastavalt bakteris ja inimeses	count_anneals.pl (GTESTER)
14-15	SDSS parameetrid vastavalt bakteris ja inimeses	sdss.pl (fractions_sdss), sdss_blast.pl (BLAST)
16	PCR-i temperatuur	primer3 (muudetud)
17	Järjestustega mitte seotud tunnused (laboritunnused)	-

2. Liigispetsiifiliste kordusjärjestuste otsimine

2.1. Liigispetsiifiliste kordusjärjestuste otsimise eesmärk

Bakterite tuvastamisel PCR-i abil on peamiseks puuduseks tundlikkuse madal määr. Fakt, mida praimerite disainimisega parandada ei saa ja, millega tuleb arvestada, on kliinilistes proovides on detekteeritava genoomi DNA kontsentratsioon enamasti madala tasemega ja proov sisaldab

varieeruvat hulka erinevaid genoome. Püütakse disainida praimereid, mis võimaldaks tundlikumat bakteriaalsete patogeenide detekteerimist. Tundlikkuse tõstimiseks on kujundatud praimereid ribosomaalseid geene sisaldavatele genoomijärjestustele, kuid ka see ei taga piisavat tundlikkuse määra, lisaks võib tekkida probleeme ribosomaalsete operonide heterogeensusega. Püütud on PCR-i tundlikkust tõsta disainides praimerid bakteriaalsetele plasmiididele, kuid sellise lähenemise puuduseks on aeg-ajalt esinev plasmiidse DNA kadumine rakkudest, mistõttu pole võimalik tuvastada antud liiki proovist.

Käesolevas töös uurisime võimalust kasutada tundlikkuse suurendamiseks liigispetsiifilisi kordusjärjestusi. Selleks leidsime igast bakteriaalsest genoomist liigispetsiifilised kordusjärjestused ning disainisime PCR-i praimerid leitud kordusjärjestustele. Üks töö eesmärkidest on hinnata kordusjärjestuste mõju PCR-i õnnestumisele ja tundlikkusele.

2.2. Kordusjärjestuste leidmine

Genoomid, millest kordusjärjestusi otsiti on toodud sektsioonis ALGANDMED JA METOODIKA punktis 2 (5 liiki). Otsitakse unikaalseid kordusjärjestusi alates pikkusest 200 bp (jagasime genoomi 100 bp pikkusteks lõikudeks ning liikusime 100 bp pikkuse aknaga ning 100 bp sammuga, tsükli erinevates sammudes suurendasime otsitava lõigu pikkust 100 bp; otsimise metoodika on toodud detailselt eelmises sektsioonis alampunktis 3) kuni 500 bp, kuna optimaalne PCR-i produkti pikkuse vahemik, mida PCR-i abil detekteerida, on 200-500 bp. Lühemate produktide detekteerimine on tihti problemaatiline, kuna lühemad produktid kontamineeruvad süsteemis ning nende detekteerimise on tõenäosus on väike.

Korduste otsimisel välditakse selliseid kordusi, millele leidub sarnane järjestus teatud teises genoomis, nii et PCR-i praimerite disainimine sellistele järjestusele võib põhjustada PCR-i valepositiivsete tulemuste tekkimise. Selliseid kordusi nimetatakse ebaspetsiifilisteks kordusteks (JOONIS 8). Spetsiifilise korduse näide toodud Lisas 1.

JOONIS 8. Näide ebaspetsiifilisest kordusest (CLUSTALW joondus). Huvi all olev liik on *Chlamydia trachomatis D/UW-3/CX* (kordus sinine, GenBank ID AE001273), mittespetsiifilisust põhjustav genoom on *Chlamydia muridarum Nigg* (kordus punane, GenBank ID AE002160).

Leitud kordusjärjestuste koopiate arv korduse kohta varieerub 2-18ni (st ühele kordusjärjestusele vastav konsensusjärjestus esineb genoomis nt 16 koopiana). Kõige vähem ning lühemad kordusjärjestused leiti liigile *M. genitalium,* enam kordusi leiti suurema genoomiga bakteritüvedele. Praimerite disainiks otsiti ka ühe koopiana genoomis leiduvaid liigispetsiifilisi järjestusi. Ühekoopialised kordused võimaldavad võrrelda PCR-i tulemusi, kui kasutatakse praimereid, mis on disainitud ühe või rohkema järjestuse peale (tahame teada, kas amplikoni korduste arv mõjutab PCR-i tulemusi). Leitud kordusjärjestuste arv liikide ja erineva koopiaarvuga kordusjärjestuste pikkuste lõikes on toodud TABELIS 3.

TABEL 3. Leitud vähemalt kahe koopiana esinevate korduste arvud liigiti

Bakteritüvi	Korduste pikkus	Korduste arv
Mycoplasma genitalium G-37	200	8
	300	3
	400	1
	500	1
Chlamydia trachomatis D/UW-3/CX	200	1
	500	3
Neisseria gonorrhoeae FA 1090	200	12
	300	10
	400	8
	500	67
Helicobacter pylori 26695	200	7
	300	5
	400	6
	500	76
Treponema pallidum subsp. pallidum str. Nichols	200	8
	300	4
	400	3
	500	38

Kordusjärjestuste asend geenide suhtes. Vaadeldi ka kordusjärjestuste asendit valke kodeerivate ja RNA geenide suhtes. Uuriti iga kordusjärjestuse korral mitu antud kordusjärjestuse koopiat olid geenidega ülekattes ning mitmel kordusjärjestusel oli vähemalt 1 koopia teatud geeniga ülekattes. Viimase kohta on toodud andmed TABELIS 4.

TABEL 4. Korduste asend valku kodeerivate ja RNA geenide suhtes.

Bakteritüvi	**Geeni tüüp**	**Ülekattes kordused/ korduste koguarv*, protsentides**
Mycoplasma genitalium G-37	valku kodeeriv	48.39
	rRNA-d kodeeriv	6.90
Chlamydia trachomatis D/UW-3/CX	valku kodeeriv	0.00
	rRNA-d kodeeriv	81.80
Neisseria gonorrhoeae FA 1090	valku kodeeriv	83.67
	rRNA-d kodeeriv	2.04
Helicobacter pylori 26695	valku kodeeriv	82.80
	rRNA-d kodeeriv	6.90
Treponema pallidum subsp. pallidum str. Nichols	valku kodeeriv	53.80
	rRNA-d kodeeriv	48.70

*korduste arvu, millel vähemalt üks koopia teatud geeniga ülekattes, suhe kõigi leitud korduste arvu;

vastavate geenide koordinaadid on alla laetud NCBI andmebaasi GenBank ftp-serverist

3. Praimerite valik eksperimentaalseteks katseteks

3.1. Kandidaatpraimerite disain.

Praimerite disainimiseks valiti kolm (*N. gonorrhoeae, M. genitalium, H. pylori*) genoomi viie genoomi hulgast, kellele kordusjärjestused olid leitud. Genoomid valiti vastavalt leitud kordusjärjestuste arvule ning vastavalt sellele, kui huvi pakkuvad antud bakteritüved diagnostiliste uuringute seisukohast on.

Igale kordusjärjestusele disainiti ülimalt 10 praimeripaari programmiga PRIMER3. Osadele kordustele vähem juhul, kui kordusjärjestuse konsensusjärjestus ei sisaldanud piisavalt pikka mitte-redundantseid järjestikuseid nukleotiide sisaldavat ala ehk korduste vahel polnud piisavalt identset ala. Vähem kui 10 praimeripaari võidi praimereid disainida ka sõltuvalt järjestuse nukleotiidsest koostisest tulenevalt, st ei suudetud disainida praimeripaare, mis vastaks etteantud PRIMER3 kriteeriumitele. PRIMER3 kriteeriume muudeti minimaalselt, et poleks hilisema analüüsi käigus vaja arvestada PRIMER3 poolt vaadatavate tunnustega ning antud

analüüsi tulemusena tekkiva mudeli abil oleks kasutajatel, kes disainivad PRIMER3-e abil praimereid selle vaikimisi väärtusi kasutades võimalus jõuda mudeli koostajate poolt ennustatud tulemusele; eeldame, et PRIMER3 eemaldab vaatluse alt sellised praimerid, mis võiks teatud viisil endaga või teise praimeriga paarist või produktiga interkateeruda mitte-oodatud viisil. Programmi PRIMER3 muudetud parameetrid on järgnevad: produkti pikkuse vahemikud (baseerudes kordusjärjestuste pikkustele), praimeri pikkuste vahemikud, praimerite sulamistemperatuuride vahemikud, kasutatav monovalentsete katioonide kontsentratsioon, oligote kontsentratsioon.

Rohkema kui ühe praimeripaari disainimine igale kordusjärjestusele on vajalik selleks, et eksperimentaalseteks katsetusteks oleks meil piisavalt palju erinevate omadustega kandidaatpraimereid. Kokku disainiti 1796 erinevat kandidaatpraimeri paari. Ülevaade disainitud praimeritest on toodud TABELIS 5.

TABEL5. Uuritavatele organismidele disainitud praimerite arvud ning võimalikke valeprodukte genereerivate praimeripaaride arvud.

