BIOLOOGIA-GEOGRAAFIA TEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOTEHNOLOOGIA ÕPPETOOL

Reidar Andreson

Erinevate *in silico* meetodite võrdlus PCR praimerite kvaliteedi parandamiseks

Magistritöö

Juhendajad**:** Maido Remm, Ph.D.
prof. Andres Metspalu, M.D., Ph.D.

TARTU
2002

Sisukord

Lühendid ja mõisted

score	väärtus (skoor)
cutoff	katkestuspunkt
gap	vahe kahe järjestuse vahel (mittekattuvad nukleotiidid)
PCR	polümeraasi ahelreaktsioon (*Polymerase Chain Reaction*)
DNA	desoksüribonukleiinhape
bp	aluspaar
kb	tuhat aluspaari
APEX	oligonukleotiidmaatriksil põhinev praimerekstensioon (*Arrayed Primer EXtension*)
SNP	ühenukleotiidne polümorfism (*Single Nucleotide Polymorphism*)
Gb	gigabait
MS-DOS	*Microsoft Disk Operating System*
RAM	arvuti operatiivmälu

Sissejuhatus

Järjest laienenud teadmised genoomis esinevatest varieeruvustest ja pidevalt täiustuv kaasaegne tehnoloogia, võimaldavad püstitada kogu genoomi hõlmavaid küsimusi ning vaadelda inimgenoomi kui tervikut. Laiaulatuslikud mass-genotüpiseerimise projektid lubavad suurte koguste SNP'de (ühenukleotiidne polümorfism) analüüsi (Wang *et al.*, 1998). Genotüpiseerimise meetoditest on kõige levinumad praimerekstensioonil põhinevad tehnoloogiad, mille eelisteks on lihtsus, kiirus, automatiseeritavus ja järjestuse spetsiifiline reaktsioon (Landegren *et al.*, 1998). Kõige selle juures on vajalik valmistada tuhandeid PCR produkte paralleelseks detekteerimiseks geenikiibil (Tõnisson *et al*., 2000b; Pastinen *et al*., 2000; Kurata *et al*., 2000).

PCR praimerite disainiks on loodud palju erinevaid programme ja algoritme, mis kasutavad enamasti traditsioonilisi meetodeid kvaliteedi kontrolliks – sulamistemperatuur 62°C, GC sisaldus ~50%, sekundaarstruktuuride tekke võimalus jne. Sageli ei leita optimaalses parameetride vahemikus soovitavaid parameetreid ja siis tekib kasutajal küsimus, milliseid parameetreid oleks õige laiendada.

Lisaks on genoomsete PCR praimerite disainil oluline temperatuuri ja järjestuse optimiseerimise kõrval ka nende unikaalsus vastavas genoomis. Varem ei olnud vajadust sellise seostumiskohtade otsingu järele, sest eksperimentides kasutati tihtipeale *template*'na plasmiidset DNA'd ning praimerite ja produktide unikaalsuse kontroll polnud nii hädavajalik. Genoomse DNA kasutamisega eksperimentides on suurem oht PCR praimeril seostuda alternatiivsetesse kohtadesse ja seega vähendada PCR kvaliteeti.

Käesoleva töö eesmärgiks on anda ülevaade olemasolevatest enam levinud praimeri disaini ja järjestuste võrldemise meetoditest. Teiseks püütakse leida optimaalseid parameetreid PCR kvaliteedi tõstmiseks, uurides praimerite GC sisalduse, produktide ja praimerite pikkuste korrelatsiooni kvaliteediga. Kolmandaks uuritakse seostumiskohtade ja alternatiivsete produktide arvu mõju PCR kvaliteedile ja vaadeldakse järjestuste võrdlemise meetodite kiirust, mälu kasutust ja tulemuste kvaliteeti, et selekteerida välja optimaalseimad parameetrid kogu protsessi jaoks PCR praimerite disainist kuni seostumiskohtade otsinguni välja.

I Kirjanduse ülevaade

1. Genotüpiseerimine ja genoomne PCR

Laiaulatuslikud mass-genotüpiseerimise projektid hõlmavad tuhandete kuni miljonite SNP markerite analüüsi ja praeguseks lubavad kaasaegsed genotüpiseerimismeetodid efektiivselt detekteerida üksikuid kuni tuhandeid SNP-sid ühes katses (Wang *et al.*, 1998). Geenikiibil põhineva tehnoloogia kasutamine võimaldab meil teha kogu genoomi hõlmavat miniatuurset paralleelanalüüsi, seega on SNP markerite laiaulatusliku infopanga koostamise üheks võimaluseks tulevikus geenikiibil põhinevate meetodite kasutamine. Põhiliselt kasutatakse SNP-de analüüsimiseks kas ainult hübridisatsioonil või hübridisatsioonil ja praimerekstensioonil põhinevaid geenikiibi tehnoloogiaid (Landegren *et al.,* 1998).

Genotüpiseerimise meetoditest on kõige levinumad praimerekstensioonil põhinevad tehnoloogiad. Praimerekstensiooni eeliseks on lihtsus, meetod on kiire, automatiseeritav ja toimub järjestuse spetsiifiline reaktsioon. Praimerekstensiooniga on detekteeritavad praktiliselt kõik mutatsioonide tüübid, v.a. praimerist pikemad kordusjärjestused ja trinukleotiidsed kordused (Tõnisson *et al.,* 2000b, Kurg *et al*., 2000).

Teaduslike ja diagnostiliste genotüpiseerimise eksperimentide juures on paljude meetodite puhul (sh. APEX - *Arrayed Primer EXtension*) vajalik valmistada tuhandeid PCR produkte paralleelseks detekteerimiseks geenikiibil (Tõnisson *et al*., 2000b; Pastinen *et al*., 2000; Kurata *et al*., 2000). Genoomsete PCR praimerite disainil lisaks on temperatuuri ja järjestuse optimiseerimisel oluline ka nende unikaalsus vastavas genoomis. Kui eksperimentides kasutati plasmiidset DNA'd või kui puudusid andmed genoomse nukleotiidse järjestuse kohta, ei olnud unikaalsuse kontroll väga oluline.

2. PCR kvaliteeti mõjutavad parameetrid

DNA dupleksi ehk praimeri ja DNA ahela seostunud struktuuri stabiilsuse põhiliseks määrajaks on nukleotiidipaaride koostis/järgnevus. Praimerite disaini juures arvestatakse enamasti 2 põhilist parameetrit – sulamistemperatuur (Tm) ja produkti pikkus. Lisaks vaadatakse veel praimeri pikkust, GC nukleotiidide sisalduse protsenti (%) praimeris, sekundaarstruktuuride tekke võimalus ja praimeri 3' otsa

nukleotiidset koostist. Praimeri pikkus ja GC sisalduse % on sõltuvuses sulamistemperatuurist (Haas *et al*., 1998).

2.1. Praimerite sulamistemperatuur

Praimeri järjestuse sulamistemperatuuri arvutamiseks on mitmeid viise, milles kõige lihtsam – A/T nukleotiidi paar lisab 2 °C ja G/C 4 °C. DNA dupleksi sulamistemperatuuri arvutamise täpsemaks algoritmiks on *Nearest-Neighbor* meetod (Borer *et al*., 1974). Breslauer kaasautoritega arvutasid aga välja kõigi 10 võimaliku nukleotiidipaaride seostumise energia väärtused (Joonis 1.) (Breslauer *et al*., 1986).

Interaktsioon	ΔH°(kcal/mol)	ΔS°(cal/K per mol)	ΔG°(kcal/mol)
AA/TT	9.1	24.0	1.9
AT/TA	8.6	23.9	1.5
TA/AT	6.0	16.9	0.9
CA/GT	5.8	12.9	1.9
GT/CA	6.5	17.3	1.3
CT/GA	7.8	20.8	1.6
GA/CT	5.6	13.5	1.6
CG/GC	11.9	27.8	3.6
GC/CG	11.1	26.7	3.1
GG/CC	11.0	26.6	3.1

Joonis 1. *Nearest-neighbor* testi tulemused vastavate DNA dupleksite lõhkumisel 1 M NaCl, 25 °C, pH 7 juures. DNA dupleksite stabiilsus (ΔG°) arvutati valemi järgi: ΔG° = ΔH° – TΔS°, kus T = 293 K. ΔG° väärtused vastavad energiale, mis kulub vastavate nukleotiidipaaride vesiniksidemete lõhkumiseks (Breslauer *et al*., 1986).

Näiteks järjestuse GGAT korral on *nearest neighbour* termodünaamika $\Delta H°(GGAT)_{total}$ = ΔH°(GG) + ΔH°(GA) + ΔH°(AT) = 11.0 + 5.6 + 8.6 = 25.2(kcal/mol) (Kämpke *et al*., 2001).

Valem sulamistemperatuuri arvutamiseks praimer-DNA interaktsiooni korral:

$$Tm_{primer} = \Delta H°_{total}/(\Delta S°_{total} + R * \ln(c/4)) - 273.15 + 16.6 \log[Na^+],$$

kus ΔH ja ΔS on praimeri-DNA kaksikhübriidi moodustumise entalpia ning entroopia väärtused, R universaalne molaarse gaasi konstant (=8.31 J/mol = 1.99cal/°C mol), c on oligonukleotiidide totaalne molaarne kontsentratsioon. Leiti, et (c) 250 pM oligonukleotiidide empiiriline kontsentratsioon andis kõigi testitud praimerite korral head eksperimentaalsed tulemused (Rychlik *et al*., 1990).

2.2. PCR produktide pikkus

Produkti pikkus on oluline eelkõige genotüpiseerimise projektides, kus samade tingimuste ja PCR tsükli aegade juures on korraga vaja üles amplifitseerida palju PCR produkte. Seega tuleb leida optimaalne produktide nukleotiidne pikkus, mille korral ei halveneks PCR kvaliteet. Samuti on lühematel PCRi produktidel väiksem võimalus anda valesid seostumisi APEX praimerite vastu. Geenikiibil toimuvas reaktsioonis neid kokku segades on oluline, et praimerid ei seostuks soovimatute produktidega (Tõnisson *et al.,* 2000b).

2.3. Praimerite GC sisaldus

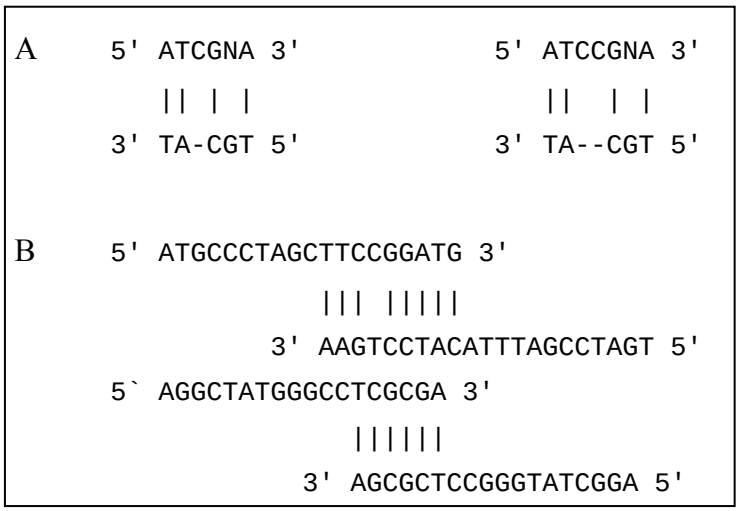
Praimeri spetsiifilise seostumise DNA järjestusega määrab tema 3' ots, täpsemalt GC nukleotiidide sisaldus 3' otsas. Seni levinud praktika nõuab GC-rikka 3'otsa kasutamist (nn. GC-clamp). Samas on näidatud, et kõrge GC nukleotiidide sisaldus praimeri 3' pooles võib põhjustada vale seostumist ja vähendada sellega praimeri spetsiifilisust (Beasley *et al*, 1999; Li *et al*., 1997).

