BIOLOOGIA-GEOGRAAFIATEADUSKOND
MOLEKULAAR – JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Triinu Kõressaar

Automaatne PCR praimerite disain kõikidele inimese geenidele

Keskastme lõputöö

Juhendaja: Maido Remm, Prof., PhD

TARTU
2003

Sisukord

Lühendid ja mõisted ... 3
Sissejuhatus ... 4
I Teoreetiline osa ... 5
1.Genoomide annoteerimine ja andmebaasid ... 5
2.Geenide ennustamine ... 6
2.1 Homoloogia kaudu ennustus ... 7
2.1.1 EST-ide baasil ennustus ... 7
2.1.2 Valkude baasil ennustus ... 8
2.1.3. DNA baasil ennustus ... 8
2.1.4 mRNA baasil ennustus ... 8
2.2 *Ab initio* geeniennustus ... 9
2.2.1 Signaali ennustamine ... 9
2.2.2 Järjestuste omaduste ennustamine ... 10
2.2.3 Kanditaateksonite ennustamine ... 10
2.2.4 Genscan, FGENES ja GenomeScan ... 11
2.3 Ensembli geeniennustus-süsteem ... 12
2.4 Automaatset ennustust teostavate meetodite täpsus ... 12
3. Kordusjärjestused ja genoomi maskeerimine ... 13
3.1 DUST ... 14
3.2 RepeatMasker ... 14
3.3 GenomeMasker ... 15
4. PCR primerite disain ... 15
4.1 Olulised PCR parameetrid ... 16
4.2 Praimerite seostumiskohtade kontroll ... 17
4.3 Praimerite disaini programmid ... 18
5. Ülesande püstitus ... 18
II Praktiline osa ... 19
Töö eesmärgid ... 19
Meetodid ... 19
1. Andmete päritolu ja struktuur ... 19
2. Kasutatud programmide kirjeldus ... 19
3. Vajalik mälumaht ja tööaeg ... 20
Tulemused ... 22
1. Eksonite kirjeldus ... 22
2. Programmide tööpõhimõte ... 23
2.1. Amplikonide konstrueerimine ... 25
2.2. Praimerite konstrueerimine amplikonidele ... 25
2.3. Praimerite seostumiskohtade unikaalsuse kontroll genoomis ... 26
3. Tulemuste kirjeldus ... 27
4. Andmebaasi loomine ... 31
5. Kasutajaliidese loomine ... 32
Arutelu ... 34
Kokkuvõte ... 36
Kasutatud kirjandus ... 37

Lühendid ja mõisted

PCR -	polümeraasi ahelreaktsioon (*Polymerase Chain Reaction*)
bp -	aluspaar
cDNA -	mRNA-lt pöördtranskriptaasi abil sünteesitud mRNA-ga komplementaarne DNA (*complementary DNA*)
EST -	mRNA-le komplementaarne lühike DNA järjestus (*expressed sequence tags*)
ab initio geeniennustus -	DNA nukleotiidse järjestuse andmetel põhinev geeniennustus
CDS -	eksoni kodeerivad alad, mis ei sisalda mittetransleeritavaid regioone (*coding sequence*)
UTR -	eksoni mitteransleeritavad alad (*untranslated region*)
Kodeerivad Statistikud -	geeni võimaliku lokalisatsiooni indikaatorid (*coding statistic*)
Sensitiivsus -	suurus, mis kajastab algoritmi võimet leida "õigeid" positiivseid (Sn, *sensitivity*)
Spetsiifilisus -	suurus, mis kajastab algoritmi ettevaatlikkust "vale" positiivsete leidmisel (Sp, *specificity*)
Flanking piirkond -	amplikonidega külgnev DNA järjestus, kuhu konstrueeritakse praimerid
Mb -	megabait
Amplikon -	ühe PCR-i produktina amplifitseeritav DNA järjestus
Kromosoomi bänd -	kromosoomiala, mis kromosoomi värvimisel naaberaladest selgelt eristub, olles neist heledam või tumedam

Sissejuhatus

Kuna saadaval on suur hulk erinevate genoomide järjestusandmeid, siis on vajadus automaatset järjestusanalüüsi teostavate vahendite järele möödapääsmatu. Kuigi tänaseks on välja töötatud palju erinevaid algoritme ja koostatud programme, et muuta võimalikuks genoomi automaatne annotatsioon, on siiski veel jätkuvalt mitmeid probleeme ja puudujääke järjestuste automaatsel analüüsimisel. Seepärast toimub jätkuv automatiseerimise arendamine.

Ulatuslik inimese genoomse DNA järjestuse informatsiooni akumuleerumine võimaldab suuremahuliste eksperimentide läbiviimist. Organismide jaoks, kellel on geenistruktuur identifitseeritud on võimalik huvipakkuvad geenid otse PCR-i abil üles amplifitseerida genoomselt DNA-lt tagades nii ka cDNA raamatukogudes mitte-esinevate geenide paljundamise. Saavutamaks geenide amplifitseerimist on relevantse tähtsusega tasuta sobivate praimerite süstemaatilist disainimist teostavate tarkvarade kättesaadavus (Varotto et al., 2001).