	N. gonorrhoeae	*M. genitalium*	*H. pylori*
Korduste erinevad koopiate arvud	1,2,3,4,5,6, 7,8,15,16,18	1,2,4	1,2,3,5
Disainitud praimeripaaride koguarv	686	120	990
Erinevate vähemalt kahe koopiaga korduste arv, millele õnnestus praimereid disainida (erinevate vähemalt kahe koopiaga korduste arv, millele ei õnnestunud praimereid disainida)	57(41)	7 (1)	81(13)
Praimeripaaride arv, mis võivad genereerida valeprodukte	85	25 (sihtmärk-genoomist)	55 (sihtmärk-genoomist)
Korduste arv (ka 1 koopiaga), millele disainitud praimerid võivad genereerida valeprodukte	13	4	23
Erinevate praimerite arv, mis genereerivad valeprodukti *N. meningitidis*est (korduste arv, millele disainitud praimerid annavad valeprodukti)	50 (9)	-	
Erinevate praimerite arv, mis genereerivad valeprodukti *N. gonorrhoeae*st (korduste arv, millele disainitud praimerid annavad valeprodukti)	43 (7)		

3.2. Uuritavate tunnuste väärtuste arvutamine

Disainitud praimeritele leiti ning arvutati nende kvaliteeti iseloomustavad tunnused ja väärtused. Tunnuste iseloom ja arvutamise põhimõtted on toodud vastavalt antud sektsioonis alampunktis 1.2 ja eelmises sektsioonis alampunktis 5. Erinevate tunnuste arvutamine on erineva ressursi kuluga (aeg, operatiivmälu). Kõige enam võtab aega vabaenergia väärtustel põhinevate tunnuste välja arvutamine - GRUPP 10-11. Võrreldes GRUPPe 10 ja 11 omavahel võtab inimese genoomist termodünaamikal põhinevate tunnuste väärtuste välja arvutamine oluliselt kauem aega kui antud bakteri genoomist vastavate tunnuste välja arvutamine (kui arvutada vastavaid väärtusi ühe protsessina, siis bakterite tunnuste välja arvutamise aega võib mõõta päevades, samal ajal kui inimesel kuudes). Sellest tulenevalt ollakse huvitatud sellest, kas on tarvilik termodünaamikal põhinevaid tunnuseid PCR-i tulemuste ennustamiseks

arvutada; kas on tarvilik vastavate tunnuste väärtuste arvutamine inimese genoomist.

3.3. Katses kasutatavate praimerite valik.

Eksperimentaalseteks katsetusteks oli võimalik valida 200 erinevat praimerit. Kuna katseks valitavate praimerite arv oli piiratud, siis püüti võimalike kandidaatpraimerite hulgast katseks valida võimalikult erinevate tunnuste väärtustega (võimalikult informatiivsed) praimeripaarid. Esmalt vähendati tunnuste arvu klasterdades kokku sarnast infot andvad tunnused. Selleks kasutati (vt 'ALGANDMED JA METOODIKA' alampunkt 6) k-keskmist klasterdamise meetodit. Praimeripaaridele välja arvutatud tunnuste klasterdamisel osutus mõttekaimaks (vt 'ALGANDMED JA METOODIKA') klastrite arvuks 12, seega vähendasime arvutatud tunnuste arvu 12ni. Leitud 12 klastri vastavate klastrite tsentritele kõige lähemad 12 tunnust peaksid klasterdamise eelduste kohaselt piisavalt korrektselt (st et hiljem oleks võimalik eksperimentaalsetest andmetest ja arvutatud-fikseeritud tunnustest leida statistiliselt olulist infot) kirjeldama meie andmestiku varieeruvust. Järgnevalt kasutasime saadud 12 tunnust kõigi 1796 kandidaatpraimeri paari klasterdamiseks.

Kuna täpsuse mõõtmiseks sobivaid alternatiivseid produkte genereerivaid praimeripaare oli vähe, siis vastavate praimerite valimiseks ei kasutatud eespool kasutatud statisilist meetodit, vaid valiti kõik potensiaalsed valeprodukti sünteesivad praimeripaarid eksperimentaalseteks katsetusteks (39 erinevat praimerit).

Kokkuvõttes disainitud 1796 kandidaatpraimeripaari hulgast valiti välja meetodiga k-keskmist 196 erinevat praimerit, millest moodustati 120 erinevat praimeripaari.

4. PCRi tulemuste intensiivsuste tasemeid mõjutavate tunnuste leidmine

4.1. Gruppide olulisemad esindajad.

Sõltumatute tunnuste hulgast peamiste tunnuste, mis mõjutavad kahte erinevat PCR-i tulemuste omadust (õnnstumine, tundlikkus), välja selgitamiseks kasutati logistilise regressiooni

analüüsi. Üksikute tunnuste mõju (mille Wald statistiku p-väärtus>0.0001) nimekiri gruppide kaupa on toodud lisas LISA 2; näidatud on kõigi tunnuste individuaalne mõju hii-ruut väärtuse kaudu (tunnused on gruppides sorditud hii-ruut väärtuse kahanemise suunas). Gruppide liikmed, milles peituv info kirjeldab ka teiste grupi liikmete mõju PCR-i tulemustele, on toodud TABELIS 6. Tabel on tagurpidi sammuhaavalist tunnuste valimismeetodit kasutava SAS protseduuri LOGISTIC gruppide kaupa lähenemise tulemus.

TABEL 6. Gruppide siseselt statistiliselt olulised (p<=0.0001) tunnused.

GNR	*tunnus (tundlikkus)*	*kirjeldus*
1	-	-
2	c_of_bacterial_DNA	bakteriaalse DNA kontsentratsioon
3	g_6_min_pr g_14_min_pr g_16_min_pr g_8_max_pr	praimeri 3' otsa minimaalsed ja maksimaalne deltaG väärtus vastavalt akendes 6,14,16 ja 8 nukleotiidi.
4	gc_min8_pr gc_max14_pr	praimeri 3' otsa minimaalne ja maksimaalne GC nukleotiidide sisaldus vastavalt akendes 8 ja 10 nukleotiidi.
5	gc_local_max_prod	maksimaalne lokaalne produkti GC nukleotiidide sisaldus
6	len_repeat	amplikoni pikkus
7	-	-
8	max12b3p max16b3	maksimaalsed praimeri seondumised loetuna praimeri 3' otsast 1,2 jne nukleotiidi edasi ja täpselt 3' otsast vastavalt 12 ja 16 nt-s aknas bakteris
9	max10h3 max14h3 max12h3p	maksimaalsed praimeri seondumised loetuna praimeri 3' otsast ja 3' otsast 1,2 jne nukleotiidi edasi vastavalt 10, 14 ja 12 nt-s aknas inimeses.
10	max_b3dgp_10 max_b3dg_14	maksimaalsed praimeri seondumised deltaG põhises aknas loetuna praimeri 3' otsast 1,2 jne nukleotiidi edasi ja täpselt 3' otsast vastavalt aknas 10 ja 14 nt bakteris (täpsemalt arvutamise metoodikast vt 'ALGANDMED JA METOODIKA')
11	min_dgregions_h	minimaalne deltaG põhine seondumiste arv inimeses (vt 'ALGANDMED JA METOODIKA')
12	-	-
13	prod10h	produktide arv arvestades praimerite 3' otsa 10 nt-seid seondumisi inimeses
14	sdss_b_min	minimaalsed SDSS põhised seondumised bakteris
15	sdss_h_min	minimaalsed SDSS põhised seondumised inimeses
16	diff_of_Ta_Tm_lower	madalama ennustatud sulamistemperatuuriga praimeri ja kasutatud PCR-i seondumistemperatuuri erinevus
17	-	-

4.2. Praimerite seostumiskohtade arvu mõju PCRi intensiivsuse tasemetele

(individuaalsete tunnuste mõju PCR-i intensiivsuse tasemetele vt LISA 2)

Praimerite seondumiste arvud bakteris (GRUPP8).

1. Akna pikkused. Vaadates tunnuseid individuaalselt ning erineva akna pikkusega praimerite 3' otste seondumiste mõju PCR-i intensiivsuse tasemetele, siis avaldavad statistiliselt olulist ($p<0.0001$) mõju peamiselt 12,14,16 aknaga mõõdetud seondumised, suurimaid Wald statistiku väärtusi omavad seondumised mõõdetuna 16-ses aknas (DF=1 korral). Võrreldes 8se ja 10se aknaga seondumiste arvude olulisust omavahel, omavad 8-se aknaga seondumiste arvud väiksemat mõju (kuigi statistiliselt olulist mõju $p>=0,0002$). Analüüsides antud grupi liikmeid koos, siis on näha, et näiteks ainult 16 nukleotiidi pikkuse aknaga praimeri 3' otsa seondumised üksi ei kirjelda seondumiste mõju PCR-i intensiivsuse tasemetele, teatud osa jääb kirjeldada ka 12 nukleotiidise aknaga loetud seondumistel (TABEL 7).

TABEL 7. PCR-i intensiivsuse tasemetele statistiliselt olulisi ($p<0.0001$) grupi 8 liikmeid kasutades saadud mudel. Mudeli c-statistiku väärtus 0.713, hälbimus/DF 0.9619

Parameeter*	***DF***	***Wald hii-ruut***	***p***
log10(max16b3)	1	34.6386	<0.0001
$max12b3p^2$	1	16.9267	<0.0001

*max16b3 - täpselt praimeri 3' otsa 16 bp seondumiste arv bakteris, max12b3p - praimeri 3' otsast alates 2sest, 3ndast jne nukleotiidist loetud seondumised bakteris, DF - vabadusastmete arv

2. Otsitava praimeri järjestuse asend praimeri 3' otsa suhtes antud aknas. Vaadates tunnuseid individuaalselt antud aknas, kas olulisemad on praimerite täpselt 3' otsa või 3' otsa 2-st, 3-st jne nukleotiidist loetud seondumised, siis reeglina on olulisemad (sama p väärtus, kuid kõrgem Wald statistiku väärtus DF=1 korral) seondumised loetuna täpselt 3' otsast. Vaadates antud grupi liikmeid koos selgub, et siiski on vajalik (täpsema mudeli saamiseks) ka praimeri 3' otsast kaugemate seondumiste arvu lugemine (TABEL 7).