2.4. Praimerite omavahelised interaktsioonid ja seostumiskohad genoomis

Vajalik on testida praimerite hübridiseerumist iseendaga (*self annealing*) ja vastaspraimeriga täispikkuses ning praimeri 3' otsa seostumist (*self-end annealing*) enda ja vastaspraimeri 3' poole järjestusega (Joonis 2). Taoline kontrollimine on PCR kvaliteedi tõstmisel oluline, sest iseendaga hübridiseerumiste tõttu väheneb *template* DNA'ga seostuvate praimerite kontsentratsioon PCR reaktsiooni segus ja võivad tekkida praimerite dimeerid (kaks praimerit seostuvad üksteise külge ja nende pealt toimub ekstensioon) (Kämpke *et al*., 2000).

Lisaks eelnevatele parameetritele tuleb kontrollida praimerite võimalikke seostumiskohti ja vastavalt tekkivaid lisaprodukte genoomse DNA peal. Enamasti on alternatiivsete seostumiskohtade põhjustajaks praimerid, mis sisaldavad genoomseid DNA kordusjärjestuste motiive. Hoolimata sellest, et kandidaatpraimereid on võimalik mitmetes andmebaasides teadaolevate korduste vastu kontrollida, ei ole need kollektsioonid täielikud ja pidevalt avastatakse uusi kordusjärjestusi. Praimerite

seostumist kordusjärjestustesse on proovitud vältida nende võrdlemisega STS andmebaasi vastu (Schuler *et al*., 1997). Teine võimalus on vältida praimeri tegemisel piirkondi milles asuvate oktameersete sõnade sagedus on genoomis kõrge (Chenal *et al*., 1996).



Joonis 2. Praimerite iseendaga ja 3' otsaga hübridiseerumise näited. Rozen kaasautoritega kasutavad oma programmis PRIMER3 praimerite 3' otste ja iseendaga hübridiseerumise testimiseks maksimaalset lubatud skoori (*score*). Iga komplementaarne nukleotiid annab skooriks 1.00, paardumine N (N'iga tähistatakse teadmata nukleotiidid) nukleotiidiga –0.25, mittekomplementaarne paar annab –1.00 ja *gap* (tähistatud '-'ga) –2.00. Joonisel on A näitel tegemist praimeri iseendaga hübridiseerumise variantidega, kus esimese lõpp skooriks on 1.75, kuid teisel on 0, sest negatiivseid lõpp skoore ei arvestata. B poolel, mis näitab praimeri 3' otsa iseendaga hübidiseerumist, on ülemise skoor 7 ja alumisel vastavalt 6.

3. Erinevad praimerite disaini programmid

Joonisel 3 on välja toodud hetkel enam levinud meetodid PCR praimerite disainiks nii tasuta (*freeware*) kui kommertsiaalsete programmide näol. Kajastatud pole loomulikult kõik võimalikud programmid, sest paljude algoritmid on vananenud, arendustöö peatunud või programmi töökeskkond pole paljudele kasutajatele sobiv (*MS-DOS*).

Praimeri disaini programmid võib esiteks jaotada kahte gruppi: need, mis on mõeldud suuremahuliste projektide jaoks nagu PRIMER3 (Rozen *et al*., 1998), PRIDE (Haas *et al*., 1998), PRIMO (Kenneth *et al*., 1996), PRIMEARRAY (Raddatz *et al*., 2000), GST-PRIME (Varotto *et al*., 2001) ja programmid, mis on mõeldud väiksemate koguste praimerite valmistamiseks – OLIGO (Rychlik *et al*., http://www.oligo.net/), DOPRIMER (Kämpke *et al*., 2000), PRIME (http://www.accelrys.com/ endine PRIMA), PRIMERSELECT (http://www.dnastar.com/products/PrimerSelect.html).

nimi	Aadress
DOPRIMER	http://doprimer.interactiva.de/pro/frameset.html
GST-PRIME	leister@mpiz-koeln.mpg.de
MEDUSA*	http://www.cgr.ki.se/cgr/MEDUSA/
OLIGO	http://www.oligo.net/
OLIGOARRAY*	http://berry.engin.umich.edu/oligoarray/
PC-RARE*	http://bioinformatics.weizmann.ac.il/software/PC-Rare/windows/
PRIDE	http://www.dkfz-heidelberg.de/tbi/services/Pride/prideform
PRIME	http://www.accelrys.com/products/gcg_wisconsin_package/program_list.html#Primer
PRIMEARRAY*	christoph.dehio@unibas.ch
PRIMER MASTER*	ftp://ftp.ebi.ac.uk/pub/software/dos/primer-master/
PRIMER PREMIER 5	http://www.premierbiosoft.com/primerdesign/primerdesign.html
PRIMER3*	http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
PRIMERSELECT	http://www.dnastar.com/products/PrimerSelect.html
PRIMO*	http://innovation.swmed.edu/primo.htm

Joonis 3. Enam levinud praimeri disaini programmid. Tärniga (*) märgistatud programmid on tasuta saadaval akadeemilistele kasutajatele. Tabelis on välja toodud enam levinud/tsiteeritud programmid ja tegelikkuses on erinevate programmide hulk suurem.

Teine võimalus programmide klassifitseerimiseks on nende automatiseeritus, ehk kasutaja võimalikult minimaalne vahelesekkumine praimeri disaini juures. PRIMER PREMIER 5 (http://www.premierbiosoft.com/) ja PRIDE on praimateks näideteks, kus esimesel juhul muudetakse praimeri disaini parameetreid dünaamiliselt lõdvemaks sobivate kandidaatide mitte leidmisel ning teisel juhul ei eemaldata ühtegi praimeri kandidaati enne kui on välja arvutatud kõik parameetrid kõikide praimerite kohta (Haas *et al*., 1998).

Kolmandaks jaotuse kriteeriumiks võib olla praimeri järjestuse võrdlemine DNA järjejestusega. Suurte genotüpiseerimise projektide juures, kus PCR praimerite sünteesiks kasutatakse genoomset DNA'd, on oluline kontrollida nende unikaalsust erinevalt eksperimentidest, kus kasutatakse plasmiidset DNA'd. Genoomsete järjestuste pidev kogumine/täiendamine suurtes andmebaasides võimaldab praegu uurida molekulaarseid protsesse laiemas plaanis, kui varem (Lander *et al*., 2001).

Genoomse DNA kasutamisega eksperimentides on suurem oht PCR praimeril seostuda alternatiivsetesse kohtadesse ja seega vähendada PCR kvaliteeti. PRIDE kontrollib praimeri seostumise stabiilsust sisestatud DNA järjestuse vastu (kontiigid), mille pealt praimereid sünteesida soovitakse ning vastava kontiigiga seotud järjestuste vastu. Iga seostumiskoha juures määratakse praimeri 8 viimase nukleotiidi (3´ otsas) ja DNA järjestuse vaheline stabiilsus ning originaalasukoha ja kõige stabiilsema alternatiivse koha vaheline suhe määrab praimeri unikaalsuse (Haas *et al*., 1998). OLIGOARRAY kasutab praimerite kontrolliks BLAST meetodit vastava

genoomse järjestuse vastu, mille pealt oligonukleotiide sünteesitakse (Rouillard *et al*., 2002).

Enamuse ülalnimetatud programmide juures on võimalik kasutada alternatiivsete praimerite seostumiskohtade vältimiseks kordusjärjestuste andmebaasi päringut. MEDUSA võimaldab aga maskeerida korduvad piirkonnad kohe kasutaja poolt sisestatud *input* järjestuses (Podowski *et al*., 2001). Hoolimata praimeri disaini programmide suurest valikust, ei ole järjestuste genoomse DNA vastu võrdlemine nende algoritmides levinud.

4. DNA järjestuse sarnasuse ja identsuse otsingu meetodid

Järjestuste võrdlemisel tuleb eristada kolme terminit: identsus, sarnasus ja homoloogia. Järjestuste identsus tähendab, et järjestuse X positsioonis n (X_n) esineb täpselt sama nukleotiid (või aminohape valkude puhul), mis võrreldava järjestuse Y_n positsioonis. Sarnasuse korral võetakse arvesse ka kaudsemad identsuse kohad kahe järjestuse vahel, mille vahel esinevad *gap*'id, ja arvutatakse välja vastav skoor, mis näitab tõenäosust, et tegemist on kas sarnaste või mittesarnaste järjestustega. Kahe järjestuse homoloogia juures ei ole järjestused X ja Y mitte ainult suure tõenäosusega sarnased, vaid ka nende organismide eellaste järjestused omasid taolisi järjestusi (Pertsemlidis *et al*., 2001).

4.1. Dünaamiline programmeerimine

Algoritmid, millega võrreldakse kahte järjestust, põhinevad dünaamilisel programmeerimisel (*dynamic programming*). Esimest korda kirjeldati eelnimetatud programmeerimise meetodeid 1950 aastatel väljaspool bioinformaatika teadust ja 1970 aastal kasutasid Needleman ja Wunsch neid bioloogilises kontekstis. Dünaamiline programmeerimine tähendab olemasolevale probleemile optimaalse lahenduse leidmist jagades selle alamprobleemideks ning kus lõpuks leitakse kõige madalamal tasemel lihtsaim resultaat, mis kantakse üle algse probleemi lahendamiseks. DNA järjestuse juures on probleemiks nende optimaalseima kokkulangemise (*optimal alignment*) leidmine. Jaotades järjestusi järjest väiksemateks "alamjärjestusteks" jõutakse lõpuks 1 nukleotiidi tasemeni välja. Seega koostatakse nukleotiidsel tasemel maatriks, milles püütakse leida kõige lühem tee diagonaali mööda maatriksi ühest nurgast teise (Pertsemlidis *et al*., 2001).

Ülal kirjeldatud dünaamilist programmeerimist kasutasid ära Smith ja Waterman (1981), kes optimiseerisid matemaatiliselt seda meetodit, mis suudaks anda kõrgeima skooriga *alignment* järjestuste vahel ühe arvutusprotsessiga. Siiski on tegemist aeganõudva ja suurt arvutivõimsust vajava algoritmiga, kuid samas on garanteeritud parim võimalik järjestuste *alignment* (Smith *et al*., 1981; Arslan *et al*., 2001). Sellisel *Smith-Waterman* nimelisel algoritmil põhineb programmi SSEARCH töö (Pearson *et al*., 1988; Pertsemlidis *et al*., 2001).