I Teoreetiline osa

1.Genoomide annoteerimine ja andmebaasid

Jätkuvalt toimub geenide ja teiste funktsionaalsete kodeerivate ja mittekodeerivate alade nii automaatne kui ka manuaalne annotatsioon. Genoomide annoteerimine on kiirelt arenev ala tänu tormiliselt genereeritavatele mahukatele genoomi järjestuste andmetele ja pidevalt täiustuvale arvutitehnikale. Kuigi manuaalselt annoteeritud järjestuste täpsust loetakse kõige usaldusväärsemaks, pole siiski reaalselt mõeldav erinevate organismide DNA järjestuste käsitsi analüüsimine. Inimese genoomi jaoks on olemas manuaalsed annoteerimistulemused ainult 14., 20. ning 22. kromosoomi jaoks. Suuremahulist manuaalset annoteerimist on läbi viinud vaid vähesed töögrupid. Märkimist väärivad siinkohal Dr. Dunham'i grupp, Wellcome Trust Sangeri Instituudis (22. kromosoomi annotatsioon), HAVANA group, Wellcome Trust Sangeri Instituudis (20. kromosoom), Genoscope, CNRS (14. kromosoom). Saadud andmed on kokku pandud VEGA-nimelisse (The Vertebrate Genome Annotation) andmebaasi, mis on loodud tsentraalseks panipaigaks kõikidele manuaalselt annoteeritud selgroogsete genoomide järjestustele (http://vega.sanger.ac.uk/). Lisaks VEGA andmebaasile on veel tähelepanuväärne ENSEMBL andmebaas, mis võimaldab kätte saada erinevate järjestuste ja genoomide parimaid võimalikke automaatse annotatsiooni andmeid (http://www.ensembl.org/). ENSEMBL andmebaasi projekt pakub erinevate järjestuste mitmekesist analüüsi ja järjestuste ning andmete erinevaid visuaalseid vaateid. Andmebaasi laiendatakse pidevalt vertikaalse ja horisontaalse andmesisestuse teel. Vertikaalne andmesisestus hõlmab erinevate andmetüüpide integreerimist andmebaasi koosseisu (ka manuaalselt annoteeritud geenide struktuurid, mis on saadud välistest allikatest) ja horisontaalne haarab erinevate võrdlevate genoomijärjestuste (mitmesuguste selgroogsete genoomid) vaadete loomist (Clamp et al., 2003). Automaatseid annoteerimistulemusi pakub ka NCBI (National Center for Biotechnology Information) RefSeq andmebaas (http://www.ncbi.nlm.nih.gov/RefSeq/), kus annoteerimisel on kasutatud väga palju erinevaid lahendusi (http://www.ncbi.nlm.nih.gov/genome/guide/build.html), (Pruitt et al., 2001).

Genoomide automaatse annoteerimise efektiivsus sõltub arvutisüsteemide võimsusest, mis on kokku pandud sekveneerimisprojektide jaoks, et suurte

andmehulkadega toime tulla (Lewis et al., 2000). Samuti on oluline automaatset annotatsiooni läbiviivate programmide ja algoritmide töötamise korrektsus, kuna genoomijärjestuse väärtus on üks-üheses seoses annoteerimisvahendite kvaliteedi tasemega. The Genome Annotation Assessment Project (GASP) viidi läbi, et uurida automaatset ennustust realiseerivate vahendite täpsust ja vastavalt tulemustest lähtuvalt parandada olemasolevaid tarkvaravahendeid ning leida uusi lahendusi efektiivsemaks automaatseks analüüsiks. Uuringu tulemusena jõuti järeldusele, et praegused geeniennustuse meetodid on märkimisväärselt paranenud (Reese et al., 2000).

Annoteerimisprotsessi eesmärk on jäikadele genoomijärjestustele omistada võimalikult palju infot. See hõlmab nii geeni lokalisatsiooni määramist genoomis kui ka geenistruktuuri paika panemist, erinevate markerite ja regulatoorsete üksuste identifitseerimist jpm. Viimase saavutamiseks on kasutusele võetud erinevaid lähenemisi; *ab initio* geeniennustus, homoloogiaotsingud nii samast kui ka teistest organismidest teadaolevate geenijärjestustega, mRNA osalise pikkuse või täispikkuse joondamine genoomsele DNA-le, motiivide ja signaalide detekteerimine ja erinevate lähenemiste kombineerimine. Annoteerimise kõrgem eesmärk on identifitseeritud geenide funktsionaalne klassifitseerimine, mis käesoleval ajal sõltub enamasti teadaoleva funktsiooniga geenide homoloogia leidmisest (Reese et al., 2000).

2.Geenide ennustamine

Geenide ennustamine on kõige tähtsam osa genoomide annoteerimisel. Inimese genoomi sekveneerimise peamine põhjus on määrata täielikult eksisteerivate geenide komplekt. Inimese genoomijärjestuse teadmine ei tähenda automaatselt transkriptoomi (genoomi poolt kodeerivate transkriptide kogumi) tundmist (Burge et al., 2001). Seepärast on välja töötatud palju erinevaid lahendusi transkripte kodeerivate alade, so geenide struktuuri ja lokalisatsiooni määramiseks. Geenide ennustamisel kasutatakse jätkuvalt häid tulemusi andvat eksperimentaalset lähenemist, kuid isegi kui ühel päeval kõik inimese geenid saadakse sel viisil identifitseeritud, on siiski oluline mõista geenistruktuuri organisatsiooni ja, kuidas põhimõtteliselt geene ära tunda. Seepärast on väga oluline produkte kodeerivate alade automaatset ennustust võimaldavate meetodite välja töötamine. Kuigi erinevaid geenileidmisprogramme on paralleelselt genoomi

andmete laekumisega rohkelt tekkinud, on siiski käesoleval ajal geeniennustamise täpsus kaugel rahuldavast (Zhang et al., 2002). Viimast eelkõige genoomi organisatsiooni suure varieeruvuse tõttu, sest raske kui mitte võimatu on üheselt määrata geenistruktuuri. Näiteks geenide ekspresseerumisel ja identifitseerimisel üliolulist rolli mängivate regulatoorsete üksuste määramine on nende regioonide paljususe (transkriptsiooni alguskoht, võimendajad, vaigistajad) ja geeni suhtes erinevate võimalike asetuste (eespool või tagapool transkriptsiooni alguskohta) tõttu äärmiselt komplitseeritud.