Praimerite seondumiste arvud inimeses.(GRUPP9)

1. Akna pikkused. Analüüsides antud grupi statistiliselt usaldusväärseid ($p<0.0001$) tunnuseid individuaalselt osutuvad enamasti võrdväärselt olulisemateks praimerite 8, 10, 14 nukleotiidi pikkused seondumised. Kuigi ka 16 nukleotiidi pikkuste akende seondumised on statistiliselt olulised, siis võrreldes teiste akendega, on nad Wald statistiku (DF=1) väärtuse suhtes, kõige väiksema informatiivsusega. Analüüsides antud grupi liikmeid koos on võimalik kirjeldada ühe akna pikkusega praimerite seondumiste arvu mõju PCR-i intensiivsusele (TABEL 8).

TABEL 8. PCR-i intensiivsuse tasemetele statistiliselt olulisi ($p<0.0001$) grupi 9 liikmeid kasutades saadud mudel. Mudeli nr 1 c-statistiku väärtus 0.695, hälbimus/DF 0.9356, mudeli nr 2 c-statistiku väärtus 0.684, hälbimus/DF 0.9298

Mudeli nr	***Parameeter***	***DF***	***Wald hii-ruut***	***p***
1	max10h3^2	1	20.9195	<0.0001
1	max10h3p^2	1	21.8986	<0.0001
2	log10(max12h3p)^2	1	26.7040	<0.0001
2	max10h3^2	1	25.0355	<0.0001

* max10h3 ja max10(12)h3p vastavalt maksimaalne praimerite seondumiste arv lugedes praimeri täpselt 3' otsast ja 3' otsa 2st, 3ndast jne nukleotiidist aknas 10 (12) nukleotiidi, DF - vabadusastmete arv

2. Otsitava praimeri järjestuse asend praimeri 3' otsa suhtes antud aknas. Võrreldes antud aknas loendatud praimerite seondumiste maksimaalseid arve individuaalselt on sama vabadusastmete ja p väärtuse juures Wald hii-ruudu väärtuse järgi olulisemad praimerite seondumised loetuna täpselt 3' otsast. Vaadates antud grupi piires vastavaid seondumisi omavahel, osutub oluliseks vaadata praimerite maksimaalseid seondumisi lisaks täpselt 3' otsast loetutele ka praimeri 3' otsast edasi loetud seondumisi.

Praimerite maksimaalsed seondumiste arvud bakteris võrrelduna vastavate seondumistega inimeses. Võrreldes gruppi 8 ja 9 eraldi, tuleb ilmsiks, et bakterites on olulisemad praimeri

maksimaalsed seondumised kasutades pikemat akent, inimese genoomi korral on optimaalne vaadata seondumisi 10 nt akna korral. Nii inimese kui bakteri genoomi korral jääb teatud osa PCR-i intensiivsuse tasemest kirjeldada ka seondumiste arvul, mis on loetud praimeri 3' otsast 2, 3 jne nukleotiidi edasi. Analüüsides gruppi 8 ja 9 koos osutuvad oluliseks eelmiste punktidega kooskõlas olevalt praimerite maksimaalsed seondumised nii bakteris kui inimeses. Üks võimalik lahendus on toodud tabelis 9 (samaväärseid mudeleid on võimalik koostada kasutades teisi antud gruppide tunnuseid, mis sisaldavad endas võrdväärset infot).

TABEL 9. PCR-i intensiivsuse tasemetele grupi 8 ja 9 koosmõju. Mudeli c-indeks on 0.730, hälbimus/DF 0.9020.

Parameeter	***DF***	***Wald hii-ruut***	***p***
$max10h3p^2$	1	55.0599	<0.0001
$max12b3p^2$	1	38.1720	<0.0001

* max10h3p (max12b3p) - maksimaalsed praimeri 3' otsa 2st, 3st jne nukleotiidist loetud seondumised 10 (12) nt pikkuse aknaga loetuna inimeses (bakteris). DF - vabadusastmete arv.

Vabaenergia põhised seondumised bakteris. (GRUPP10)

1. Akna pikkus. Analüüsides erinevaid seondumiste arvu leidmiseks kasutatud akna pikkuseid on statistiliselt olulisemad lühemad akna pikkused (8,10 nt), kuigi ka pikemad akna pikkused on statistiliselt olulise p-väärtusega ($p<0.0001$). Vaadates akna pikkuseid grupisiseselt, siis on võimelised pikema aknaga loendatud seondumised kirjeldama PCR-i intensiivsuse taseme, mis on põhjustatud praimerite seondumistest bakteriaalses genoomis (TABEL 10).

2. Otsitava praimeri järjestuse asend praimeri 3' otsa suhtes antud aknas.

Vaadatades antud aknas praimerite seondumiste arve loendatuna praimeri 3' otsast ja 3' otsa 2st, 3ndast jne nukleotiidist ei saa teha järeldusi esimese või teise lähenemis eelistuse suunas. Grupisisesel analüüsil ilmneb, et samaväärne mudel on võimalik koostada, kui kasutada tunnuseid, mis loendavad praimerite seondumisi 3' otsast 2, 3 jne nukleotiidi edasi või, kui vastavaid tunnuseid mitte kasutada (TABEL 10, vastavalt mudelid 1 ja 2).

3. DeltaG põhiste seondumiste arvutamise metoodika

Võrreldes omavahel 'ALGANDMED JA METOODIKA' alampunktis 5 toodud erinevaid deltaG põhiseid seondumisi (so maxG+maxGp, max_dg+max_dgp, max_dgreg+max_dgregp, max_dg_tot, sum_dg_tot, max_dgregions+min_dgregions) individuaalselt, on olulisemad (võrreldes Wald hii-ruut väärtust sama p ja DF väärtuse korral) maxG-maxGp, max_dg-max_dgp, sum_dg_tot ning max_dg_tot põhised seondumised. Grupi siseseses analüüsis võistlevad PCR-i intensiivsuse kirjeldamises peamiselt max_dg(p), sum_dg_tot ja maxG(p), millest parem mudeli loovad kaks esimest tunnust.

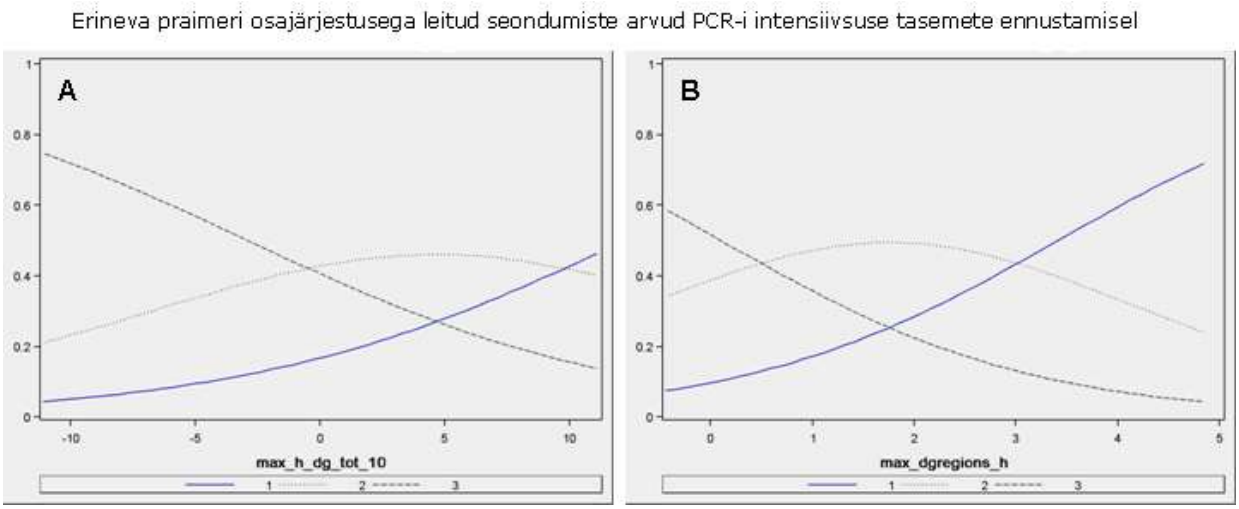
TABEL 10. Grupi 10 võimalikud mudelid PCR-i tundlikkuse tasemete kirjeldamiseks.

Mudeli nr	***Parameeter***	***DF***	***Wald hii-ruut***	***p***	***mudeli c-indeks; mudelihälbimus/DF***
1	max_b3dg_14^2	1	36.2181	<0.0001	0.749; 0.9051
1	sum_b_dg_tot_16^2	1	16.4130	<0.0001	
1	log10(max_b3dgp_10)	1	16.5057	<0.0001	
2	log10(sum_b_dg_tot_16)^2	1	19.2527	<0.0001	0.702; 0.9301
2	max_b3dg_14^2	1	43.4515	<0.0001	

*max_b3dg_14, sum_b_dg_tot_16, max_b3dgp_10 - vt 'ALGANDMED JA METOODIKA' alampunkt 5 vastavalt 10.3, 10.8, 10.4

Vabaenergia põhised seondumised inimeses. (GRUPP11)

1. Akna pikkus. Vaadates tunnuste individuaalseid mõjusid PCR-i intensiivsuse tasemetele avaldavad suuremat mõju (võrreldes Wald hii-ruut väärtusi, kui p<0.0001 ja DF=1) täispika ja pikema aknaga (16, 14 bp) leitud praimeri seondumiste arv (JOONIS 9, A,B).