4.2. BLAST2

Enim levinud ja tsiteeritud meetodiks järjestuste võrdlemisel on BLAST (*Basic Local Alignment Search Tool*), mis sarnaselt varasema meetodiga FASTA kasutab heuritsilisi algoritme otsimaks päringule lähimat vastet järjestuste andmebaasist (Altschul S., 1999**;** Altschul *et al*., 1990**;** Pearson *et al*., 1988). Varasema BLAST versiooniga võrreldes on uuem BLAST2 (ver 2.0, *Gapped BLAST*) algoritm 3 korda kiirem ja võimaldab otsingus kasutada ka *gap*'e. Kuna *gap*'id modelleerivad evolutsioonis toimunud insertsioone ja deletsioone on tulemused ka bioloogiliselt mõistlikumad kui *gap*'ideta versiooni korral (Altschul *et al*., 1997). BLAST2, mis on optimiseeritud kiiruse suhtes, kasutab fikseeritud sõnapikkust – 11 nukleotiidi – ja otsing toimub mõlemalt DNA ahelalt erinevalt FASTA3'st, kus on vaja teha kaks päringut sama resultaadi saamiseks (Pearson, 1998).

BLAST2 algoritm töötab põhimõttel, et kõigepealt leitakse kõik sõnad pikkusega *W* ja skooriga *T*. Kui 2 eelnevale kriteeriumile vastavat sõna satuvad sama diagonaali peale otsingu aknas ($A < 40$), pikendatakse mõlemaid sõnasid mõlemas suunas piki diagonaali seni, kuni skoori *S* absoluutväärtus ei lange alla 22 *bit*'i või ei lange üle 7.5 *bit*'i senisest parimast väärtusest. Seejärel sooritatakse *gapped alignment* ja katkestatakse kui skoor langeb alla 15. Eelkirjeldatud algoritmi nimetatakse BLAST2 *two-hit* meetodiks, sest varasem BLAST võimaldas teha ainult *one-hit* meetodit, mis põhines ainult 1 sõna leidmisel ja pikendamisel ning ei kasutanud *gapped alignment*'i (Altschul *et al*., 1997; Pertsemlidis *et al*., 2001).

BLAST2 skoori arvutamiseks kasutatakse järgmist valemit, kus λ ja K tähistavad Karlin-Altschul parameetreid ning S_{raw} esialgset skoori järjestuste identsuse võrdlemisel (Tatusova *et al*., 1999):

$$S_{bits} = [\lambda * S_{raw} - \ln K] / \ln 2.$$

Lisaks BLAST2 skoorile on teiseks olulisemaks väärtuseks E (*expected value*). E-väärtus näitab mitu antud skooriga järjestust võiks antud suurusega andmebaasist tänu juhuslikule kokkusattumisele (ilma bioloogilise tähenduseta) leida. Valem E-väärtuse arvutamiseks

$$E = mn2^{-S},$$

kus m tähistab andmebaasi järjestuse, n päringu pikkust ja S võrdub S_{bits}. Suuremate andmebaaside korral on E-väärtus kõrgem kui väiksematel. E-väärtus võimaldab võrrelda erinevates andmebaasides tehtud otsinguid ja on bioloogiliselt mõtestatud ainult siis, kui andmebaas koosneb peamiselt juhuslikest ja juhusliku nukleotiidse koostisega järjestustest. Paljudel juhtudel on andmebaasis liiga palju mittejuhuslikke, "valitud" järjestusi ning tegemist on nihkega (*biased*) andmebaasis (Altschul *et al*., 1997; Pertsemlidis *et al*., 2001).

BLAST2 otsingut on lühemate (7-20 aluspaariliste) järjestuste juures võimalik kiirendada, kui ühendada näiteks PCR praimerite järjestused üheks pikaks eraldatud päringuks. Üheks automaatselt selliseid päringuid teostavaks programmiks on *Perl*'is kirjutatud MPBLAST. Nimetatud programm koosneb skriptide paketist, mis ühendab lühikesed järjestused üheks kokku ja filtreerib pärast tulemustest välja igale alamjärjestusele vastavad resultaadid (Korf *et al*., 2000).

4.3. MEGABLAST

Teine võimalus kiiremaks alignmentide leidmiseks pikkade nukleotiidsete järjestuste vahel on MEGABLAST. MEGABLAST kasutab nukleotiidsete järjestuste võrdlemisel *greedy* algoritmi dünaamilise programmeerimise asemel ja suuremate sõnapikkuste juures võib programmi kiirus ületada traditsioonilisi järjestuste sarnasuse leidmise meetodeid kuni 10 korda. *Greedy* algoritmi peamine erinevus dünaamilisest programmeerimisest seisneb selles, et paljud maatriksi punktid jäetakse kõrvale, kuna ainult ühte diagonaali pikendatakse mõne punkti võrra ja MEGABLAST kasutab standartsena (*default*) pikemat sõnapikkust kui BLAST2. MEGABLAST'is on sõna pikkus minimaalne identse järjestuse pikkus, mille algoritm üles leiab ja kõige efektiivsemalt töötab programm 16 nukleotiidse järjestusega, kuigi võimalik on minimaalselt 8 pikkust kasutada. Kui *W* väärtus jagub 4, garanteerib meetod selle, et kõik perfektsed järjestuse *alignment*'id pikkustega *W* + 3 leitakse

üles. Vastasel juhul võetakse aluseks lähim 4 jaguv väärtus ja proovitakse *alignment*'i pikendada soovitud sõna pikkuse suunas. Seega MEGABLAST on sobilikum 95% ja enama identsusega järjestuse leidmiseks. MEGABLAST vajab ~1Gb arvuti mälu (*RAM*) järjestusi inimese genoomse DNA kontiigide vastu kontrollimiseks (Zhang *et al*., 2000).

Lisaks on olemas programm BLAT (*BLAST-Like Alignment Tool*), mis ei ole seotud BLAST perekonnaga, kuid on mõeldud järjestuste 95% ja rohkema sarnasuse leidmiseks. Kuid täpsemad tulemused saadakse pikemate päringute korral järjestuse pikkusega alates 40 nukleotiidist ja seega statistika lühemate järjestuste jaoks võib puududa. BLAT koostab ja hoiab kogu genoomi indeksid mälus, mis sisaldab infot kõigi mittekattuvate 11-meeride (11 nukleotiidi pikkusega sõnad) kohta genoomis (Jim Kent, http://genome.ucsc.edu/cgi-bin/hgBlat; Kent, 2002).

4.4. SSAHA

Ning kaasautoritega kasutavad oma programmis SSAHA sellist meetodit, kus eelnevalt genoomne DNA indekseeritakse ja kirjutatakse suurde *hash* tabelisse. Hiljem otsingut sooritades loetakse tabel arvuti mällu ja päringu võrdlemine indeksi vastu toimub kiiresti (Ning *et al*., 2001).

Joonisel 4 on näha *hash* tabeli loomise põhimõte, kus arvuti mällu luuakse kaks andmestruktuuri: kõigi võimalike *pointer*'ite jada *A* ja nende esinemis positsioonide loetelu *L*. Erinevate *pointer*'ite arv sõltub sõna pikkusest $k - 4^k$. Järjestuse otsimine indekseeritud andmebaasist toimub sõna pikkuse k haaval. Koostatakse esialgne loetelu *H*, kuhu lisatakse iga päringu sõna pikkuse k kohta vastav positsioonide jada *L*. *Master* loetelu *M* sisaldab juba *H* jada esimese ja teise liikme järgi sorteeritud positsioone (*index* ja *shift*), mille järgi detekteeritakse asukoht andmebaasi järjestuses otsides üles sellised positsioonid, kus *index* ja *shift* omavad samu väärtusi (Ning *et al*., 2001).

Kõige rohkem aega nõuab seostumiskohtade otsingul *hash* tabelist *M* loetelu sorteerimine. Meetodi tööd on võimalik kiirendada, kui piirata sama *pointer*'i kohta tekkiva jada *A* suurust. Niimoodi on võimalik piirata korduvate järjestuste esinemist *hash* tabelis ja otsingu sooritamine muutub vastavalt kiiremaks. Siiski kulub tavakasutajal palju aega varem kirjutatud tabeli mällu lugemine arvuti kettalt, sest kogu indeksi mälus hoidmiseks on vaja ~16Gb *RAM*'i (Ning *et al*., 2001).

S_1 = ATCGCAATCCAGCTTCTAGA
S_2 = GTCGAATTGAGCGGACGCTGGT

w	***pointer'*id**	**positsioonid**
AA	0	(2,5)
AC	1	(2,15)
AG	2	(1,11)
AT	3	(1,1) (1,7)
CA	4	(1,5)
CC	5	(1,9)
CG	6	(1,3) (2,3)
CT	7	(1,13)
GA	8	(1,19) (2,9)
GC	9	(2,11) (2,17)
GG	10	(2,13)
GT	11	(2,1) (2,21)
TA	12	(1,17)
TC	13	(1,15)
TG	14	(2,19)
TT	15	(2,7)

Joonis 4. 2 tähelise (2-*tuple*) sõna pikkusega *hash* tabeli loomine kahe järjestuse S_1 ja S_2 põhjal. *Hash* tabel sisaldab 4^k *pointer*'ite jada *A*, kus *k* tähistab sõna pikkust ning , ja positsioonide loetelu *L*. Positsioonide jadas olev esimene number (sulgude sees) tähistab järjestuse numbrit, mille pealt vastav 2 täheline sõna on leitud. Teine number näitab positsiooni vastavas järjestuses.

4.5. Suffiksi puud (*Suffix-trees*)

Viimastel aastatel on esile tõusnud *suffix-tree*'del põhinevad algoritmid just tänu nende kiirusele ja arvutite võimsuse kasvule. Delcher kaasautoritega valmistasid programmi, mis samuti kasutab suurte järjestuste võrdlemisel *suffix-tree*'l põhinevat algoritmi, kuid mälu kasutus üle 30 baidi ühe aluspaari kohta on inimese genoomi mastaabis liialt suur maht praegu kasutusel olevate arvutite jaoks (Delcher *et al*., 1999). *Suffix-tree*'de suurt arvuti mälu nõudlust näitasid oma töös ka Aach kaasautoritega (Aach *et al*., 2001).

II Eksperimentaalne osa

Töö eesmärgid

Käesoleva töö eesmärgiks on anda ülevaade olemasolevatest enam levinud praimeri disaini ja järjestuste võrldemise meetoditest. Teiseks püütakse leida optimaalseid parameetreid PCR kvaliteedi tõstmiseks, uurides praimerite GC sisalduse, produktide ja praimerite pikkuste korrelatsiooni kvaliteediga. Kolmandaks uuritakse seostumiskohtade ja alternatiivsete produktide arvu mõju PCR kvaliteedile ja vaadeldakse järjestuste võrdlemise meetodite kiirust, mälu kasutust ja tulemuste kvaliteeti, et selekteerida välja optimaalseimad parameetrid kogu protsessi jaoks PCR praimerite disainist kuni seostumiskohtade otsinguni välja.

Kasutatud meetodid

1. PCR praimerite disain

Käesolevas magistritöös kasutasime PCR praimerite kvaliteediandmeid, mis saadi 1278 erineva SNP analüüsil inimese kromosoom 22 pealt. SNP'de valimine teostati 1,5K projekti raames koostöös Sanger'i keskusega Inglismaal (http://www.sanger.ac.uk/). Praimerite kvaliteeti hinnati agaroosgeeli piltide vaatlemise tulemusena, kus negatiivseks loeti praimeri paarid, mis andsid mitu produkti või olid ilma produktita. Praimerite seostumis kohtade leidmise programmide parameetrite optimiseerimiseks kasutasime 100 negatiivset (PCR kvaliteet alla 50%) ja 100 positiivset (kvaliteet üle 95%) praimerite paari (Joonis 5.).