Automaatsel ennustamisel võib eristada homoloogia kaudu ennustust ja *ab initio* geeniennustust. Praktikas kasutatakse parima tulemuse saamiseks nende kahe lähenemise kombineerimist.

2.1 Homoloogia kaudu ennustus

Järjestuste sarnasuste otsingutel põhinevad programmid, mis baseeruvad järjestuste vaheliste konserveerunud alade võrdlusel, on heal tasemel ja efektiivsed geeniavastamise vahendid. Kahe järjestuse märkimisväärne sarnasus viitab sellele, et nad on homoloogsed ja omavad ühte evolutsioonilist esivanemat ja seega ka sarnast funktsiooni (Rogic et al., 2001).

Homoloogia kaudu ennustamisel võib välja tuua järgmised kasutatavate andmete poolest erinevad lähenemised: EST-ide, mRNA, sama liigi valkude ja erinevatest liikidest pärit valkude baasil homoloogia otsingud ja ka otseselt genoomide (DNA) vahelisel võrdlusel põhinevad meetodid.

2.1.1 EST-ide baasil ennustus

EST-ide otsese joondamise abil genoomile on tuvastatud tuhandeid inimese geene. Kuigi see meetod on ennast edukalt tõestanud, on sellel olulised puudujäägid. Peamine limitatsioon seisneb selles, et meetod põhineb põhiliselt stohhastiliselt detekteeritud transkripide järjestustel, mille kättesaadavus sõltub suuresti vastava geeni ekspressiooni tasemest. Seepärast on kõrgelt ekspresseeruvad geenid tihti üleesindatud EST-ide andmebaasides, kusjuures geenid, mis ekspresseeruvad madala tasemega või vähestes kudedes võivad puududa (Burge et al., 2001). Kasutades EST-e saab väga

kiiresti ja lihtsalt teada uuritava tundmatu järjestuse kohta seda, kas teda transkribeeritakse, kuid nagu eelpool mainitud, suudab anda vaid vihjeid geenistruktuuri kohta (Rogic et al., 2001).

2.1.2 Valkude baasil ennustus

Kõige tavapärasemalt kasutatakse lokaalsel joondamisel põhinevaid algoritme, et identifitseerida võimalikke kodeerivaid piirkondi. Näiteks BLASTX, mis kasutab sarnasust genoomsest järjestusest oletatavast geeni struktuurist kontseptuaalselt transleeritud valgu ja teadaoleva valgu järjestuse vahel. BLASTX-stiili sarnasust teadaoleva valguga on kombineeritud globaalse joondamisega GeneWise algoritmis, et täpsemalt määrata eksonite omavahelisi seoseid (Burge et al., 2001).

2.1.3. DNA baasil ennustus

Mitmed meetodid kasutavad genoomide vahelisel sarnasusel baseeruvat ennustust, kus eeldatakse, et genoomi kodeerivad alad on evolutsiooniliselt konserveerunumad kui mittekodeerivad regioonid. Mistõttu järeldatakse, et erinevate liikide vahelised homoloogsed piirkonnad on kodeerivad alad.. Sellised on näiteks ExoFish, mis kasutab TBLASTX homoloogial põhinevat võrdlust kerakala (*Tetraodon nigroviridis*) ja inimese genoomijärjestuse vahel, et leida konserveerunud kodeerivaid regioone ja GLASS/ROSETTA, mis kombineerib inimese ja hiire genoomsete järjestuste globaalse joondamise dünaamilise programmeerimisega, et saada selgust õigetest geenistruktuuridest mõlemas organismis (Burge et al., 2001).

2.1.4 mRNA baasil ennustus

Programmide idee, mis baseeruvad mRNA järjestustel, seisneb selles, et mRNA sisaldab otse geenilt kodeeritud komplementaarset transleeritavat järjestust, mida on lihtne uuritavalt genoomilt homoloogiaotsinguid kasutades üles leida ja üheks terviklikuks geeniks kombineerida. Meetodi suurteks puudusteks on rakkudest saadud mittetäielik mRNA-de komplekt, kuna paljud geenid avalduvad koespetsiifiliselt ja osad on alaekspresseeritud, nii et neid pole võimalik detekteerida. Lisaks sellele on

täheldatud ka eukarüootsete geenide kõrget alternatiivse splaissumise taset, see tähendab, et ühes geenis sisalduvaid eksoneid on võimalik omavahel erinevalt küpseks mRNA-ks kombineerida. Teada on ka teisi raskendavaid juhtumeid, näiteks üks geen on lokaliseerunud teise geeni intronisse. Sellised geenide iseärasused vähendavad oluliselt ainult mRNA-del tuginevate programmide täpsust.

2.2 *Ab initio* geeniennustus

Kuigi homoloogial põhinevaid meetodeid loetakse efektiivseteks, on näidanud hiljutine 22. kromosoomi reannoteerimine, et vaid 50% ennustatud valkudest omavad homoloogiat seniajani teadaolevate valkudega (Rogic et al., 2003). Seega on väga oluline homoloogiast sõltumatute meetodite juurutamine. *Ab initio* geeniennustus põhineb eelkõige genoomi DNA järjestusel. Oluline eeldus korrektseks ennustuseks on see, et genoomijärjestus oleks nukleotiidi tasemel täpselt määratud. Need meetodid pole otseselt sõltuvad geenide ekspressiooni tasemest ja suudavad efektiivsemalt ennustada harva ekspresseeruvaid geene (Burge et al., 2001).