JOONIS 9. Praimerite deltaG põhiste seondumiste arvu mõju PCR-i intensiivsuse hindamisele vaadates erinevaid praimeri osajärjestuste seondumiste arve. x-teljel max_h_dg_tot_10 ja max_dgregions_h vastavalt praimerite 10 nt pikkuse akna ja täispika praimeri seondumiste arvud on toodud sõltuvuse paremaks illustreerimiseks teisendatud andmetega. y-teljel on toodud PCR-i intensiivsuste tõenäosused, sinine joon tähistab madalat, punane keskmist ja roheline kõrget PCR-i intensiivsuse taset.

2. Otsitava praimeri järjestuse asend praimeri 3' otsa suhtes antud aknas. Vaadates seondumiste arve praimeri 3' otsa suhtes antud aknas avaldavad valdavas osas mõju (Wald hii-ruut väärtust jälgides, kui $p<0.0001$) praimerite seondumiste arvud, mis on loetud praimeri 3' otsast 2,3 jne nukleotiidi edasi. Üldiselt praimerite seondumised väiksema aknaga ja täpselt praimeri 3' otsast on statistiliselt ebaolulised.

3. DeltaG põhiste seondumiste arvutamise metoodika. Vaadates tunnuste individuaalset mõju PCR-i intensiivsuse tasemetele omavad suuremat tähtsust 'ALGANDMED JA METOODIKA' punktis 5 alampunktis 11.5 (max_dgregions, min_dgregions), 11.4 (sum_dg_tot), 11.3 (max_dg_tot) arvutatud tunnused. Võrreldes rühmade min_dgregions+max_dgregions, sum_dg_tot, max_dg_tot, max_dgp tunnuseid rühmasiseselt, osutuvad kõige informatiivsemateks rühmade max_dg_tot ja max_dgp liikmed (TABEL 11). Tabelis 11 toodud kolm esimest rühma on võrdväärsed, st liikmed ühest rühmast ei anna lisainformatsiooni juurde teisele rühmale PCR-i tundlikkuse tasemete ennustamisel.

TABEL 11. Erinevat deltaG põhist seondumiste leidmiste metoodikat kasutavate tunnuste mõju PCR-i intensiivsuse tasemetele

Mudeli (rühma) nimi	***Parameeter****	***DF***	***Wald hii-ruut***	***p***	***mudeli c-indeks; mudeli hälbimus/DF***
dgregions	$max\_dgregions\_h^2$	1	44.6256	<0.0001	0.625; 0.9753
max_dg_tot	$log10(max\_h\_dg\_tot\_14)^3$	1	22.7242	<0.0001	0.660; 0.9800
	log10(max_h_dg_tot_14)	1	14.1208	<0.0001	
max_dgp	log10(max_h3dgp_14)	1	20.9407	<0.0001	0.680; 0.9585
	$max\_h3dgp\_14^3$	1	34.7255	<0.0001	
sum_dg_tot	$log10(sum\_h\_dg\_tot\_14)^3$	1	31.0853	<0.0001	0.559; 0.9987
max_dg	$max\_h3dg\_14^3$	1	20.3359	<0.0001	0.590; 1.0153

* vt tekstist

Vabaenergia põhised seondumised bakteris ja inimeses.

PCR-i intensiivsuse taseme kirjeldavad kõige paremini tunnused, mis kätkevad endas seondumiste arve nii inimeses kui bakteris. Näiteks mudel, mis sisaldab kahte liiget - sum_h_dg_tot_10 ja max_b3dg_14 omab c-indeksit väärtusega 0.796, mis on suurem kui tabelis 10 ja 11 toodud vastavalt bakteri ja inimese genoomis leitud seondumiste arvule vastavad mudelid.

4.3. Teiste tunnuste mõju PCR-i tulemsusetele.

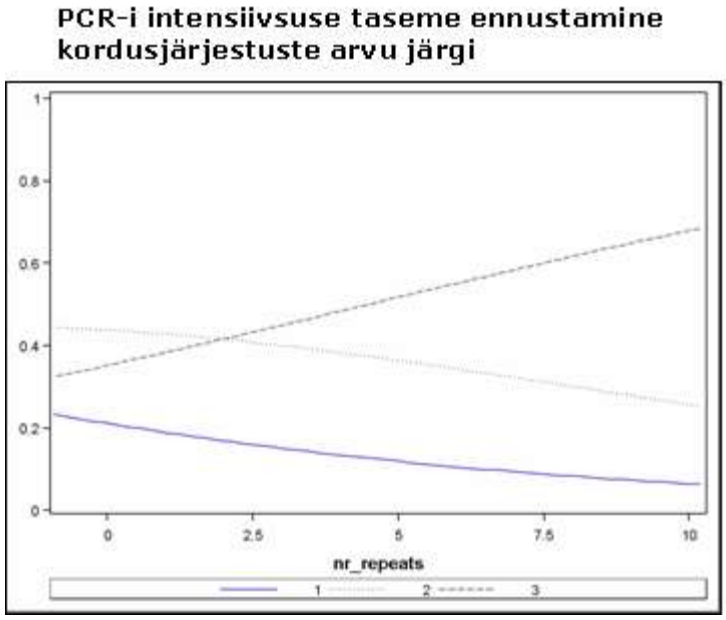
Siinkohal tuuakse ülevaade mõne uurijale huvipakkuvama grupi tähtsusest PCR-i sensitiivsuse tasemetele, ülejäänud tunnuste individuaalne mõju PCR-i intensiivsuse tasemtele on toodud Lisas 2.

Korduste arvu ja korduste pikkuse mõju PCR-i intensiivsuse tasemetele (GRUPP 6, 7)

Korduste arvul ja igal intensiivsuse tasemel eraldi seost ei ole - esimene ja teine intensiivsuse tase on ennustatavad ühesuguse tõenäosusega teatud korduste arvu korral ning kolmas

intensiivsuse tase on eristatav korduste arvu järgi iseseisvalt (JOONIS 10).

Korduste pikkusel on PCR-i intensiivsusele mõju - mida pikemad kordusjärjestused, seda kõrgem intensiivsus.

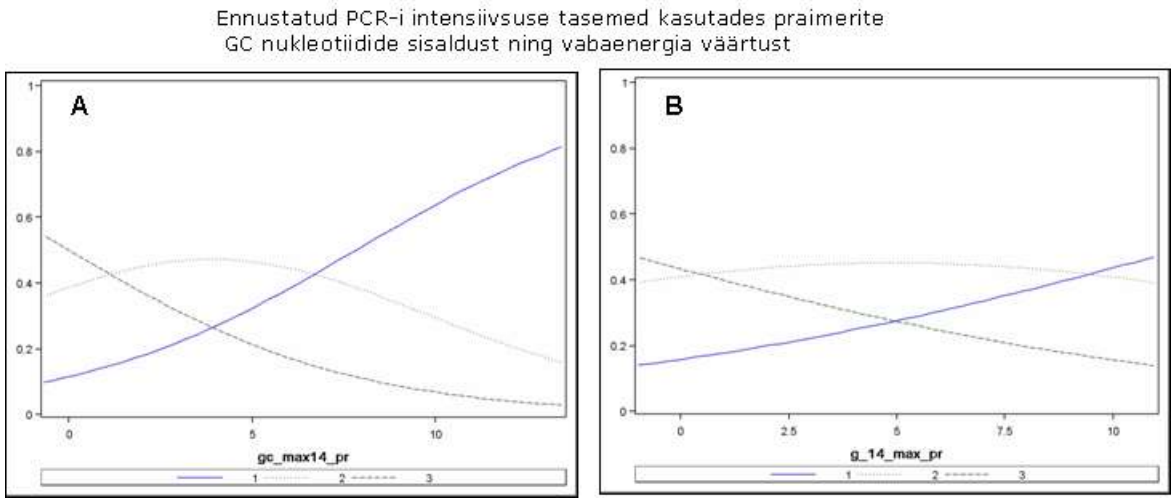


JOONIS 10. Korduste arvu mõju PCR-i intensiivsuse tasemetele.

Inimeses ja bakteris ennustatud produktide arvu mõju (GRUPP 12, 13)**.** Bakteris ennustatud produktide arv erineva pikkusega praimeri 3' otsa täieliku seondumise korral PCR-i intensiivsusele statistiliselt olulist mõju ei avalda. Viimane tuleneb eelkõige sellest, et praimerid disainiti liigispetsiifilistele kordusjärjestustele ja sellega eemaldati suures osas liigisisene võimalike mitte-spetsiifiliste produktide tekkimine. Ennustatud produktide arvud inimeses on olulised ($p<0.0001$) arvutatuna 8 ja 10 nukleotiidi pikkuses praimeri aknas.

Praimeri järjestuse mõju (GRUPP 3, 4)**.** Arvutatud deltaG ja GC sisaldusel põhinevatest praimeri tunnustest on mõlemad statistiliselt olulised. Praimerite deltaG väärtused on olulised lühema pikkusega akendes (6,8,10), pikemate akna pikkuste korral on olulisemad praimeri (praimeripaarist) minimaalsed väärtused (JOONIS 11), GC sisaldus on oluline pikemates praimeri 3' otsa akendes (16,14,12). Viimast tõenäoliselt seetõttu, et deltaG väärtus kirjeldab

lühemates akendes adekvaatsemalt praimeri 3' otsa stabiilsust (muutub samade nukleotiidide erineva järjestikuse asetsemise korral).