Ligikaudu 300 SNP jaoks jaoks olid PCR praimerid disainitud käsitsi, ülejäänud disainisime automaatselt. Praimerite disaini programmiks valisime PRIMER3, kuna tegemist on akadeemiliste projektide jaoks tasuta programmiga ja mis töötab UNIX'i keskkonas. Teise valikuna oli ka PRIMA EMBOSS bioloogiliste programmide paketist (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/), kuid sisendväljade väikese valiku tõttu eelistasime PRIMER3'e.

Põhilised parameetrid, mida kasutasime PRIMER3 juures olid praimeri pikkus – 18, 20, 25 (minimaalne, optimaalne ja maksimaalne) nukleotiidi, praimeri GC sisaldus – 20%, 50%, 80% ning PCR produkti pikkus – 100, 200, 600 nukleotiidi. Kordusjärjestuste raamatukogu (*repeat library*) praimerite disainimise juures ei

kasutatud. Ülejäänud parameetreid kasutati vaikimisi (*default*) programmi poolt etteantuna.

Praimeri disaini erinevate parameetrite ja PCR kvaliteedi vahel püüdsime leida korrelatsiooni, et välja selgitada optimaalseimad väärtused ja analüüsiks kasutasime 714 praimeri paari. 564 praimeri paari ei kasutatud, kuna nendega oli sooritatud kordus eksperimente alla 10 korra.

2. Genoomi-test

Genoomi-testiks nimetame protsessi, kus otsitakse üles kõik võimalikud PCR praimerite seostumiskohad genoomis ja ennustatakse nende pealt võimalike produktide teket. Testi kasutasime nii 714 praimeri paari genoomsete seostumiste leidmiseks kui ka 2 valimi (100 pos. ja 100 neg. praimeri paari) korral (Joonis 5.).

Testi läbiviimiseks kasutati nelja programmi: BLAST2, SSAHA, MEGABLAST ja käesoleva töö raames *Perl*'is (http://www.perl.com/) kirjutatud programm Identsuse Otsing (IS – *Identity Search*). Kui IS, BLAST2 ja SSAHA kasutasid inimese assembleeritud kromosoomide järjestusi, siis MEGABLAST võrdles praimerite järjestusi ENSEMBL'i "*golden-path*" (http://www.ensembl.org/) kontiigide järjestuse vastu. Tegelikud positsioonid kromosoomides leiti hiljem spetsiaalse *parser*-programmi abil andmebaasist päringuid sooritades. Kõikide programmide väljundis arvestati lõpptulemuses ainult selliste praimerite seostumiskohtadega, kus oli kinnitunud 3' ots. Võimalike PCR produktide arvu leidmiseks lahutati DNA vastasahela (*antisense*) praimeri koordinaadist kodeeriva ahela (*sense*) praimeri koordinaat ja arvestati neid produkte, mille pikkus ei ületanud 1000 aluspaari.

3. Genoomi-testi uuritavad parameetrid

BLAST2 programmi testiti nii parameetri –g T kui ka –g F korral, mis tähendavalt vastavalt, kas vahed (*gaps*) on lubatud või mitte. Ilma *gap*'ideta BLAST2 tulemusi ei saanud arvestada, kuna programm ei töötanud korrektselt, kui oli tegemist suurearvuliste seostumiskohtadega.

Teiseks parameetriks, mida testiti, oli minimaalne vajalik PCR praimeri pikkus 3' otsast eraldamaks positiivseid negatiivsetest. Erinevad pikkused, mida vaadeldi, olid pikkusega 12 nukleotiidist kuni 20. Neid numbreid saab võrrelda

BLAST2 ja MEGABLAST väärtustega (*bitscore*) 24, 26, …, 40, ehk siis iga identse nukleotiidi kohta antakse 2 punkti.

Kolmandaks võrreldi genoomi-testi meetodite kiirust erinevate praimeri hulkadega: 1, 10, 100 ja 1000 (Joonis 5). Kiiruse testid sooritati PIII Intel 700 MHz protsessoriga ja 2 GB mälumahuga arvuti peal. Aega arvutati programmi käivitamise hetkest kuni lõpptulemuste kättesaamiseni. Aeg sisaldab ka tulemuste interpreteerimisprogrammide (*parser*) tööd, mis filtreerisid välja praimerite 3' otsa seostumised ja arvutasid võimalikke produkte. Kuna *parser*'ite töötamise algoritm oli kõigi meetodite juures sarnane, ei esinenud nende meetodite tööajas märkimisväärset erinevust.

Genoomi testi korral kasutati *ENSEMBL* 3.26.1 (Jaanuar 2002) andmebaasi inimese genoomi DNA järjestust ja SNP'de andmeid samast andmebaasist.

Tulemused

1. PCRi kvaliteedi seos praimeri omadustega

Kuna sulamistemperatuur oli fikseeritud, testiti praimerite disaini juures järgmisi parameetreid, mis võiksid mõjutada PCR tulemuste kvaliteeti: GC sisaldus (praimeri 3' pooles ja viimastes nukleotiidides), praimeri pikkus ja produkti pikkus. PCR praimerite kvaliteedi piiriks määrati väärtus 80%, millest madalama protsendiga loeti mittekvaliteetseks praimeri paariks ja kõrgema protsendiga kvaliteetseks.

1.1. GC sisaldus ja praimerite pikkus

Levinud seisukohad on, et praimeri 3' otsas peaks olema soovitavalt C või G nukleotiid, et hübridiseerumine DNA ahela külge oleks tugevam ja praimeri pikkus võiks olla 21-26 aluspaari vahel. Praimerite pikkused varieerusid käesolevas töös 18-26 aluspaarini ja võis täheldada vähest kvaliteedi tõusu 21 aluspaari pikkusest alates.

Joonisel 6 on võrreldud praimerite 3' poole (praimerite järjestus jagati keskelt kaheks) GC nukleotiidide sisalduse korrelatsiooni PCR kvaliteediga. Selgus, et üle 50% GC sisaldusega praimerite kvaliteet langeb alla 80%. Arvatavasti põhjustab suurem GC sisaldus praimerite 3' osas mittespetsiifilist seostumist teistesse genoomi regioonidesse. Erinevalt varasematele arusaamadele labori praktikast leiti, et PCR kvaliteeti ei mõjuta oluliselt praimerite 3' otsa viimase 3 nukleotiidi GC sisaldus (Joonis 7.). Kuigi praimerite 3' otsa viimaste nukleotiidide GC sisalduse suurenedes PCR kvaliteet langes vähesel määral, püsib graafik defineeritud kvaliteedi *cutoff*'ist (80%) kõrgemal.

1.2. PCRi praimerite seostumine genoomsele DNA'le

PCR praimeri paaride juures uurisime nii summa (seostumiste arvud liidetuna iga praimeri paari kohta), minimaalse (võrdlusel kasutati väiksema seostumiste arvuga praimerit igast paarist) kui ka maksimaalse seostumiste arvu sõltuvust PCR kvaliteediga. Selgus, et summa korral saime *cutoff*'i 10 ehk siis praimeri paarid, mille seostumiste koguarv ületab 10, annavad PCR kvaliteediks vähem kui 80%. (Joonis 8.) Minimaalse *cutoff*'i leidmisel oli vastav number 3. Järelikult kui väiksema genoomsete seostumiste arvuga praimeri väärtus on alla 3 ja praimeri paari

seostumiste arv ei ulatu üle 10, võib praimerit lugeda positiivseks (genoomi-testi järgi kvaliteetseks). Seostumiskohtade tulemused saadi kõiki praimereid kontrollides genoomse DNA järjestuse vastu 18 sõnapikkusega 3' otsast.

1.3. PCR produktide ennustamine genoomselt DNA'lt

Lisaks uurisime tekkivate PCR produktide arvu ja pikkuse suhet PCR kvaliteeti (Joonis 9). Labori eksperimentides on enamasti oluline, et üles amplifitseeritakse ainult üks ja õige produkt ja iga alternatiiv vähendab selle kontsentratsioni ning põhjustab APEX reaktsiooni analüüsil mittespetsiifilisi (vale-positiivseid) signaale. Seetõttu tunnistame ebasobivaks kõik praimerid, mis võivad anda rohkem kui ühe produkti. Lisaks leidsime, et produkti pikenedes üle 400 nukleotiidi väheneb PCR kvaliteet samuti alla 80%. Kuna meie projektis kasutati kõigi praimerite jaoks sama PCR programmi (tsüklite pikkused samad), võib pikema produkti süntees katkeda ning lisaks eelnevale on vigade tekke võimalus suurem.

2. Genoomi-testide tulemused

Eelpool uuritud praimerite parameetrid võib põhimõtteliselt jagada kaheks - esimesed (GC%, praimerite pikkus) on seotud ainult praimerite omadustega, teised (seostumiste arv genoomis, produktide arv genoomis) aga *template* DNAga, meie näite puhul inimese genoomse DNAga. Et selgitada milline genoomi-testide läbiviimise metoodika on efektiivseim ja millised on parimad parameetrid, korraldasime täiendava katse. Valisime 100 positiivset ja 100 negatiivset PCR praimerit ning sooritasime nendega genoomi testi. Võrdlesime 4 erinevat programmi ja erinevaid praimerite 3' otsa pikkusi.

Joonisel 10 võib täheldada, et meetodid annavad sarnased tulemused ja umbes 50% negatiivsetest praimeritest on võimalik leida seostumiskohtade arvu kasutades. Kuigi erinevused pole praimeri pikkuste skooride vahel suured, saab optimaalseimaks pidada 16 ja 17 nukleotiidseid praimeri pikkusi. Analüüsil kasutati seostumiskohtade arvuna praimeri paaride seostumiste summat, kuna seostumiste minimaalne hulk ja produktide arv suutsid lahutada maksimaalselt 20% negatiivsetest praimeritest ja maksimaalne arv andis sarnased tulemused. Praktikas kasutamiseks on 16 sõna pikkus ka optimaalseim, sest 16 nukleotiidist koosnevaid sõnu saab tüüpilise arvuti peal väljendada täpselt 4 baidi abil. $4^{16} = 2^{32} =$ 4 baiti.

2.1. Kiiruse test 4 meetodiga

Teiseks uurisime, kaua võtab aega erinevate koguste praimeri paaride testimine genoomi vastu kõigi 4 meetodi korral. Tulemusi analüüsides selgus, et väikese hulga praimeri paaride korral (1-5) on kiireim MEGABLAST ja teised kolm üksteisest oluliselt ei erine (Joonis 11). Kuid 10 praimeri paari korral võis täheldada, et BLAST2 ja IS (*Identity Search*) programmi kiirus oli 2 korda aeglasem kui SSAHA'l ja 4 korda aeglasem kui MEGABLAST'il. 100 praimeri paari juures langes MEGABLAST'i kiirus samale tasemele SSAHA'ga, samas kui 1000 paari korral oli SSAHA teistest juba tunduvalt kiirem. SSAHA aegluse põhjuseks väiksemate andmemahtude juures on see, et programm kulutab suure hulga ajast indekseeritud tabelite lugemisele kettalt (inimese genoomi korral 16 Gb) ning 100 ja enama praimeri paari juures tema suhteline kiirus võrreldes teiste meetoditega tõuseb. Mälu kasutuselt on ressursinõudlikumad MEGABLAST ja SSAHA, mille puhul sooritades inimese 1. kromosoomi pealt päringut vajatakse ~1Gb mälu ruumi.