Eukarüootses geenis saab eristada struktuurseid üksusi ja signaale, nagu intronid, eksonid, promooteri piirkond, splaissaidid jpt, kusjuures igaühte neist iseloomustab mingi iseäralik omadus. Sellest tulenevalt on välja töötatud ka erinevad algoritmid erinevate geeni osade ennustamiseks. *Ab initio* geeniennustuse võib jagada olulisemateks etappideks, seejuures iga etapi läbimiseks on välja töötatud erinevad algoritmid, mis olenevalt ennustatavast geeni osast panevad rõhku erinevatele momentidele. Nii võib eraldi välja tuua signaalide, järjestuse erinevate omaduste ja kanditaateksonite ennustust ning eksonite järjekorra leidmist teostavad programmid või algoritmid.

2.2.1 Signaali ennustamine

Signaale ennustavad meetodid püüavad leida mõne nukleotiidi pikkusi fragmente, mis tuntakse ära raku masinavärgi poolt ja on olulised kindlate protsesside initsiatsioonil. Geene-leidvate programmide poolt tavaliselt modelleeritavad signaalid on promooteri elemendid, start- ja stopkoodonid, splaissaidid ja polü-(A) saidid. Kuna genoomijärjestus on niivõrd varieeruv, siis ei saa ka kirjeldatud signaalide leidmisel

järjestustest neile väga suurt kaalu omistada, sest tegelikult pole võimalik eristada päriselt funktsioneerivaid signaale signaale mimikriseeritavatest signaalidest (Rogic et al., 2001).

2.2.2 Järjestuste omaduste ennustamine

Järjestuste erinevate omadustega arvestavad meetodid võtavad kasutusele nn kodeerivad statistikud (*coding statistic*), mis reageerivad erinevalt kodeerivatele ja mittekodeerivatele aladele. Kodeerivaid statistikuid on nimetatud ka valke kodeerivate alade indikaatoriteks. Tuntuimaks ja ka tõhusaimaks statistikuks loetakse spetsiifiliste heksameeride esinemist teatud lugemisraamides (Fickett et al., 1993).

2.2.3 Kanditaateksonite ennustamine

Eukarüootse geeni eksonid võib tinglikult ennustamise seisukohast jaotada nelja klassi: geeni 5’ ja 3 ‘ eksonid, geeni sisemised eksonid (iteksonid, *internal translated exons*) ja introniteta geenid, mis sisaldavad vaid ühte eksonit. Kuna selgroogsete geenid sisaldavad palju eksoneid ja introneid, siis on ka enamus programme fokuseeritud just iteksonite ennustamisele. Oluline on mainida, et mugavuse huvides mõeldakse eksonite mõiste all siinkohal tegelikult eksonite kodeerivaid alasid (CDS, *coding sequence*), kust on välja arvatud mittetransleeritavad piirkonnad (UTR, *untranslated region*). Eksoneid ennustavates programmides püütakse arvestada eksonite erineva pikkusejaotusega geeni erinevates osades, splaissaitidega, polü-(A) signaalidega, promooteri piirkonnaga, et täpsemalt identifitseerida erinevatesse klassidesse kuuluvaid eksoneid. Kui kanditaateksonid on ennustatud, siis kasutades kas dünaamilist programmeerimist (DP), nt programm FGENES (Solovyev et al., 1995) või tõenäosuslikke mudeleid (HMM), nt programm Genscan (Burge and Karlin, 1997), et assambleerida kokku õige geenistruktuur.

2.2.4 Genscan, FGENES ja GenomeScan

Tuntud *ab initio* geeniennustust läbiviiv programm on Genscan, mis ennustab eksonite (st kodeerivaid alasid ehk CDS-e) ja geenide lokalisatsiooni genoomsel järjestusel võttes arvesse erinevat eksonite ja intoronite ülesehitust (ka eelpool nimetatud kodeerivaid statistikuid) ja modelleerib RNA splaissimisel olulised signaalid. Programm põhineb HMM-il (Hidden Markov Model), kus erinevatele seisunditele (välimised eksonid, sisemised eksonid, promooteri piirkond, intergeensed alad, mittetransleeritavad alad jt) ja transitsioonidele (ühest seisundist teise minemine) pannakse täielikult tõenäosuslikke skoore. Kuna Genscan-i analüüs ei sõltu ekspressiooni tasemest ega homoloogsete geenide või proteiinide olemasolust, siis seda saab rakendada igale sekveneeritud genoomi regioonile (Burge et al., 2001).

Võrdväärse täpsusega ennustab geene ka FGENES, mis kasutab dünaamilist programmeerimist, et ennustatud eksoneid, promootereid ja polü(A)-saite optimaalseks geeniks kokku kombineerida. Programm konstrueerib antud järjestuselt geenide kogumiku, arvutab igale kanditaatgeenile kindlate kriteeriumite alusel skoori ja suurima skoori väärtuse saanud kanditaatgeen loetakse optimaalseimaks geeniks. Programm on väga paindlik, mistõttu võimaldab ennustada osaliselt või täielikult üksikuid ja tandeemselt korduvaid geene (Rogic et al., 2001).

GenomeScan (http://genes.mit.edu/genomescan/) algoritmi aluseks on eelnevalt välja töötatud ekson-intron ja splaissumissignaalide konsensusmudel, mis on moodustatud teadaolevate geenistruktuuride põhjal. GenomeScani laialdane testimine näitab, et programm annab vähe “vale” positiivseid tulemusi ja identifitseerib täpselt ekson-intron struktuuri geenis (Yeh et al., 2001). Programmi kasutatakse ka NCBI RefSeq andmebaasis olevate geenide ennustamiseks (Pruitt et al., 2001).

Kuigi kättesaadavad on paljud erinevad geenieennustuse programmid, pole nad võimelised õigesti või üldse mitte ennustama regulatoorseid elemente, polü-(A) saite, ei arvesta 5’ ja 3’ mittetransleerivate aladega, alternatiivse splaissimisega ja ei detekteeri ülekattuvaid või üksteisega põimuvaid geene.