JOONIS 11. Praimerite 3' otsa 14 nukleotiidi pikkuse akna maksimaalne GC nukleotiidide sisaldus ja maksimaalne deltaG väärtus PCR-i intensiivsuse tasemete ennustamisel. Praimeri GC sisaldus ennustab PCR-i intensiivsuse taset paremini kasutades praimeri 3' otsa pikemat järjestust, deltaG väärtus ennustab intensiivsust paremini, kui on kasutatud praimeri 3' otsa lühemaid akna pikkuseid.

Produkte iseloomustavate tunnuste mõju (GRUPP5)

Antud grupist on omab kõrgeimat Wald statistiku väärtust ($p>0.0001$ korral) produkti lokaalne GC nukleotiidide sisaldus ning statistiliselt ebaoluline on üldine GC nukleotiidide sisaldus produktis.

1.5. PCRi praimerite intensiivsuse tasemeid ennustav mudel

PCR-i tulemuste intensiivsust ennustatakse juhul, kui edukust ennustav mudel on PCR-i õnnestumiseks andnud tõenäosuse, mis vastab statistiliselt usaldusväärselt PCR-i õnnestumisele, st intensiivsuse ennustamisel on eelduseks PCR-i õnnestumine. Modelleeriti PCR-i intensiivsust kolmel tasemel - 1 - nõrk, 2 - keskmine, 3 - tugev;

Tulemusena saadi järgnevad kolm regressioonijoone funktsiooni:

$$g_{1,2,3}(x)=\beta_0 + \beta_1(\text{gc\_local\_max\_prod}) + \beta_2(\text{max\_h3dgp\_14}) + \beta_3(\text{max12h3p})$$

kus g(x) on logit funktsioon antud intensiivsuse tekkimise tõenäosusest. Mudelis olevad

parameetrid ning nende kordajate ennustatud väärtused koos Wald hii-ruudu väärtuste ning vastava tõenäosusega on toodud tabelis 8.

TABEL 8. Analüüsi tulemusena saadud mudeli parameetrid ja parameetrite ennustatud väärtused

Parameeter		sümbol	Ennustus	Wald hii-ruut	p
Vabaliige	**1**	β_0	-3.7895	119.8253	<.0001
Vabaliige	**2**	β_0	-0.9194	13.4086	0.0003
log10(gc_local_max_prod)		β_1	-14.1211	75.5099	<.0001
log10(max_h3dgp_14)^3		β_2	0.0118	12.2685	0.0005
max12h3p^2		β_3	-0.3275	25.3807	<.0001

*p -Wald statistiku tõenäosus

Kasutades ennustatud parameetrite kordajate väärtusi saame järgnevad tõenäosusvalemid:

$$\text{logit}(p_1) = -3.7895 - 14.1211\beta_1 + 0.0118\beta_2 - 0.3275\beta_3$$

$$\text{logit}(p_1+p_2) = -0.9194 - 14.1211\beta_1 + 0.0118\beta_2 - 0.3275\beta_3,$$

kust p_3 avaldamine on ilmne ($p_3=p_1+p_2$).

Valminud mudeli hälbimus on 546.4287 (DF=677). Hälbimust kasutati erinevate mudelite võrdlemiseks. Mudeli c-statistiku väärtus on 0.819, mis näitab mudeli adekvaatsuse olemasolu, st mudel on võimeline eristama erinevaid intensiivsuse tasemeid.

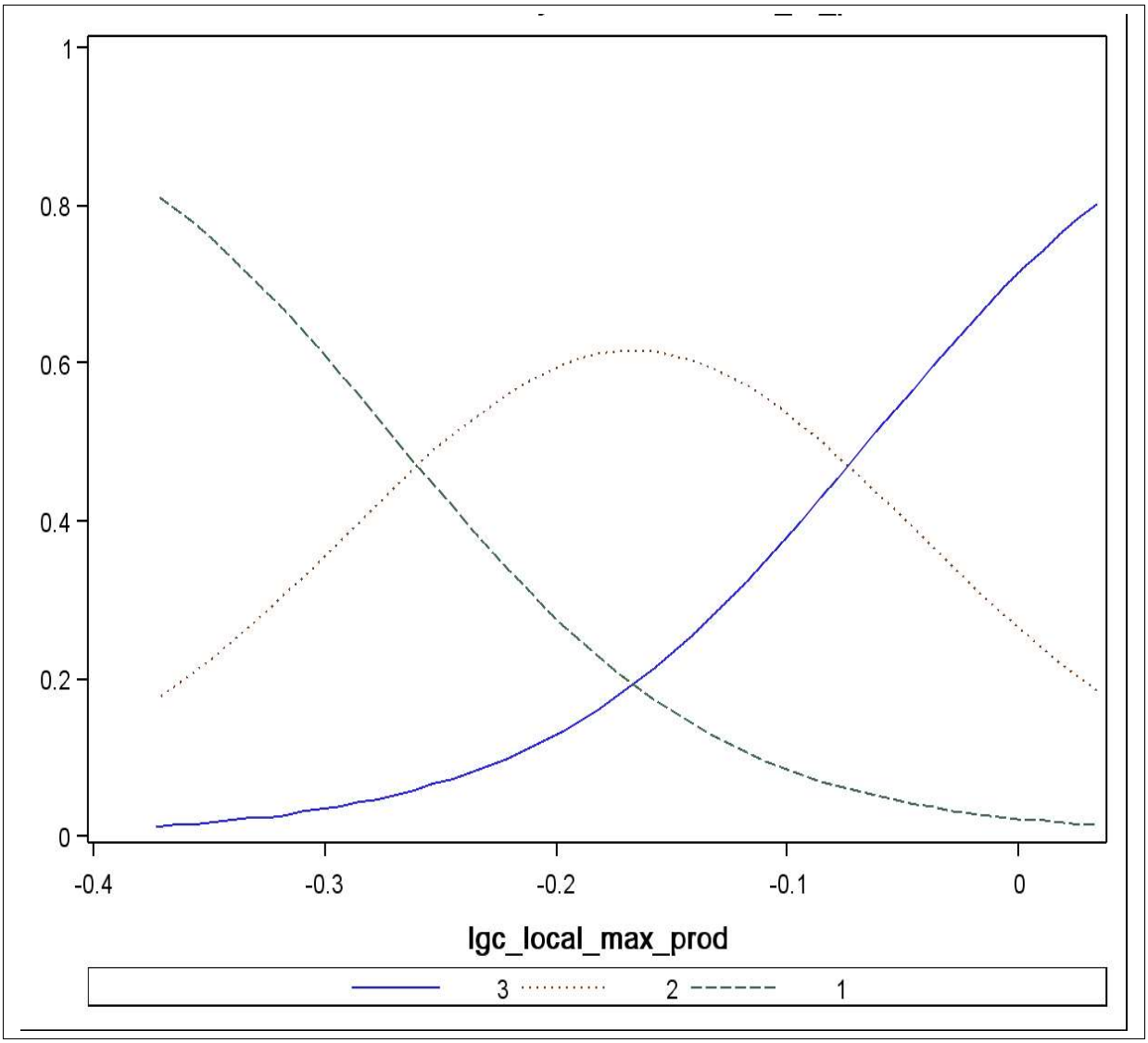
Mudeli parameetrid: 1. gc_local_max_prod - PCR-i produkti maksimaalne lokaalne GC nukelotiidide sisaldus (GRUPP5). Kui produktid omavad teatud piirkonnas kõrget GC nukleotiidide sisaldust, siis PCR-i vastavates tsüklites temperatuuri tõstes ja alandades produkti-dupleksid on üksteisega teatud regioonidest tugevalt interakteerunud ning ei sula lahti või kleepuvad kiiresti kokku vastusena temperatuuri tõstmisele-alandamisele.

2. max_h3dgp_14 -praimeri 3' otsa alates 2.,3. jne nukleotiidist 14 nukleotiidi pikkused

maksimaalsed seondumised deltaG põhises aknas inimese genoomis (GRUPP11). Praimerite nn kadu ning müra tekitamine pärsivad õige PCR-i produkti paljundamist.

3. max12h3p - maksimaalsed praimeri 3' otsa alates 2.,3. jne nukleotiidist 12 nukleotiidi pikkused maksimaalsed seondumised inimese genoomis (GRUPP9).

Näide mudeli prognoosijõust on toodud joonisel 9, kus on illustreeritud parameetri 'gc_local_max_prod' (abstsiss) mõju PCR-i õnnestumisele (oordinaat).



JOONIS 9. Koostatud mudeli poolt ennustatud kolme intensiivsuse (int_of_product) taseme tõenäosused. lgc_local_max_prod on produkti maksimaalne GC nukleotiidide lokaalne sisaldus (väärtus kümnendlogaritmi kujul). 3 - kõige tugevama intensiivsuse tase, 1 kõige madalam. Mida kõrgem on vastav GC nukleotiidide sisaldus, seda tugevam on intensiivsus. (y-teljel on toodud sündmuse tõenäosus, x-teljel kümnendlogaritmitud tunnuse lgc_local_max_prod väärtused)

Tähtis on mainida, et saadud valemid pole ainuvõimalikud ja võrreldava ennustusjõuga mudeleid on võimalik koostada kasutades selleks teisi meie poolt välja arvutatud tunnuseid. Samuti peab arvestama asjaoluga, et valemid on koostatud antud katsesüsteemi ja andmestikku kasutades, viimane on eriti tähtis just PCR-i edukuse hindamisel, kuna andmestik sisaldas valdavalt õnnestunud PCR-i tulemusi (andmestik pole piisavalt varieeruv).