2.2. Optimaalsed *cutoff*'id seostumiskohtade otsingu programmide jaoks

Lähtuvalt eelnevast testist valiti kaks kiiremat meetodit, millega edasi töötada ja püüti leida neile vastavad optimaalseimad praimeri pikkuste *cutoff*'id positiivsete ja negatiivsete praimerite eraldamiseks. Joonisel 12 on näha kahe kiirema meetodi – SSAHA ja MEGABLAST – võrdlust positiivsete ja negatiivsete praimerite seostumiste arvude tulpade näol, kus 16 nukleotiidse praimeri pikkuse juures 3' otsast on selleks *cutoff*'iks ~30 seostumist genoomis. Antud juhul tähendab vastav *cutoff* seda, et 95% positiivsetest praimeritest omab seostumiskohti genoomis <= 30. Hiljem sooritati samad katsed ka kahe ülejäänud programmiga (BLAST2 ja IS), ning tulemused olid sarnased (*cutoff* ≈ 30) eelnevate meetoditega. Seega on määravaks meetodite vahelise paremuses programmi kiirus.

Arutelu

Tulemuste järgi otsustades olid olulisteks PCR kvaliteeti mõjutavateks parameetriteks parimeri 3' poole GC sisaldus, produkti pikkus ja seostumiste arvud genoomsel DNA'l (nii praimerite kui produktide korral). Võrreldes tüüpiliste levinud parameetritega leidsime, et oluline on vähendada GC sisaldust disainitavates PCR praimerites ning piirata produktide pikkusi. Sarnaselt Haas ja kolleegide tööle leiti, et praimeri pikkus ei ole otseses korrelatsioonis PCR väljatulemise kvaliteediga (Haas *et al*., 1998). Samas pole vaja otseselt muuta ka viimaste nukleotiidide sisaldust praimeri 3' otsas.

Sageli ei leita optimaalses parameetride vahemikus soovitavaid parameetreid ja tekib küsimus milliseid parameetreid oleks õige lõdvemaks lasta. Siis tuleks esimeses järjekorras laiendada just praimeri või produkti pikkuse parameetreid, praimeri 3' otsa viimaste nukleotiidide GC sisaldust (*GC_CLAMP*), sest antud parameetrid ei mõjutanud oluliselt PCR kvaliteeti. Alles seejärel võiks järjekordse negatiivse tulemuse korral laiendada üldist GC sisaldust, muuta Tm'i või seostumiste arvu *cutoff*'i genoomis.

Sulamistemperatuuri optimumi ei olnud meil võimalik määrata, kuna Tm'id olid labori poolt kindlalt fikseeritud ja polnud mõistlik tellida mitu erinevat praimerite komplekti. Beasley kaasautoritega kasutasid oma töös praimerite disainil Tm väärtusena 62 °C ja meie püüdsime valida praimereid just selle väärtuse läheduses (Beasley et al, 1999). Samas oleks erinevate Tm'ide korral parameetrite optimume ka keerulisem võrrelda ja leida.

Hoolimata genoomi-testi optimaalsete praimeri pikkuste *cutoff*'ide leidmisest, jäi siiski hulk negatiivseid praimereid "detekteerimata". Põhjusteks võivad olla vead labori eksperimentides ja agaroosgeeli piltide interpreteerimisel (vale-negatiivsed). Samuti annab tulemusi parandada GC sisalduse parameetri kitsendamisega piiridesse 20-50% ja produkti pikkuse piiramisega vahemikku 100-400 nukleotiidi. Siiski võib olla veel teisi tegureid, mis mõjutavad reaktsiooni käiku ning seega PCR kvaliteeti.

Meie töös kasustusel olevad genoomi-testi programmid põhinevad järjestuse identsuse otsingu algoritmidel, mis aga ei anna võibolla nii täpset tulemust või head korrelatsiooni PCR kvaliteedi ja praimerite seostumiste arvu vahel genoomis. Alternatiiv oleks *Nearest-Neighbour* algoritmi kasutamine, mis arvestaks praimeri ja DNA järjestuse vahelise identsuse asemel termodünaamilist aspekti. Näiteks seostub

lühem GC rikas ning pikem AT rohke praimer DNA'ga sama temperatuuri juures võrdse efektiivsusega.

Optimaalseimaks seostumiste otsingu sõnapikkuseks saime 16 nukleotiidi praimeri 3' otsast ja *cutoff*'ideks 30 seostumist summana. Kuigi 17 nukleotiidi andsid parema tulemuse eelistasime siiski 16, kuna 32 bit'istes arvutites käib andmete mälus hoidmine 4 baidi kaupa ja 16 tähe mälus hoidmiseks kulub täpselt 4 baiti. Seega on 16 sõnapikkusega otsingu sooritamine efektiivsem ja vähem ressursinõudlik kui 17 nukleotiidiste sõnade kasutamine.

Genoomi-testi alternatiiviks on maskeerida kordusjärjestused genoomses DNA's, et vältida juba disaini ajal praimerite sattumist korduspiirkondadesse. Maskeerimise käigus otsitakse üles korduspiirkonnad ja asendatakse need järjestused muu tähemärgiga (tavaliselt 'N'). Enim levinud selliste järjestuse genereerimiseks on REPEATMASKER. Kuid programmi puudusteks on esiteks aeglus, kuna ta kasutab järjestuste võrdlemiseks dünaamilisel programmeerimisel põhinevat *Smith-Waterman*'i algoritmi. Teiseks on maskeerimise tulemused suuresti sõltuvuses genoomi korduspiirkondade valikust (*RepBase* andmebaas – http://www.girinst.org/Repbase_Update.html) ja kolmandaks maskeerib REPEATMASKER korduses oleva ala terves pikkuses, kuigi see võib sisaldada tegelikkuses lühikesi "häid" praimeri disaini regioone (Jurka, 2000; Smit ja Green http://ftp.genome.washington.edu/RM/RepeatMasker.html). Alternatiivseks võimaluseks on kirjutada uus programm, mis koostaks nimekirja kõigist võimalikest 16 nukleotiidi pikkustest järjestustest inimese genoomis, mis teeb koguarvuks ≈ $4*10^9$. Erinevaid järjestusi, mis annavad >10 seostumiskoha genoomis, on kokku ~40 000 000. Nende vältimine juba praimerite disaini algstaadiumis parandaks praimerite kvaliteeti märkimisväärselt ja muudaks genoomitesti lihtsamaks (vajalik ainult produktide arvu kontroll).

Kokkuvõte

Käesolevas töös anti ülevaade enamlevinud PCR praimerite disaini programmidest ja eksperimentides (1278 praimeri disainil) kasutati laialt levinud ning akadeemilistele asutustele vabalt saadaval olevat programmi PRIMER3. Tulemuste analüüsis selgus, et PCR kvaliteedi tõstmisel on olulisteks parameetriteks praimeri 3' poole GC sisaldus, produktide pikkus. Soovitav vahemik GC% jaoks oleks 20-50 ja produktide pikkus 100-400 aluspaari. Märkimisväärset korrelatsiooni ei täheldatud nii praimeri pikkuse kui ka 3' otsa viimaste nukleotiidide GC sisalduse ja PCR kvaliteedi vahel. Seega tuleks praimerite mitteleidmisel esimesena suurendada just praimerite või PCR produkti pikkuse vahemikku ja praimeri 3' otsa (*GC_CLAMP*) väärtusi ja alles seejärel muuta praimerite üldist GC% või sulamistemperatuuri. Viimast parameetrit ei testitud, kuna labori poolt oli kindel Tm enne paika pandud.

Teiseks leiti, et praimerite seostumisel üle 10 korra 18 sõna pikkusega otsides ja üle 30 korra 16 sõna pikkuse juures langes PCR kvaliteet alla 80%. Sama juhtus ka 2 või enama PCR produkti tekkimise korral. Kuigi 17 nukleotiidi andsid parema tulemuse eelistati siiski 16, kuna 32 bit'istes arvutites käib andmete mälus hoidmine 4 baidi kaupa ja 16 tähe mälus hoidmiseks kulub täpselt 4 baiti. Seega on 16 sõnapikkusega otsingu sooritamine efektiivsem ja vähem ressursinõudlik kui 17 nukleotiidiste sõnade kasutamine. Genoomi-test suutis 100 negatiivsest praimerist välja selekteerida ~50%. Põhjuseks võimalikud vead eksperimentides või agaroosgeeli piltide interpreteerimisel. Lisaks võib vähendada nende vale-negatiivsete arvu veel GC protsentide ja PCR produkti pikkuste muutmine.

Seostumiskohtade otsingul oli 1-100 praimeri paari andmemahu juures kiireim MEGABLAST, kuid alates 100 paarist ületas kiiruselt teda SSAHA. Nelja meetodiga erinevate sõnapikkustega päringuid tehes (BLAST2, MEGABLAST, SSAHA ja *Perli*'s kirjutatud idenstuse otsingu programm) saadi kõigi puhul sarnased tulemused ja võrreldes negatiivsete praimerite eraldamisvõimet 4 meetodi juures, ei olnud märkimisväärset vahet meetodite juures märgata.

Edaspidi tuleks katsetada korduste raamatukogu lisamist maskeerides korduvad genoomsed DNA järjestused 'N' tähtedega 3' otsast. Üheks võimaluseks on kirjutada uus programm, mis koostaks nimekirja kõigist võimalikest 16 nukleotiidi pikkustest järjestustest inimese genoomis.

Summary

Single nucleotide polymorphism's (SNP's) are the most widely used markers for large genotyping projects. Mass-genotyping platforms now allow testing of up to 100 000 SNPs from each individual. However, most genotyping methods require the PCR amplification procedure. To ensure efficient genomic PCR from the human genome we need to test PCR primers against the genome for additional binding sites. Also, it is important to deselect such primer pairs that generate more than one product, because they could add false signals in genotyping analysis.

Most of the available primer design programs do not check the uniqueness of designed primers and products against the genomic DNA. Therefore, we decided to compare some methods that could be used for efficient detection of all primer binding sites in the human genome. We compared several sequence similarity search programs to optimise primer binding site detection: SSAHA (Ning *et al*., 2001), MEGABLAST (Zhang *et al*., 2000), BLAST2 (Altschul *et al*., 1997) and our own IDENTITY SEARCH (IS) program, written in *Perl*.

We used 714 experimentally tested PCR primers from a large genotyping 1,5K project in a cooperation with Sanger Centre (http://www.sanger.ac.uk/). Primers were designed with PRIMER3 (S. Rozen, H.J. Skaletsky, 1996-98). Repeat library of the PRIMER3 was not used in primer design process. For additional analysis, a positive and a negative test set (100 PCR primers each) was created based on observed PCR amplification quality. Primers in the positive ("good") dataset had amplification quality over 95%. Primers in the negative ("bad") dataset had amplification quality below 50%.