2.3 Ensembli geeniennustus-süsteem

Ensembli geenistruktuuri määramise süsteem hõlmab erinevaid meetodeid – *ab initio* geeniennustus, homoloogia otsingud, ka manuaalsete annoteerimistulemustega arvestamine.

Ensembli andmebaasi projekt kasutab kolmeastmelist protsessi geenide asetamiseks genoomile. Esiteks leitakse kõik teadaolevad inimese valgud SPTREMBL-ist, kasutatakse "fast protein to DNA matcher'it" (pmatch). Järgnevalt teostatakse lihvimine programmiga GeneWise, et tagada õige geenistruktuur, arvestatakse ka mittetransleeritavate alade ehk UTR-idega. Saadud geene kutsutakse teadaolevateks geenideks (*known genes*). Teiseks teostatakse analoogne protsess eelmisega, kuid kasutatakse paraloogseid inimese valke ja valke teistest organismidest, et ennustada uudseid geene (*novel genes*). Kolmanda võimalusena annoteerimaks geene kasutatakse *ab initio* programmi GenScan, mida jooksutatakse üle genoomi, et luua genscan peptiidide komplekt. Seejärel assambleeritakse geeniks need eksonid, millele leitakse homoloog BLASTi otsingutega valkude, selgroogsete mRNA-de ja UniGene clusterite vastu. Selliselt saadud geenid kannavad nimetust genescan geenid (*genescan genes*). "Ensembl core" andmebaasi geene peetakse vigadeta ennustatud geenideks, kuna neid toetavad mitmed erinevad eksperimentaalsed katsed (Hubbard et al., 2002).

2.4 Automaatset ennustust teostavate meetodite täpsus

Kuna *ab initio* geeniennustusmeetodites kasutatakse kaudseid tõendeid geenistruktuuri määramiseks, mis on tuletatud statistiliselt, siis on tähtis, et arvuti abil ennustatud geenid verifitseeritaks mingisuguste eksperimentaalsete katsetega (Burge et al., 2001). Samuti tuleb ennustusalgoritme treenida nn training data set abil, et kohandada programme bioloogiliste andmete iseärasustele. Kasutatakse nii positiivseid training set'te, mis on "õiged objektid" nagu näiteks eksonid ja negatiivseid training set'te, mis on "valed objektid" nagu pseudogeenid ja suvalised mittekodeerivad alad.

Programmide usaldatavuse hindamiseks kasutatakse enamasti kahte väärtust, need on sensitiivsus (Sn, *sensitivity*) ja spetsiifilisus (Sp, *specificity*). Sensitiivsus on suurus, mis näitab algoritmi võimet leida objekte, mis on tõesti olemas ehk "õigete" positiivsete (TP, *true positives*) leidmise võimet.

$$Sn=TP/(TP+FN),$$

kus FN tähistab “vale” negatiivseid (*false negatives*)

Spetsiifilisus on suurus, mis näitab algoritmi ettevaatlikkust “vale” positiivsete (FP, *false positives*) leidmisel.

$$Sp=TP/(TP+FP)$$

Antud suuruste väärtused kõiguvad vahemikus [0,1], mida enam läheneb suurus väärtusele üks, seda täpsem on programm. Täiuslikult ennustava programmiga on tegu juhul, kui mõlemad suurused on võrdsed ühega. Praktikas on näidatud, et kui muuta parameetreid, mis suurendavad spetsiifilisust, siis väheneb selle arvelt programmi sensitiivsus ja vastupidi (Collins et al., 2003).

3. Kordusjärjestused ja genoomi maskeerimine

Inimese genoomis on umbes 60% mittekorduvat, 30% mõõdukalt korduvat ja 10% kõrgelt korduvat DNA-d. Mõõdukalt korduv DNA kujutab endast suhteliselt lühikesi järjestusi, mida on korratud 10-1000 korda ja kõrgelt korduv DNA esineb väga lühikeste järjestustena, mida esineb tuhandetes kordades inimese genoomis. Kordusjärjestuste leidmine on oluline järjestuste assambleerimisel, homoloogiaotsingutes, PCR praimeite valimisel, isiku tuvastamisel (mikrosatelliidid), genoomide evolutsiooni uurimisel jpt.

Enamus korduvast DNA-st asetseb ekstrageenses alas, mida on 75% inimese genoomist. Eristatakse tandeemselt korduvat mittekodeerivat DNA-d (ka võib tinglikult eristada tandeemselt korduvat kodeerivat DNA-d, milleks on tandeemsed geenid, nt rRNA geenid ja histoone kodeerivad geenid) ja hajuskordusjärjestusi. Esimesena mainitud võib jaotada nelja klassi: megasatelliidid, satelliidid, minisatelliidid (VNTR – *variable number of tandem repeat*) ja mikrosatelliidid (STR – *simple tandem repeat*). Kui megasatelliidid ja mikrosatelliidid võivad lokaliseeruda erinevates genoomi piirkondades, siis minisatelliidid on eelkõige detekteeritavad kromosoomide telomeersetes alades ja satelliit-DNA kromosoomi tsentromeersetes alades. Hajuskordusjärjestused on üle genoomi laiali paisatud, erinevast pikkusest ja esinemissagedusest tulenevalt klassifitseeritakse hajuskordusjärjestused mitmeks

alamklassiks – SINE'd (tuntuim esindaja Alu-kordusjärjestused), LINE'd (Kpn-perekond e. L1), LTR ja DNA transposoonid. Kui SINE'sid (*short interspersed nuclear elements*) iseloomustab kõrge GC protsent, siis LINE'sid (*long interspersed nuclear elements)* iseloomustab vastupidiselt kõrge AT sisaldus (Makalowski , 2001)..