ARUTELU

Käesoleva töö eesmärgiks on bakterite molekulaardiagnostikas kasutatavate PCR- i praimerite tundlikkuse taseme parandamine. Tundlikkuse taseme parandamiseks disainisime PCR-i praimerid liigispetsiifilistele kordusjärjestustele. Igale disainitud praimeripaarile arvutati välja erinevad skoorid, mis on ennustatavalt seotud PCR-i tulemuste intensiivsuse tasemtega. Kasutasime ülesande lahendamiseks statistilist lähenemist. On näidatud, et statistiliste mudeli kasutamine praimerite valimiseks eksperimentaalseteks katsetusteks parandab SNP-ide genotüpiseerimise edukuse taset (86, 87). Kasutades logistililise regressiooni analüüsi leidsime kõigi arvutatud tunnuste mõju PCR-i intensiivsuse tasemetele. Suurt mõju PCR-i intensiivsuse tasemele avaldab maksimaalne lokaalne GC nukleotiidide sisaldus PCR-i produktis (Wald hii-ruut 87.3257, $p<0.0001$), mis on arvutatud kui maksimaalne GC nukleotiidide sisalduse tase kasutatud praimerite pikkusest väiksema pikkusega praimeri pikkusest alates kuni produkti pikkuseni. Asjaolu võib olla seletatav sellega, et sünteesitavad PCR-i produktid on praimeriekstensiooni faasis tugevamalt koos ning ei põhjusta praimerite amplikonilt ‚lahti hüppamisi' produktidupleksite ebastabiilsuse tõttu.

PCR-i intensiivsuste tasemega regressioonis on ka praimerite seondumiste arvud loendatuna nii inimese kui ka bakteri genoomis kasutades selleks nukleotiid-nukleotiid kui ka deltaG põhist meetodit. Nagu näha ‚TULEMUSED' alampunktis 4.2 toodud analüüsis, on PCR-i intensiivsuse tasemetele teatud määral olulisemad nukleotiid-nukleotiid seondumiste arvud loendatuna inimeses, kuid teatud osa nukleotiid-nukleotiid seondumiste varieeruvusest jääb kirjeldada ka bakteriaalses genoomis loendatud seondumiste arvul. Sama kehtib ka deltaG põhiste seondumiste analüüsimisel. Üldiselt võib tõdeda, et suhteliselt usaldusväärne (antud statistilise analüüsi tulemusena) PCR-i intensiivsuse tasemete ennustus on võimalik saavutada kasutades vaid inimese genoomis leitud seondumiste arvu. Loomulikult saavutatakse suurem ennustusjõud kasutades selleks rohkem erinevaid informatiivseid tunnuseid, kuid sellega

kaotatakse PCR-i intensiivsuse taset ennustavat valemit realiseeriva programmi läbilaskevõimes.

PCR-i intensiivsuse tasemete ja korduste arvu vahel ei saa tuua välja statistiliselt olulist seost. Välja saab tuua seose PCR-i kõrgeima (3nda) intensiivsuse taseme ja PCR-i intensiivsuse tasemete vahel (JOONIS 7). Alates ca 6 koopiaga korduste kasutamisest saavutatakse kolmanda PCR-i intensiivsuse taseme saamise kõrgem tõenäosus võrreldes kahe ülejäänud tasemega. Võimalik, et väiksema arvu koopiatega korduste kasutamisel pole intensiivsuse tase võrreldes ühe koopialiste kordusega (st mitte-kordustega) visuaalselt eristatav.

Oluline on mõista, et meie poolt esitatud valem on vaid üks võimalik lahendus PCR-i intensiivsuse taseme ennustamiseks, ennustusjõu suhtes võrdväärseid valemeid on võimalik luua kasutades selleks erinevaid tunnuste komplekte. Vajalik on meie poolt välja töötatud parameetrite individuaalse statistilise olulisuse teadmine. Meie lähtusime mudeli loomisel peamiselt kahest põhimõttest – valem peab sisaldama võimalikult vähe tunnuseid ning võimalikult väikese arvutusmahuga leitavaid tunnuseid. Võimalikult väheste tunnuste kaasamine mudelisse on esiteks kasutajale usaldusväärsem ning vastuvõtlikum, lisaks kaasates mudelisse liialt palju tunnuseid võib see suurendada mudeli ebastabiilsust, st mudel on liialt kohandatud meie poolt kasutatud treeningandmestikule. Teise tingimuse vajalikkus on ilmne.

Mudeli kasutamisel on tarvilik arvestada ka sellega, et mudeli välja töötamiseks on kasutatud vaid ühe eksperimendiseerija tarvis tehtud katsete andmeid. Viimane loob võimaluse, et leitud mudel on antud katseprotokolli spetsiifiline või katsete läbiviimise laboratooriumi poolt põhjustatud ‚müra' (arvutuslikult võimatu ennustada) spetsiifiline. Viimase probleemi vähendamiseks planeeritakse uut korduskatsete seeriat. Võimalik, et pole võimalik universaalselt ennustada järjestusest sõltumatuid faktoreid, mis mõjutavad PCR-i intensiivsuse tasemeid, st vajalik on iga labori tarvis mõne võrra teistsuguse mudeli välja töötamine. Samuti peab märkima, et eksperimentides kasutatud praimerid on disainitud antud praimeridisaini programmiga, mis juba eelnevalt eemaldas kandidaatpraimerite hulgast teatud kriteeriumitele mitte vastavad praimeripaarid.

Käesoleva töö raames eksperimentaalsetest katsetest kogutud andemeid kasutatakse edaspidi kahe uue mudeli välja töötamiseks. Tahame leida valemeid, mis kirjeldavad PCR-i edukust üldiselt (bändi olemasolu) ja alternatiivsete produktide tekkimise asjaolusid. Seejärel arendatakse välja summaarne tarkvara, mis kasutab antud töö raames välja töötatud praimerite disainimise metoodikat liigispetsiifilistele kordusjärjestustele ning ennustab disainitud praimeripaarile PCR-i õnnestumise, PCR-i intensiivsuse taseme ning PCR-i täpsuse (valebändi olemasolu) tõenäosused.

KOKKUVÕTE

Käesoleva töö raames leidsime matemaatilise mudeli bakterite molekuaardiagnostikas kasutatavate PCR-i praimerite intensiivsuse taseme parandamiseks. Mudeli välja töötamiseks viidi läbi eksperimentaalsed katsed meie poolt disainitud praimeripaaridega.

1. Leidsime liigispetsiifilised kordusjärjestused viiele bakteriaalsele patogeenile (*Neisseria gonorrhoeae FA 1090, Helicobacter pylori 26695, Mycoplasma genitalium G37, Treponema pallidum subsp. pallidum str. Nichols, Chlamydia trachomatis D/UW-3/CX*), kellest kolmele esimesele disainisime kokku 1796 PCR-i kandidaatpraimeripaari.

2. Kasutades statistilist meetodit k-keskmist tunnuste klasterdamiseks valisime kandidaatpraimerite hulgast välja laboratooriumis testitimiseks 120 erinevat praimeripaari (196 erinevat praimerit).

3. Leidsime 145 tunnuse väärtused, mis on ennustuslikult seotud PCR-i tulemuste intensiivsuse tasemetega ja arvutasime need igale disainitud praimeripaarile. Iga katsetatud praimeripaari kohta märgiti PCR-i tulemuste intensiivuse tase asjatundjate poolt.

4. Kasutades logistilise regressiooni analüüsi leidsime iga individuaalse tunnuse jaoks tema statistilise olulisuse PCR-i tasemete intensiivuse määramisel Wald hii-ruut statistiku ja vastava p-väärtuse näol. Suure arvu tunnuste (kokku 145 + 12 eksperimentide käigus tekkinud tunnust) ning tunnuste vahelise tugeva multikollineaarsuse tõttu polnud võimalik kõiki tunnuseid mudelisse koos lisades leida usaldusväärset summaarset PCR-i intensiivsusi ennustavat valemit. Sarnase iseloomuga tunnused grupeeriti kokku ning leiti igast grupist kõige olulisemad tunnused, misjärel koostati leitud tunnustest summaarne mudel kasutades tagurpidi sammuhaavalist logistilise regressiooni analüüsi. Koostatud mudeli headust näitava c-statistiku väärtus on 0.819. Leitud matemaatiline valem sisaldab kolme parameetrit, mis on seotud produkti GC nukleotiidide maksimaalse teatud regiooni sisalduse tasemega, nukleotiid-nukleotiid seondumiste ja deltaG põhiste seondumistega inimeses.

SUMMARY

Detection of pathogenic bacteria from the clinical samples is continually relevant problem. One of the most exerted methods for detecting specific bacterial infections is the polymerase chain reaction (PCR). The quality of the results of PCR depends significantly on specificity and sensitivity of primers used in the reaction mixture.