At first we have analyzed the PCR primer design parameters and noticed that parameters such as primer GC content and PCR product length affecting the PCR quality. GC% over 50 and product length over 400 bp gave quality below 80% (Fig. 6,9). Parameters like *GC_CLAMP* (last nucleotides at primer 3' end) (Fig. 7) and primer length (data not shown) did not affect PCR quality significantly (lengths 21-26 were a bit higher quality than shorter ones).

Then we have counted the number of binding sites and alternative product numbers in the genomic DNA for all primer pairs and counted the number of detected binding sites for each method (SSAHA, MEGABLAST, BLAST2, IS) at different lengths of primer 3' end. Figure 8 shows that primer with binding sites over 10 and product number over 1 are likely ot fail in PCR reaction.

Genome-test (binding sites search programs described above) methods were compared for their efficiency to discriminate between positive and negative datasets based on the number of primer binding sites in the genome (Fig. 10). Different lengths in range between 12 and 20 nucleotides were tested. We can compare these values to bitscores from the BLAST2 and assume that they are comparable to BLAST2 bitscores 24 - 40 (1 bp = 2 bits for two identical nucleotides). Optimal query lengths were between 16 - 18 nucleotides from primer 3' end for all methods.

After that every method was tested with different number of query sequences and total time was calculated from program start to end (Fig. 11). Total time includes our parsers work which gathered binding hits and counted possible PCR products. All tests were performed at Intel P3 machine with 700 MHz processor and 2 GB RAM. Graphic in Fig. 11 shows, that for smaller queries (1-10 primer pairs) MEGABLAST is faster than others. SSAHA is slower with small queries because it spends significant time for reading hash tables from the disk (total size 16GB). With large number of primers (100 or more) the efficiency of SSAHA and relative speed compared to others is rising.

Finally at Fig. 12, a typical distribution of both "good" and "bad" PCR primers according to the number of their binding sites is shown. MEGABLAST and SSAHA hits plotted against fraction of positive and negative samples percent. Primer pairs that have more than 30 binding sites in the genome (sum of left and right primer binding sites) are likely to fail in PCR. To deselect negative primers we can select cutoff with value 30 where 95 % positive primers have binding hits less than this cutoff. Other programs had similar distribution and cutoff values. There are still many negative primers that we cannot predict with *in silico* methods. Lab experiment errors and misinterpretation of PCR quality could be behind low quality of some primers in "bad" dataset. Also it might be possible to rise the PCR quality by lowering the GC and product number parameters and using the repeat library.

For conclusions – our optimal primer design parameters are: GC% 20-50, primer length 21-26 bp, product length 100-400 bp, best genome-test methods SSAHA (for datasets larger than 100 primer pairs) and MEGABLAST (for datasets between 1-100 pairs) with 3' primer length cutoffs 16.

Tänuavaldused

Tahaksin tänada kõigepealt oma juhendajat dr. Maido Remm'i, kelle näpunäideteta ja suunavate nõuanneteta poleks käesolev töö sellise tasemeni jõudnud. Teiseks soovin tänada ka oma teist juhendajat professor Andres Metspalu, kelle juhitud laboris antud magistritöö valmis.

Samuti tahaksin tänusõnad öelda dr. Jaak Vilo'le esimeste programmeerimis alaste algteadmiste edasiandmise eest ning lisaks labori kaaslastele nii TÜ MRI biotehnoloogia õppetoolis kui ka firmas AS Asper Biotech.

Lõpuks tänan veel oma armast abikaasat ja teisi sõpru toetuse ja igakülgse abi eest.

Kasutatud kirjandus

1. Aach J., Bulyk M.L., Church G.M., Comander J., Derti A., Shendure J. Computational comparison of two draft sequences of the human genome. Nature 409:856-59, 2001.

2. Agarwal P., States D.J. Comparative accuracy of methods for protein sequence similarity search. Bioinformatics 14(1):40-7, 1998.

3. Altschul S. Hot Papers In Bioinformatics The Scientist 13(8):15-16, 1999.

4. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool J Mol Biol 215:403-10, 1990.

5. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs Nucleic Acids Res. 25:3389-3402, 1997.

6. Arslan A.N., Egecioglu Ö., Pevzner P.A. A new approach to sequence comparison: normalized sequence alignment. Bioinformatics 17(4):327-37, 2001.

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

8. Borer P.N., Dengler B., Tinoco I.Jr. Stability of Ribonucleic acid Double-stranded helices. J. Mol. Biol. 86:843, 1974.

9. Chenal V., Souque P., Markovits A., Griffais R. Choosing higly specific primers for the polymerase chain reaction using octomer frequency disparity method: application to *Chlamydia trachomatis*. Gene 176:97-101, 1996.

10. Delcher A.L., Kasif S., Fleischmann R.D., Peterson J., White O., Salzberg S.L. Alignment of whole genomes. Nucleic Acids Res 27(11):2369-76, 1999.

11. Fortna A., Gardiner K. Genomic sequence analysis tools: a user's guide. Trends Genet 17(3):158-64, 2001.

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

13. Hartemink A.J., Gifford D.K., Khodor J. Automated constraint-based nucleotide sequence selection for DNA computation. Biosystems 52(1-3):227-35, 1999.

14. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16(9):418-20, 2000.

15. Kent J.W. BLAT – The BLAST-like Alignment Tool. Gen Research 12:656-664, 2002.

16. Korf I., Gish W. MPBLAST: improved BLAST performance with multiplexed queries. Bioinformatics 16(11):1052-53, 2000.

17. Kurata K., Nakamura H. Novel method for primer/probe design and sequence analysis. Genome Informatics 11, 331-32, 2000.

18. Kurg A., Tõnisson N., Georgiou I., Shumaker J., Tollett J., Metspalu A. Arrayed primer extension: solid.phase four-color DNA resequencing and mutation detection technology. Genetic Testing 4(1):1-7, 2000.

19. Kämpke T., Kieninger M., Mecklenburg M. Efficient primer design algorithms. Bioinformatics vol.17 3:214, 2001.

20. Landegren U., Nilsson M., Kwol P.-Y. Reading bits of genetic information: methods for single nucleotide polymorphism analysis. Gen Research 8:769-76, 1998.

21. Lander E.S. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921, 2001.

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

23. Ning Z., Cox A.J., Mullikin J.C. SAHA: A fast search method for large DNA databases. Gen Research 11:1725, 2001.

24. Pastinen T., Kurg A., Metspalu A., Peltonen L., Syvänen A.-C. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. Gen Research 7:606-14, 1997.

25. Pearson W. R. Flexible sequence similarity searching with the FASTA3 program package Methods in Molecular Biology, 1999.

26. Pearson W.R., Lipman D.J. Improved tools for biological sequence comparison. PNAS 85(8):2444-8, 1988.

27. Pertsemlidis A., Fondon J.W. 3rd. Having a BLAST with bioinformatics (and avoiding BLASTphemy). Genome Biol 2(10):1-10, 2001.

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

29. Proutski V., Holmes E.C. Primer Master: a new program for the design and analysis of PCR primers. Comput Appl Biosci 12(3):253-5, 1996.

30. Raddatz G., Dehio M., Meyer T.F., Dehio C. PrimeArray: genome-scale primer design for DNA-microarray construction. Bioinformatics 17(1):98-9, 2001.

31. Rouillard J.-M., Herbert C.J., Zuker M. OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics 18(3):486-87, 2002.

32. Rychlik W., Spencer W.J., Rhoads R.E. Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acids Res 11; 18(21): 6409-6412, 1990.

33. Schuler G.D. Sequence mapping by electronic PCR. Gen Research 7:541-50, 1997.

34. Smit A.F., Green P. RepeatMasker http://ftp.genome.washington.edu/RM/RepeatMasker.html

35. Smith T.F., Waterman M.S. Identification of common molecular subsequences. J Mol Biol 147(1):195-7, 1981.

36. Tatusova T.A., Madden T.L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 174(2):247-50, 1999.

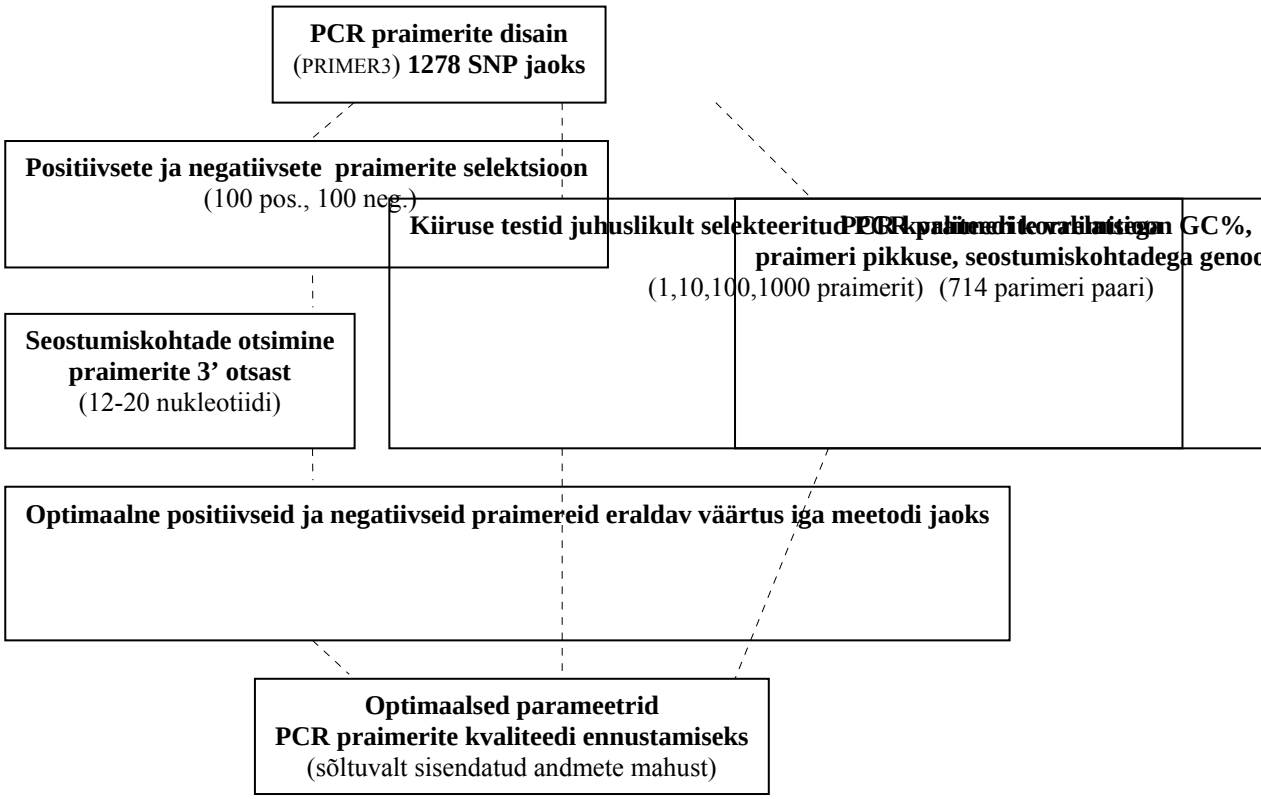
37. Tõnisson N., Zernant J., Kurg A., Pavel H., Slavin G., Roomere H., Meiel A., Hainaut P., Metspalu A. Evaluating the arrayed primer extension resequencing assay of TP53 tumor suppressor gene. PNAS 99(8):5503-8, 2002.