Kuna inimese genoom sisaldab arvukalt kordusi ja enamasti korduvas DNA järjestuses geene ei leidu, siis on oluline need piirkonnad genoomis geenide ennustamisel või praimerite disainimisel maskeerida ehk välja selekteerida ebaspetsiifilisuse vältimiseks. Maskeerimiseks on kasutusel mitmeid erinevaid programme, mis maskeerivad erinevaid kordusi erinevates pikkustes. Neist tähtsamatest tuleb juttu allpool.

3.1 DUST

DUST on programm, mis maskeerib DNA järjestuses vähese keerukusega alad, st piirkonnad, milles kohalik nukleotiidide mitmekesisus on väike. Algoritm on vaikimisi kasutatav BLASTI otsingutes (ftp://ftp.ncbi.nih.gov/pub/tatusov/dust/version1/).

3.2 RepeatMasker

RepeatMasker maskeerib vaikimisi uuritavas järjestuses kõik inimese genoomis teadaolevad kordusjärjestuste perekondade esindajad, no lihtsad, tandeemsed ja keerulised kordusjärjestused (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker). Samuti on võimalik kasutajal ise ära määrata kordusjärjestuste hulk, mille vastu tahetakse kontrollimist sooritada. Programm võrdleb uuritavat järjestust korduvate elementide vastu ja maskeerib vastavalt sellele sisestatud järjestuse. Programm tagastab töö tulemusena maskeeritud järjestuse, on serveril baseeruv ja seega pole vajalik tülikas allalaadimine. Teoreetiliselt programmi tööaeg peaks kasvama lineaarselt järjestuse suurusega, kuid reaalselt on lühikeste järjestustega töötades programm märgatavalt kiirem kui pikkade järjestuste korral (http://repeatmasker.genome.washington.edu/RM/RepeatMasker).

3.3 GenomeMasker

GenomeMasker on programmipakett, mis sisaldab korduvate järjestuste maskeerimise ja leidmise programme GMASKER ja GTESTER (Reppo et al., unpublished.) GMASKER maskeerib üleesindatud kindla pikkusega sõnad FASTA formaadis järjestuse failis vältides sel viisil praimerite disainimist korduvjärjestusi sisaldavatesse aladesse. Maskeerimisprotsess on sarnane RepeatMaskeri tööpõhimõttega, kuid võib välja tuua mitmeid erinevusi kahe programmi vahel. Esiteks GMASKER ei maskeeri kogu üleesindatud sõna, vaid ainult sõna 3' otsas oleva viimase nukleotiidi asendades suure tähega märgitud nukleotiidi väikese tähega. Võimalik on ka teiste maskeerimissümbolite kasutamine (x, N). Teiseks erinevuseks kahe programmi vahel on asjaolu, et GMASKER-iga maskeerimine on asümmeetriline, sense/antisense ja forward/reverse ahelaid on võimalik maskeerida erinevalt muutes vaid käsurea parameetreid. Kuna GMASKER kasutab hash liste korduvatest 16 nukleotiidi pikkustest sõnadest on ta tunduvalt kiirem võrreldes teiste, järjestuse joondamise algoritme kasutavate programmidega. Programm kasutab üleesindatud sõnu sisaldavat nn musta nimekirja (*black list*), kuigi võimalik on kasutada sõnu pikkusvahemikus [8,16], soovitavad programmi autorid parima tulemuse saamiseks sõna pikkuseks 16 (http://www.biodata.ee/gm/README).

4. PCR primerite disain

Funktsionaalne genoomika püüab jätkuvalt leida kõigi genoomis leiduvate geenide ülesandeid. Geeni funktsionaalsuse iseloomustamise esimeseks sammuks on tema ekspressioonimustri identifitseerimine. Enamus lähenemisi, mis mõõdavad ekspressiooni taset, baseeruvad PCR produktidel ja seega vajatakse võimalikult spetsiifilisi PCR praimereid. PCR-i praimereid vajatakse ka genotüpiseerimisel, mutatsioonide uurimisel, cDNA ja genoomse DNA kloneerimisel jpt DNA järjestust uurivates protsessides.

Disainimise käigus leitakse kaks praimerit (praimerite paar), mis seostuvad järjestuses vajalikku kohta ennustatava tugevusega ja annavad loodetud unikaalse produkti. Praimeri seostumise tugevuse all mõistetakse tavaliselt füüsilist seondumistugevust.

4.1 Olulised PCR parameetrid

Praimerite disaini juures on oluline arvestada erinevate PCR-i parameetritega, olulisemad neist on sulamistemperatuur (Tm) ja produkti pikkus. Sulamistemperatuur on temperatuur, mille juures pool DNA ahelatest on üheahelalised ja pool kaheahelalised. Kõrgema GC nukleotiidide sisaldusega DNA järjestus omab kõrgemat sulamistemperatuuri tänu rohkematele vesiniksidemetele. Selleks, et saaks kasutada mõistlikke sulamistemperatuure tuleks praimerid teha pikkusega vahemikus [18,30], kusjuures ideaalseks pikkuseks loetakse vahemikku [20,25], ja sulamistemperatuuriga 50-70 °C. Parimate praimerite saamiseks tuleks paaris kasutatavad praimerid konstrueerida sarnase sulamistemperatuuriga. Sulamistemperatuuride arvutamiseks on välja töötatud erinevaid suurusi arvestavad valemid. Lihtsaimaks ja sellest tulenevalt ka suhteliselt ebatäpseks võimaluseks on praimeris nelja erineva nukleotiidi esinemise arvuga arvestamine, mille korral valem näeb välja järgmine (Kämpke et al., 2001):

$$Tm = 4\ °C * (G + C) + 2\ °C * (A+T)$$

Kõige täpsemat sulamistemperatuuri andvaks valemiks loetakse meetodit, mis kasutab *Nearest-Neighbor* algoritmi. n-aluselise pikkusega praimeri jaoks on n-1 nn nearest neighbor kombinatsiooni. Iga kombinatsiooni jaoks arvutatakse välja kaks termodünaamilist suurust – süsteemi korrapäratust kajastav entroopia ja protsessi soojussisaldust kirjeldav entalpia. Saadud suuruste abil arvutatakse välja vastav sulamistemperatuur allpool toodud valemiga:

$$Tm = \triangle H/(\triangle S + \triangle S + R * \ln(c/4)) - 273.15 + 16.6 * \log[Na],$$

kus H – entalpia, S - entroopia, R - universaalne gaasikonstant, R= 8.31J/mol, c – oligonukleotiidide totaalne molaarne konsentratsioon (Rychlik et al., 1990).