We have developed the mathematical model for predicting the levels of intensivity of PCR from bacterial genomes. To evolve the empirical formula:

1. we have developed the BLAST-based method for finding species-specific bacterial repetitive sequences. By using this methodology we have found repetitive sequences for five bacterial pathogenic genomes (*Neisseria gonorrhoeae FA 1090, Helicobacter pylori 26695, Mycoplasma genitalium G37, Treponema pallidum subsp. pallidum str. Nichols, Chlamydia trachomatis D/UW-3/CX*).
2. We have designed primers for the repetitive sequences of three bacterial genomes (*N. Gonorrhoeae, H. Pylori, M. genitalium*). We have composed the training set with 1796 primer pairs of PCR. With statistical clustering method *k-means* we have selected 120 primer pairs for experimental tests.
3. we have developed a set of parameters (145) assumed to be correlated with the levels of sensitivity of PCR and calculated the values for every designed primer pair. The level of sensitivity of experimentally tested primer pair was marked by specialist.
4. we have applied a logistic regression analysis to combine all scored parameters into one measure predicting the overall level of sensitivity of a given primer pair of PCR. The c-statistic achieved for the model for estimateing the levels of intensivity of PCR is 0.819.

This new statistical prediction can be used to improve primer design for the identification of bacterial genomes and to evaluate the probability of levels of sensitivity of PCR.

TÄNUAVALDUSED

Südamlik tänu oma juhendaja professor Maido Remmile, kes aitas saavutada käesoleva töö positiivsed küljed ning tänu, kelle põhjalikkusele ja laiale silmaringile olen omandanud palju uusi teadmisi ning õppinud probleemide üle detailsemalt juurdlema.

Samuti tänan meie bioinformaatika töögrupi statistikut Tõnu Mölsi, kes aitas mul teha esimesi samme statistilise analüüsi valdkonnas.

Suured tänud ka eksperimentaalsete katsete läbiviijatele.

BIBLIOGRAAFIA

1. Mhlanga MM, Malmberg L (2001). Using Molecular Beacons to Detect Single-Nucleotide Polymorphisms with Real-Time PCR.*METHODS* 25 (4), 463–71.
2. Chiueh L-C, Chen Y-L, Yu J-H and Shih D. Y-C. (2001). Detection of Four Types of Genetically Modified Maize by Polymerase Chain Reaction and Immuno-Kit Methods. *Journal of Food and Drug Analyses*. 9(1), 50-7.
3. Nugent K. G, Saville B. J. (2003). Forensic analysis of hallucinogenic fungi: a DNA-based approach. *Forensic science international*. 140(2), 147-57.
4. Yancy HF, Mohla A, Farell DE, Myers MJ. (2005). Evaluation of a rapid PCR-based method for the detection of animal material. *Journal of food protection*. 68(12), 2651-5.
5. Budowle B, Bieber FR, Eisenberg AJ. (2005). Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Legal medicine*. 7(4), 230-43.
6. Zsikla V, Hailemariam S, Baumann M, Mund MT, Schaub N, Meier R, Cathomas G. (2006). Increased Rate of Helicobacter pylori Infection Detected by PCR in Biopsies With Chronic Gastritis. *The American journal of surgical pathology.* 30(2), 242-8.
7. Banks JT, Bharara S, Tubbs RS, Wolff CL, Gillespie GY, Markert JM, Blount J. (2005) Polymerase chain reaction for the rapid detection of cerebrospinal fluid shunt or ventriculostomy infections. *Neurosurgery*. 57(6), 1237-43.
8. Bidochka MJ, McDonald MA, St Leger RJ, Roberts DW. (1994). Differentiation of species and strains of entomopathogenic fungi by random amplification of polymorphic DNA (RAPD). *Curr Genet.* 25(2), 107-13.
9. Rogic S, Mackworth AK, Ouellette FB. (2001). Evaluation of gene-finding programs on mammalian sequences. *Genome Res*. 11(5), 817-32.
10. Bedell JA, Korf I, Gish W. (2000). MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics.*16(11), 1040-1.
11. Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott, MP, Zipursky L, Darnell J. (2003). Molecualr Cell Biology. 5th Edition. Chapter 9, 375-6.
12. Arezi B, Xing W, Sorge JA, Hogrefe HH. (2003). Amplification efficiency of thermostable DNA polymerases. *Analytical Biochemistry*. 321(2), 226-35.
13. Poland D, Scheraga HA. (1970). Theory of Helix-Coil Transitions in Biopolymers. Academic Press, New York.
14. Pavlov AR, Pavlova NV, Kozyavkin SA, Slesarev AI. (2004). Recent developments in the optimization of thermostable DNA polymerases for efficient applications. *Trends in Biotechnology.* 22(5), 253-60.

15. Viguera E, Canceill D, Ehrlich SD. (2001) In vitro replication slippage by DNA polymerases from thermophilic organisms. *Journal of Molecular Biology*. 312(2), 323-33.

16. Ollis D, Brick P, Hamlin R, Xuong N and Steitz T. (1985) Structure of the large fragment of Escherichia coli DNA polymerase I complexed with dTMP. *Nature*. 313(), 762–6.

17. Andricioaei I, Goel A, Herschbach D, Karplus M. (2004). Dependence of DNA Polymerase Replication Rate on External Forces: A Model Based on Molecular Dynamics Simulations. *The Biophysical Society*. 87(3), 1478-97.

18. Li Y, Korolev S,Waksman G. (1998). Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of Thermus aquaticus DNA polymerase I: structural basis for nucleotide incorporation. *The EMBO journal.* 17(24), 7514-25.

19. Turner RM Jr, Grindley ND, Joyce CM. (2003) Interaction of DNA polymerase I (Klenow fragment) with the single-stranded template beyond the site of synthesis. *Biochemsitry*. 42 (8), 2373-85.

20. Cline J, Braman JC, Hogrefe HH. (1996). PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* 24(18), 3546-51.

21. http://www.dnasoftware.com/Products/VisualOMP/CaseStudies/PCR/index.htm, mai 2006

22. Eckert KA, Kunkel TA. (1990). High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acid Res.* 18(13), 3739-44.

23. Judo MS, Wedel AB, Wilson C. (1998). Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res.* 26(7), 1819-25.

24.Summerer D, Marx A. (2004). 4' C-Modified Nucleotides as Chemical Tools for Investigation and Modulation of DNA Polymerase Function. *Synlett.* 2004(02), 217-24.

25. Jung KH, Marx A. (2005). Nucleotide analogues as probes for DNA polymerases. *CMLS.* 62(18), 2080-91.

26. Li Y, Waksman G. (2001). Crystal structures of a ddATP-, ddTTP-, ddCTP-, and ddGTP-trapped ternary complex of Klentaq1: Insights into nucleotide incorporation and selectivity. *Protein Sci.* 10(6), 1225-33.

27. Kunkel TA, Bebenek K. (2000). DNA replication fidelity. *Annu Rev Biochem.* 69, 497-529.

28. Pavlov AR, Pavlova NV, Kozyavkin SA, Slesarev AI. (2004). Recent developments in the optimization of thermostable DNA polymerases for efficient applications. *Trends in Biotechnology*. 22(5), 253-60.

29. Shinde D, Lai Y, Sun F, Arnheim N. (2003). Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.* 31(3), 974-80.

30. Kohandel M, Ha BY. (2006). Thermal denaturation of double-stranded DNA: Effect of base stacking. *Phys Rev E Stat Nonlin Soft Matter Phys.* 73(1), 011905-1- 011905-8.

31.Friedman RA, Honig B. (1995). A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophys J.* 69(4), 1528-35.

32. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Res.* 34 (2), 564-74.

33. Norberg J, Nilsson L. (1998). Solvent influense on base stacking. *Biophys J.* 71(1), 394-402.

34. Kool ET. (1997). Preorganization of DNA: Design Principles for Improving Nucleic Acid Recognition by Synthetic Oligonucleotides. *Chem Rev*. 97(5), 1473-88.

35. Kool ET. (2001). Hydrogen bonding, base stacking, and steric effects in dna replication. *Annu Rev Biophys Biomol Struct*. 30, 1-22.

36. SantaLucia J Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci*. 95, 1460–65.

37.Watson JD and Crick FHC. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature*. 171, 737-8.

38. Breslauer KJ, Frank R, Blocker H, Marky LA. (1986). Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A.* 83(11), 3746-50.

39. Owczarzy R, Vallone PM, Gallo FJ, Paner TM, Lane MJ, Benight AS. (1997). Predicting sequence-dependent melting stability of short duplex DNA oligomers. *Biopolymers*. 44(3), 217-39.

40. Ivanov V, Piontkovski D, Zocchi G. (2005). Local cooperativity mechanism in the DNA melting transition. *Phys Rev E Stat Nonlin Soft Matter Phys.* 71(4), 041909-(1-8).

41.Sundaralingam M, Ponnuswamy P.K. (2004). Stability of DNA Duplexes with Watson-Crick Base Pairs: A Predicted Model. *Biochemistry*. 43(51), 16467-76.

42. Sugimoto N, Nakano S, Yoneyama M, Honda K. (1996). Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.* 24(22), 4501-5.

43. SantaLucia J Jr, Allawi HT, Seneviratne PA. (1996). Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry*. 35(11), 3555-62.

44. SantaLucia J Jr, Hicks D. (2004). The Thermodynamicsof DNA Structural Motifs. *Annu Rev Biophys Biomol Struct.* 33, 415-40.

45. Bommarito S, Peyret N, SantaLucia J Jr. (2000). Thermodynamic parameters for DNA

sequences with dangling ends. *Nucleic Acids Res.* 28(9), 1929-34.