38. Varotto C., Richly E., Salamini F., Leister D. GST-PRIME: a genome wide primer design software for the generation of gene sequence tags. Nucleic Acids Res 29(21):4373-77, 2001.

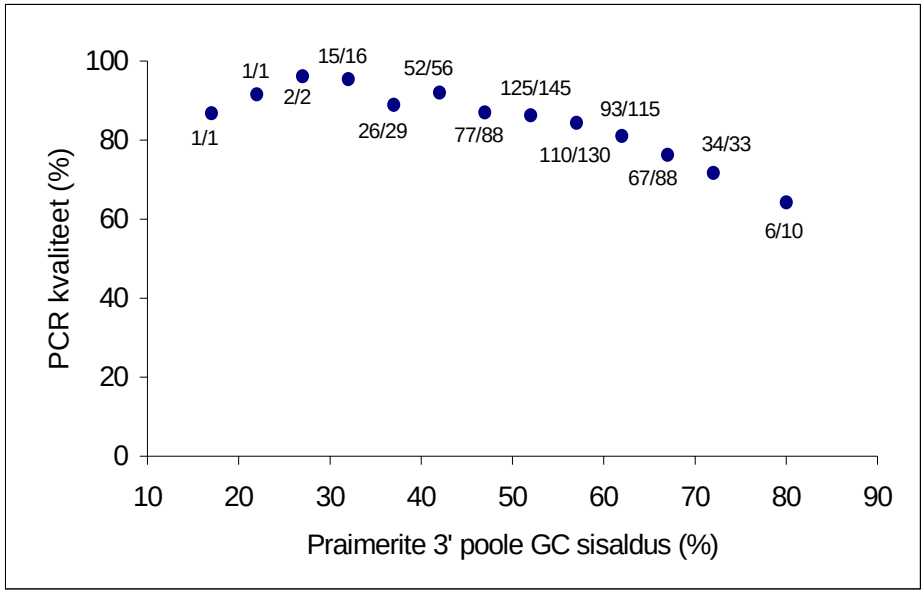
39. Wang D.G., Fan J.-B., Siao C.-J., *et al*. Large scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. Science 280:1077, 1998.

40. Zhang Z., Schwartz S., Wagner L., Miller W. A greedy algortihm for aligning DNA sequences. J Comp Biol 7(1/2):203-14, 2000.

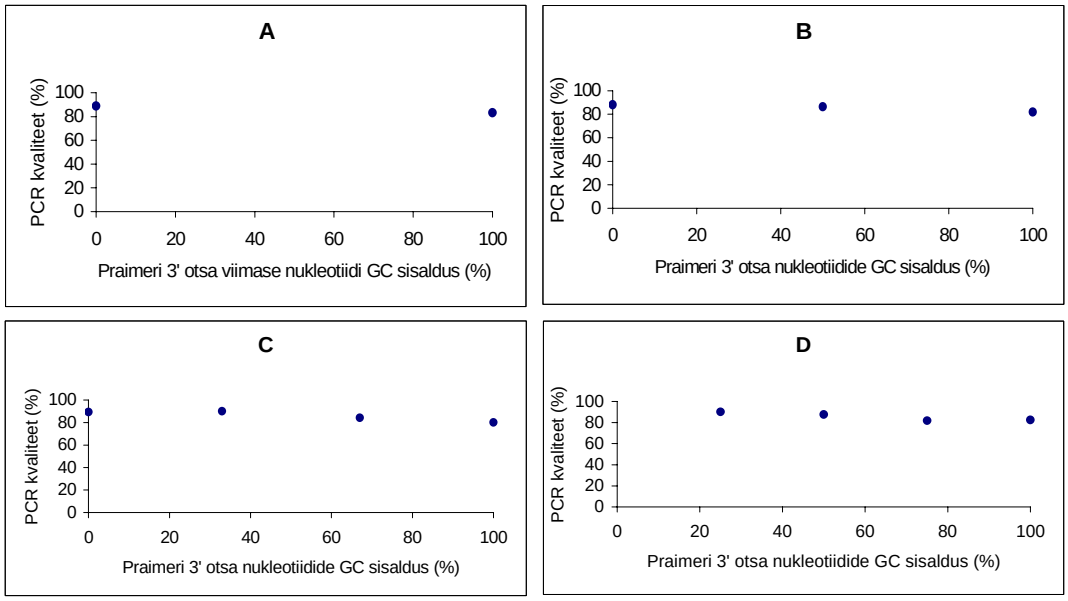
Joonised



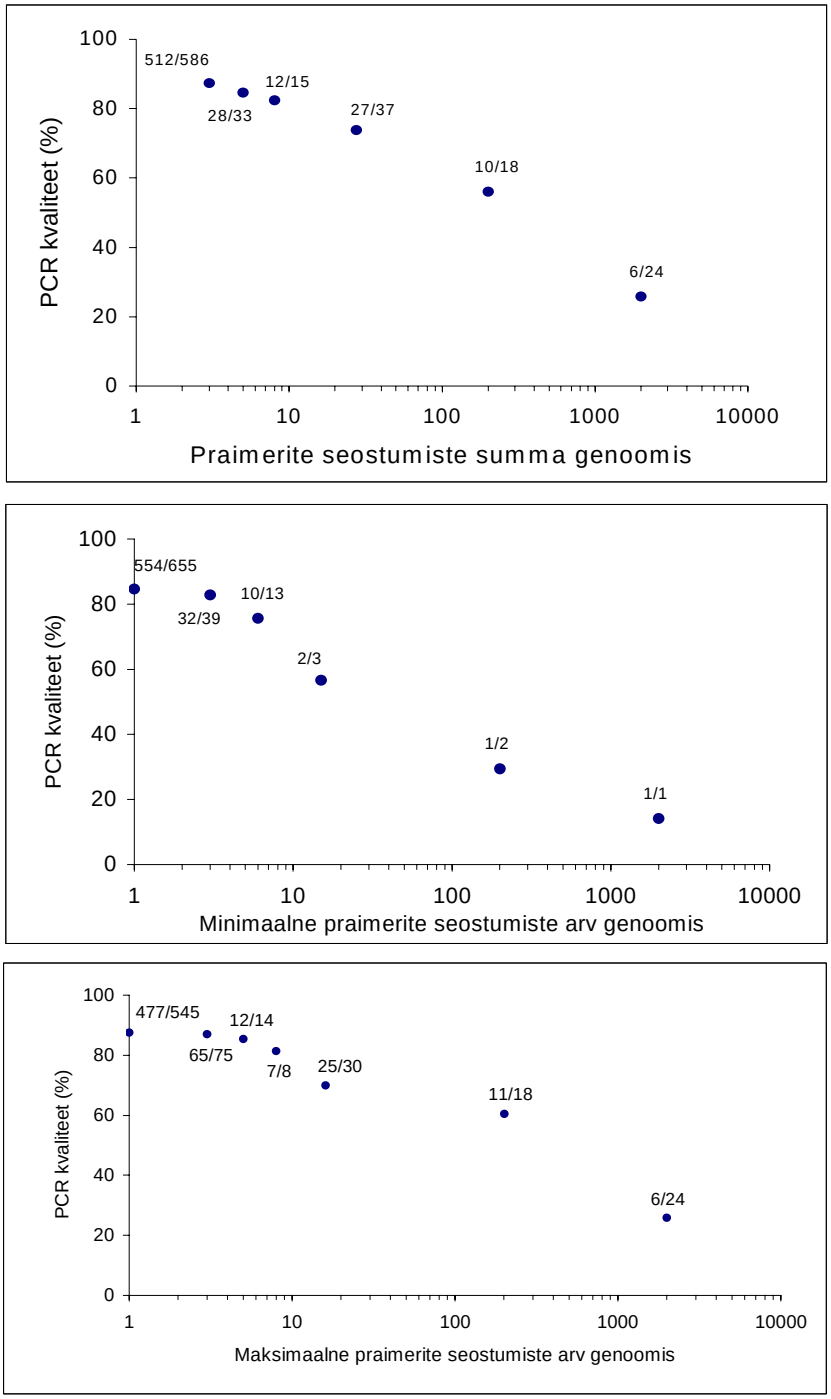
Joonis 5. Magistritöö eksperimentaalne üldskeem. Skeemil on kirjeldatud eksperimentide järjekorda, kus vasakpoolne rada kujutab endast positiivsete ja negatiivsete praimerite valimite (100 pos. ja 100 neg) jaoks optimaalse väärtuse leidmist iga genoomi testi programmiga. Parempoolne rada tähistab erinevate parameetrite võrdlust PCR kvaliteediga kõigi praimerite suhtes (714 paari) ja kiiruse testid juhuslikult valitud praimeri paaride gruppidega.



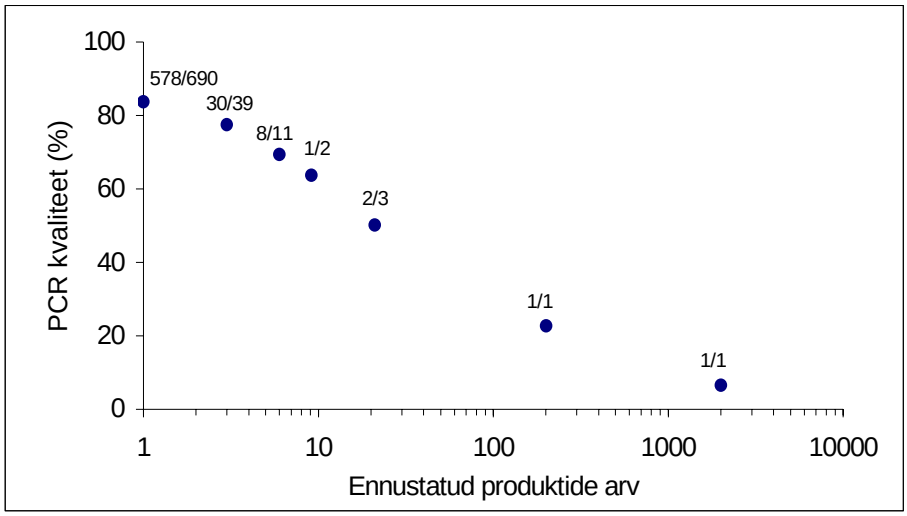
Joonis 6. Praimerite 3' poole GC sisalduse korrelatsioon PCR kvaliteediga. Kõigi praimerite korral arvutati nende 3' poole GC sisalduse protsent ja iga väärtuse kohta arvutati keskmine PCR kvaliteet. Üle 50% GC nukleotiidide sisaldus praimerites vähendab PCR kvaliteeti alla meie seatud *cutoff*'i – 80%.