Kuigi viimastel aastatel on märkimisväärselt arenenud PCR-i meetodid pikkade DNA fragmentide amplifitseerimiseks, siis sellegipoolest PCR annab paremaid tulemusi lühemate fragmentide korral. Erinevate PCR-i meetoditega saab amplifitseerida varieeruva pikkusega DNA fragmente. Produkti pikkused jäävad tavaliselt aluspaaride

vahemikku [100, 3000], kuid uued tehnoloogiad võimaldavad saada ka kvaliteetseid produkte maksimaalse pikkusega 6 kb lähedale.

Samuti on tähtsad näitajad GC nukleotiidide sisalduse protsent praimeris, sekundaarstruktuuride tekke võimalus ja praimeri 3' otsa nukleotiidne koostis. Praimerite GC sisaldus peaks olema 50-55%, sest 50%-st palju väiksema GC-sisaldusega praimereid tuleks pikendada optimaalsest praimeri pikkusest pikemaks, et saavutada vajalik sulamistemperatuur üle 50 °C. Võimalike sekundaarsete hübridisatsioonide vältimiseks tuleks vältida üksikute (või kahe) nukleotiidide kordusi, seda eriti praimeri 3' otsas. Praimeri 3' otsas võiks viimase stabiliseerimiseks olla G või C nukleotiid, kuna praimeri spetsiifilise seostumise DNA järjestusega määrab tema 3' ots (Henegariu et al., 1997).

4.2 Praimerite seostumiskohtade kontroll

Kuna väga tähtis on seostumise spetsiifilisus ja ainult ühe unikaalse produkti tekkimine, siis on oluline teostada enne praimerite disaini antud järjestuse, millele praimerit disainitakse, kordusjärjestuste maskeerimine (vt. punkt 3) ja juba disainitud praimerite seostumise kontroll genoomi vastu ja ka kontrollida ega praimer ei seostu iseenda ega teise praimeri külge (Andreson et al., unpublished).

Selleks, et kontrollida praimerite erinevaid seondumiskohti genoomil, on välja töötatud mitmeid meetodeid. Tuntumad on BLAST (Altschul et al., 1997), MEGABLAST (Zhang et al., 2000), SSAHA (Ning et al., 2001). Kuigi erinevalt BLAST-ist ja MEGABLAST-ist on SSAHA märgatavalt kiirem, nõuab ta küllalt suurt mälumahtu (1GB). Programm GTESTER GenomeMasker paketist (Reppo et al., unpublished) on väga kiire nii suurte kui väikeste praimerite koguste korral. Viimane programm testib, et praimeritel poleks ülemääraseid seondumiskohti märklaudjärjestuses ja, kui palju ning mis asukohaga PCR produkte amplifitseeritakse antud matriitsjärjestuse korral. Rohkema kui ühe produkti amplifitseerimine võib anda "vale" positiivseid signaale genotüpiseerimisel. GTESTERi tööpõhimõte on sarnane SSAHA-le, kus eelnevalt genoomne DNA indekseeritakse ja kirjutatakse suurde *hash* tabelisse, kuid esimesena mainitud programm on spetsiifiliselt disainitud fikseeritud sõnapikkusele, mis on vaikimisi 16 ja on seetõttu suurusjärgu võrra kiirem. Samuti erinevalt SSAHA-st väikeste

andmemahtude korral pole vaja indekseid lugeda füüsilisse mällu, mis teeb otsingu aeglasemaks.

4.3 Praimerite disaini programmid

PCR praimerite disainimiseks on olemas erinevaid nii kommertsiaalseid kui ka tasuta kättesaadavaid programme. Praimerite disainimisprogrammid PRIMER3 (Rozen et al., 1998), PRIDE (Haas et al., 1998), GST-PRIME (Varotto et al., 2001) on mõeldud suuremahuliste (kogu genoomi) projektide jaoks, tuntumad programmid väiksemate praimerikoguste modelleerimiseks on OLIGO (Rychlik et al., 1989), PRIME, PRIMERSELECT.

PRIMER3 disainib praimerid PCR-i reaktsiooni jaoks vastavalt kasutaja poolt valitud parameetritele. Programm PRIDE eelisteks on suur kasutajasõbralikkus, st nõuab kasutaja minimaalset vahelesegamist, ja maskeerimise vältimine, mistõttu otsitakse praimereid ka kriitilistest järjestuste piirkondadest, kus võib sisalduda hea kvaliteediga seondumiskohti. GST-PRIME erilisuseks võib lugeda seda, et ta detekteerib ja seejärel väldib introneid ja splaissaite praimerite disainimisel, seetõttu võimaldab programm efektiivset ja automaatset praimerite modelleerimist ka keerulistele genoomidele, mis sisaldavad palju introneid ja eksoneid. Pakett PRIMERSELECT võimaldab kasutajal endal sisestada oma praimerite järjestused või võib lasta eelpool nimetatud programmil disainida vastava järjestuse jaoks võimalikke praimereid ja seejärel analüüsida nende headust ja vastavaid termodünaamika parameetreid (https://www.dnastar.com/cgi-bin/php.cgi?r11.php).