46. Rychlik W, Spencer WJ, Rhoads RE. (1990). Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.* 18(21): 6409–6412.

47. Baldino F Jr, Chesselet MF, Lewis ME. (1989). High-resolution in situ hybridization histochemistry. *Methods Enzymol.* 168, 761-77.

48. Wallace RB, Shaffer J, Murphy RF, Bonner J, Hirose T, Itakura K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res.* 6(11), 3543–3557.

49. Bicout DJ, Kats E. (2004). Bubble relaxation dynamics in double-stranded DNA. *Phys Rev E Stat Nonlin Soft Matter Phys.* 70(1), 010902.

50. Tikhomirova A, Taulier N, Chalikian TV. (2004). Energetics of Nucleic Acid Stability: The Effect of *C*P. *J Am Chem Soc.* 126(50), 16387-94.

51. Ahsen von N, Wittwer CT, Schutz E. (2001). Oligonucleotide melting temperatures under PCR conditions: nearest-neighbor corrections for Mg(2+), deoxynucleotide triphosphate, and dimethyl sulfoxide concentrations with comparison to alternative empirical formulas. *Clin Chem.* 47(11), 1956-61.

52.Thomas R. (1954). Investigations into denaturation of DNA. *Biochim Biophys Acta* 14, 231-240.

53. Nakano S, Fujimoto M, Hara H, Sugimoto N. (1999) Nucleic acid duplex stability: influence of base composition on cation effects. *Nucleic Acids Res.* 27(14), 2957-65.

54. Cheng Y, Korolev N, Nordenskiöld L. (2006). Similarities and differences in interaction of K+ and Na+ with condensed ordered DNA. A molecular dynamics computer simulation study. *Nucleic Acids Res.* 34(2), 686-96.

55. Lyubartsev AP, Laaksonen A. (1998). Molecular dynamics simulations of DNA in solutions with different counter-ions. *J Biomol Struct Dyn.* 16(3), 579-92.

56.Mocci F, Saba G. (2003). Molecular dynamics simulations of A. T-rich oligomers: sequence-specific binding of Na+ in the minor groove of B-DNA. *Biopolymers.* 68(4), 471-85.

57.Owczarzy R, You Y, Moreira BG, Manthey JA, Huang L, Behlke MA, Walder JA. (2004) Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures. *Biochemistry.* 43(12), 3537-54.

58.Schildkraut, C., and Lifson, S. (1965) Dependence of the melting temperature of DNA on salt concentration, *Biopolymers.* 3, 195-208.

59.Andreson R, Reppo E, Kaplinski L, Remm M. (2006). GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics.* 7(172).

60. Versalovic J, Lupski JR. (2002). Molecular detection and genotyping of pathogens: more accurate and rapid answers. *Trends Microbiol.* 10(10), S15-21.
61. D'Aquila RT, Bechtel LJ, Videler JA, Eron JJ, Gorczyca P, Kaplan JC. (1991). Maximizing sensitivity and specificity of PCR by pre-amplification heating. *Nucleic Acids Res.*19(13), 3749.
62. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS. (1991). 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res.* 19(14), 4008.
63. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol.* 215(3), 403-10.
64. Lin B, Wang Z, Vora GJ, Thornton JA, Schnur JM, Thach DC, Blaney KM, Ligler AG, Malanoski AP, Santiago J, Walter EA, Agan BK, Metzgar D, Seto D, Daum LT, Kruzelock R, Rowley RK, Hanson EH, Tibbets C, Stenger DA. (2006). Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome Res.* 16:527-35.
65. An L, Tang W, Ranalli TA, Kim HJ, Wytiaz J, Kong H. (2005). Characterization of a Thermostable UvrD Helicase and Its Participation in Helicase-dependent Amplification. *Journal of Biological Chemistry.* 280(32), 28952-8.
66. Brandt O, Hoheisel JD. (2004) Peptide nucleic acids on microarrays and other biosensors. *Trends in Biotechnology.* 22(12), 617-622.
67. Csako G. (2006) Present and future of rapid and/or high-throughput methods for nucleic acid testing. *Clinica Chimica Acta.* 363:6-31.
68. Zhao R. (2005). From Single Cell Gene-based Diagnostics to Diagnostic Genomics: Current Applications and Future Perspectives. *Clinical Laboratory Science.*
69. Fortina P, Kricka LJ, Surrey S, Grodzinski P. (2005). Nanobiotechnology: the promise and reality of new approaches to molecular recognition. *Trends in Biotechnology.* 23(4), 168-173.
70. Holland CA, Kiechle FL. (2005). Point-of-care molecular diagnostic systems — past, present and future. *Current Opinion in Microbiology.* 8:504-509.
71. Mothershed EA, Whitney AM. (2006). Nucleic acid-based methods for the detection of bacterial pathogens: Present and future considerations for the clinical laboratory. *Clinica Chimica Acta.* 363:206-220.
72. Vincent M, Xu Y, Kong H. (2004) Helicase-dependent isothermal DNA amplification. *EMBO Reports.* 5(8), 795-800.
73. Rozen S, Skaletsky H. (1998). Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.
74. Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22(22), 4673-80.
75. Zuker M, Mathews DH, Turner DH. (1999) Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology.

J. Barciszewski and B.F.C. Clark, eds., *NATO ASI Series, Kluwer Academic Publishers*.

76. Miura F, Uematsu C, Sakaki Y, Ito T. (2005). A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. 21(24), 4363-70

77. Johnson NL. (1949). Systems of frequency curves generated by methods of translations. *Biometrika.* 36:149-76.

78. Slifker JF, Shapiro SS. (1980). The Johnson System: Selection and Parameter Estimation. *Technometrics.* 22(2), 239-246.

79. Quinn GP, Keough MJ. (2005). Experimental Design and Data Analysis for Biologists. *Cambridge University Press, 4th printing*. 13(13.2), 360-371.

80. Möls T. (2005). Linear Statistical Methods for Estonian Freshwater Waterbodies. Naturalist Handbooks II. *Eesti Loodusuurjate Selts.* 3.1.3, 85-94.

81. Statistical Analysis Sevice SAS. SAS Institute Incorporation 2006. Version 9. http://support.sas.com/documentation/onlinedoc/sas9doc.html

82. Cook N. (2003). The use of NASBA for the detection of microbial pathogens in food and environmental samples. *Journal of Microbiological Methods.* 53(2), 165-74.

83. http://archive.bmn.com/supp/tibtec/PavlovTable1.pdf, veebr 2006.

84. Wick LM, Rouillard JM, Whittam TS, Gulari E, Tiedje JM, Hashsham SA. (2006). On-chip non-equilibrium dissociation curves and dissociation rate constants as methods to assess specificity of oligonucleotide probes. *Nucleic Acids Res.* 34(3), e26;1-10.

85. Kaplinski L. Publitseerimata. *http://bioinfo.ebc.ee/download/,* mai 2006.

LISAD

Lisa 1.

Näide liigispetsiifilisest järjestusest väljavõttena mitmese-joonduse programmi CLUSTALW väljundist. Sinisega on näidatud uuritav genoom (sinine - *Neisseria gonorrhoeae FA 1090)* hele- ja tumepunasega mitte-uuritavad genoomid (helepunane – GenBank ID AL157959, *Neisseria meningitidis Z2491*; punane GenBank ID AE002098, *Neisseria meningitidis MC58)*

```
NC_002946.2064420.2064619.16          --AAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.617595.617794.16            --AAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1274979.1275178.16          --AAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1208100.1208299.16          --AAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1255933.1256132.16          --AAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.500848.501047.16            --AAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.662334.662533.16            --AAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.536965.537164.16            --AAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.204518.204717.16            --AAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.166855.167054.16            --AAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1097850.1098050.16          -GAAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1376593.1376792.16          -GAAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1357439.1357638.16          -GAAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1160599.1160798.16          -GAAAGAAGTACAGAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1123495.1123694.16          -GAAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.729041.729240.16            -GAAAGAAGTACAAAAAGAACCGCTCCGTTCTTTTGTACCGGAAGTTACC
NC_002946.1439525.1439724.AL157959    -GAAAAAAGTACAGAAAGAACTGCTCCGTTTTTT-GTACTGGAAGTTACC
NC_002946.1890989.1891188.AL157959    -GAAAAAAGTACAGAAAGAACTGCTCCGTTTTTT-GTGCTGGAAGTTACC
NC_002946.1383424.1383623.AL157959    --AAAAAAGTACAGAAAGAACTGCTCCATTTTTTTGTACTGGAAGTTACC
NC_002946.612439.612638.AE002098      AGAAAAAAGTACAGAAAGAACTGCTCCGTTTTTT-GTACTGGAAGTTACC
                                        *** ******* ******* ***** ** *** ** * **********

NC_002946.2064420.2064619.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.617595.617794.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1274979.1275178.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1208100.1208299.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1255933.1256132.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.500848.501047.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.662334.662533.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.536965.537164.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.204518.204717.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.166855.167054.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1097850.1098050.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1376593.1376792.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1357439.1357638.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1160599.1160798.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1123495.1123694.16          GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.729041.729240.16            GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCGATTCGGCGGCACT
NC_002946.1439525.1439724.AL157959    GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCAATTCGGCAGCACT
NC_002946.1890989.1891188.AL157959    GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCAATTCGGCAGCACT
NC_002946.1383424.1383623.AL157959    GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCAATTCGGCAGCACT
NC_002946.612439.612638.AE002098      GCCCGTTCTGCCGCCGATATTTTGGGTATCCATCCCAATTCGGCAGCACT
                                      ********************************** ******* *****
```