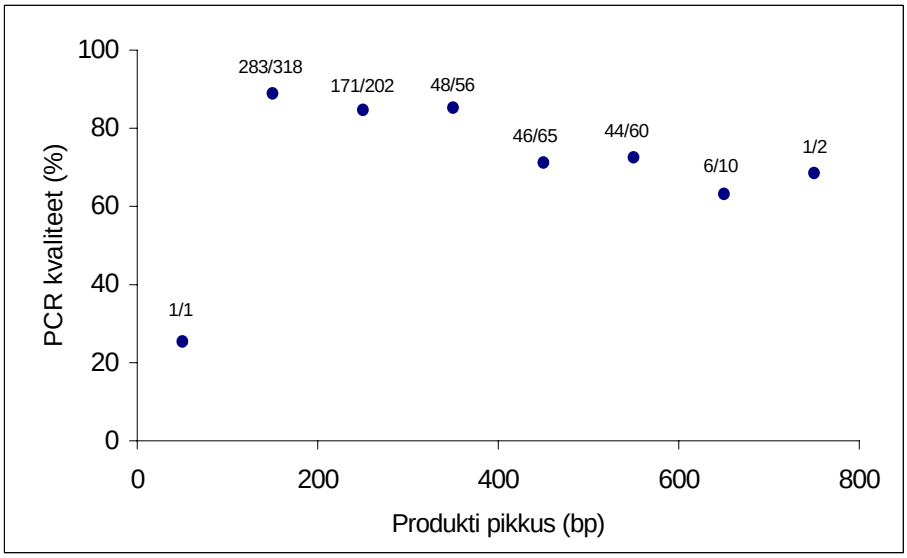


Joonis 7. Praimerite 3' otsa viimase 4 nukleotiidi GC sisalduse korrelatsioon PCR kvaliteediga. Joonisel on võrreldud praimerite viimase (A), 2 viimase (B), 3 viimase (C) ja 4 viimase (D) nukleotiidi GC sisalduse protsenti PCR kvaliteediga. Selgus, et praimerite 3' lõpu nukleotiidne koostis ei mõjuta oluliselt PCR kvaliteeti.

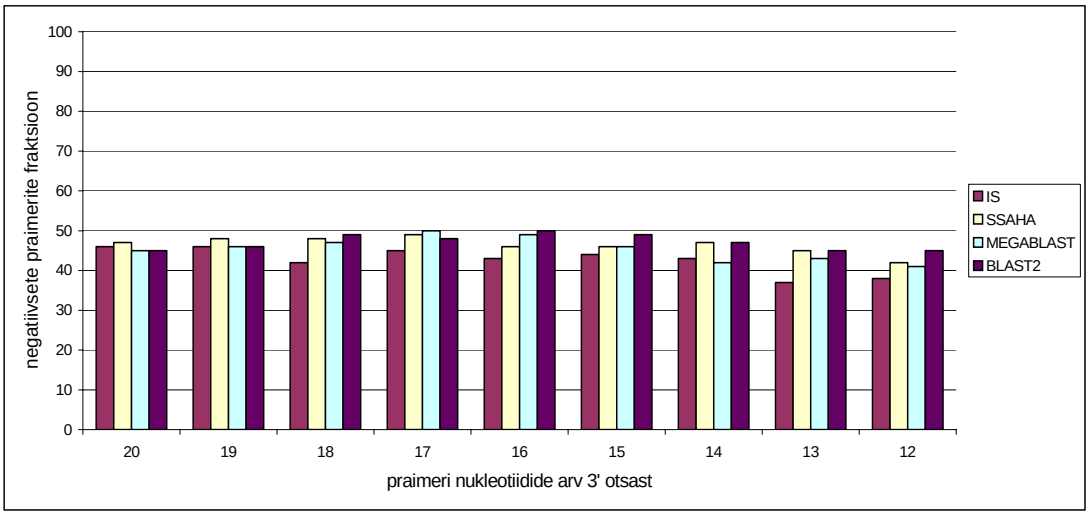


Joonis 8. PCR praimerite paaride genoomsete seostumiste summa, minimaalsete ja maksimaalsete seostumiste arvu korrelatsioon PCR kvaliteediga. Esimese graafiku X teljel on näidatud praimerite seostumiste arv summana genoomis, teisel minimaalne seostumiste arv ehk praimeri paari väiksema väärtusega seostumiste arv ja kolmandal maksimaalne seostumiste arv. Kui praimeri paari seostumiste summa või maksimaalne seostumiste arv ületab 10 ja/või minimaalne seostumiste väärtus on >3, siis langeb PCR kvaliteet alla 80 %.

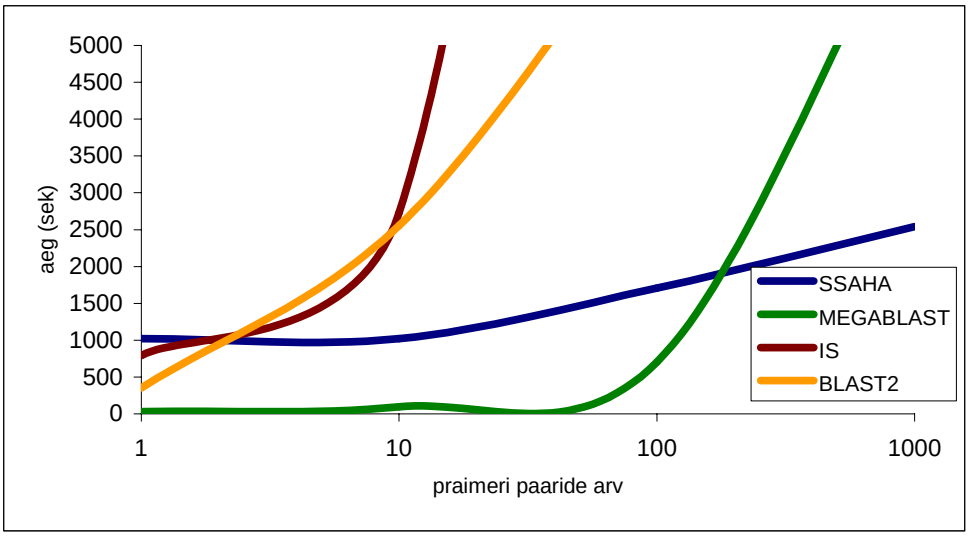




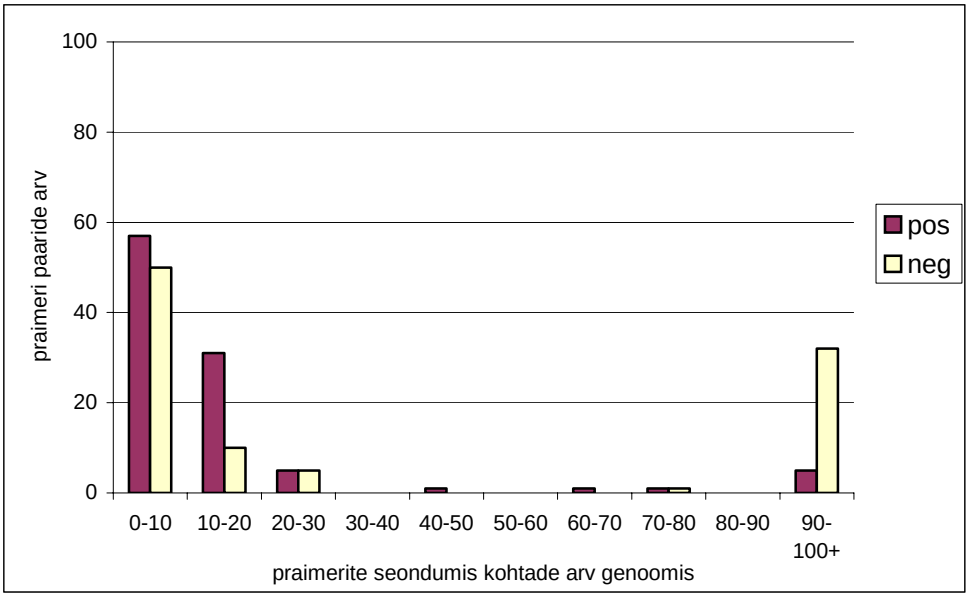
Joonis 9. Kõikide võimalike PCR produktide arvu ja pikkuste korrelatsioon PCR kvaliteediga. Kui ennustatavate produktide arv ületab 1 või 2, siis on täheldada PCR kvaliteedi langust alla 80 %. Sama langus tekib siis kui, produkti pikkus ületab 400 nukleotiidi.

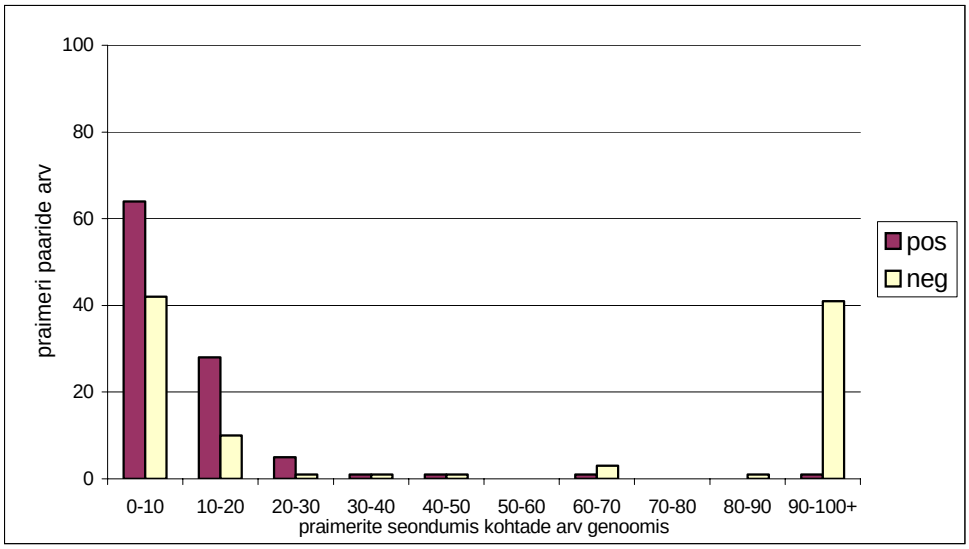


Joonis 10. 100 positiivse ja negatiivse praimeri paaride seostumiskohtade otsingu (genoomi-test) meetodite võrdlus erinevate 3' otsa pikkustega. Y telg näitab negatiivsete praimeri paaride arvu, mis detekteeriti vastava *cutoff*'i juures. *Cutoff*'iks defineeriti punkt, kus 95% positiivsetest praimeri paaridest omas arvuliselt vähem seostumiskohti genoomis kui punkti väärtus. X telje peal on nukleotiidide erinevad pikkused praimerite 3' otsast.



Joonis 11. Kiiruse test. 4 meetodi kiirust mõõdeti erinevate praimeri paaride arvu juures – 1, 10 , 100 ja 1000. Tööks kuluva aja sisse arvutati nii programmi enda tööaeg kui ka *parser*-programmide töö, mis filtreerisid välja 3' otsaga seostunud tulemused ja arvutasid võimalike produktide hulka.





Joonis 12. SSAHA (üleval) ja MEGABLAST (all) meetodite võrdlus optimaalse *cutoff*'i leidmiseks 16 nukleotiidse sõnapikkuse korral. 30 seostumiskohtade arvu juures on kaetud 95 % positiivsetest praimeri paaridest mõlema meetodi korral. Siiski on jäänud märkimisväärne hulk negatiivseid praimereid samuti selle *cutoff*'i piiresse, mida genoomi-test ei suuda välja selekteerida.