5. Ülesande püstitus

Arvestades genoomi info suurt mahtu on vaja automaatseid meetodeid praimeri disainiks. Üks olulistest probleemidest on kodeerivate alade eksonite amplifitseerimine (näiteks genotüpiseerimiseks, geneetiliste haiguste uurimiseks vms). Nende ülesannete lihtsustamiseks lõime automaatse meetodi vajalike praimerite disainiks koos vastavate programmide ja andmebaasiga. Viimastest antakse täpsem ülevaade allpool.

II Praktiline osa

Töö eesmärgid

Käesoleva töö eesmärgiks on inimese genoomi kõikide geenide ja ühtlasi ka kõikide eksonite amplifitseerimiseks vajalike praimerite automaatne disainimine ning vastavate PCR-i produktide produtseerimine. Eesmärk hõlmab ka saadud resultaadi, so kõigi disainitud praimerite kasutajale kättesaadavaks tegemist; viimane kujutab endast töö tulemuste viimist andmebaasi ja graafilise kasutajaliidese konstrueerimist, mis sisaldab endas vastavat otsingumootorit vajalike praimerite leidmiseks.

Meetodid

1. Andmete päritolu ja struktuur

Vajalikud algandmed praimerite disainimiseks saime ENSEMBL-i andmebaasist, kus käesoleval ajal on saadaval inimese genoomi andmebaasi versioon 13.31.1. Praeguse seisuga (antud töö kirjutamise ajal on andmebaasi viimati muudetud 31-03-2003) sisaldab andmebaas 24,847 ENSEMBL-i geeniennustuse geeni, 58,770 nn genscan geeni, 248,449 ENSEMBL-i eksonit. Andmete kätte saamiseks kasutasime ENSEMBL-i kodulehel pakutavat andmekaevandust võimaldava töövahendi EnsemblMart-i (lühemalt EnsMart) web-i kasutajaliidest *MartView*, mis võimaldab lihtsalt ja kiirelt endale sobival kujul vajalike andmete allalaadimist. Meie valisime inimese genoomist nii teadaolevad (known genes) kui ka uudsed (novel genes) Ensembl geenid ja Sanger geenid (praegused Vega geenid, kus on veel lisaks 14. kromosoom). Resultaatfailis soovisime näha kromosoomi nime (*Chromosome Name*), Ensembli ja Sangeri geenide ning eksonite spetsiifilisi identifikaatoreid (*Ensembl Gene ID, Sanger Gene ID, Ensembl Exon ID, Sanger Exon ID*), peale selle olid veel vajalikud eksoni algus- ja lõppkoordinaat (*Exon Start, Exon End*).

2. Kasutatud programmide kirjeldus

Põhimõtteliselt võib genereeritud praimerite disainimist teostavad programmid jagada kolmeks. Esimene programm (blokid.pl/blokid_gene.pl, vt edasi) tekitab

eksonite liitmise teel amplikonid maksimaalse pikkusega kuni 3000 aluspaari. Seejärel igale disainitud amplikonile konstrueerib ühe praimeripaari teine programm (exon_primers_31.pl/gene_primers.pl, vt edasi), mis eelnevalt maskeerib kogu genoomi ohtlikke kordusjärjestusi sisaldavatest regioonidest. Kolmas programm (gtest_31.pl/gtest_g.pl, vt edasi) testib eelmise programmi töö tulemusena saadud praimerite seondumiskohti üle genoomi ja selekteerib rohkemat kui ühte seondumiskohta omavad praimerid välja. Amplikonidele, mille praimerid eelpool kirjeldatud programmi töö tulemusena välja praagiti, disainitakse asemele 100 uut praimerit teise programmiga analoogse pragrammi abil (exon_primers100_31.pl/gene_primers100.pl, vt edasi), misjärel kontrollitakse kõigi 100 praimeripaari seostumisi üle genoomi kolmanda programmiga analoogse programmi abil (gtest100_31.pl/gtest100_g.pl, vt edasi). Sellise tööpõhimõtte järgi leitakse praktiliselt kõigile amplikonidele optimaalsed praimerid.

Lisaks sellele võib programmid jagada veel kaheks vastavalt erinevale resultaadile. Ühed programmid genereerivad geenispetsiifilisi praimereid, see tähendab üks amplikon sisaldab vaid ühest konkreetsest geenist pärinevaid eksoneid. Ja teised programmid teostavad kogu genoomi eksonite jaoks sobivate praimerite leimise, sel juhul võib üks amplikon sisaldada eksoneid ka mitmest erinevast geenist.

3. Vajalik mälumaht ja tööaeg

Programmide, mis konstrueerivad praimereid amplikonidele, vajalik mälumaht ulatub 80-100 Mb-ni, seda põhiliselt tänu antud programmides kasutatavale GMASKERile. Programmid, mis kontrollivad konstrueeritud praimerite seondumiskohtade unikaalsust genoomis vajavad tunduvalt rohkem mälumahtu, vajalik mälumaht küündib 350–450 Mb-ni ja siinkohal põhiliselt tänu kasutatavale programmile GTESTER.

Programmide tööajad on toodud Tabelis 1, kus on erinevate programmide lõikes näidatud jooksutamiseks kuluvad ajad tundides. Aega arvutati programmi käivitamise hetkest kuni lõpptulemuse kättesaamiseni. Kõige aeganõudvamad on programmid, mis kasutavad GTESTER-it, need on gtest_g.pl, gtest_31.pl ja gtest100_g.pl, gtest100_31.pl

Tabel 1. Konstrueeritud programmide tööajad

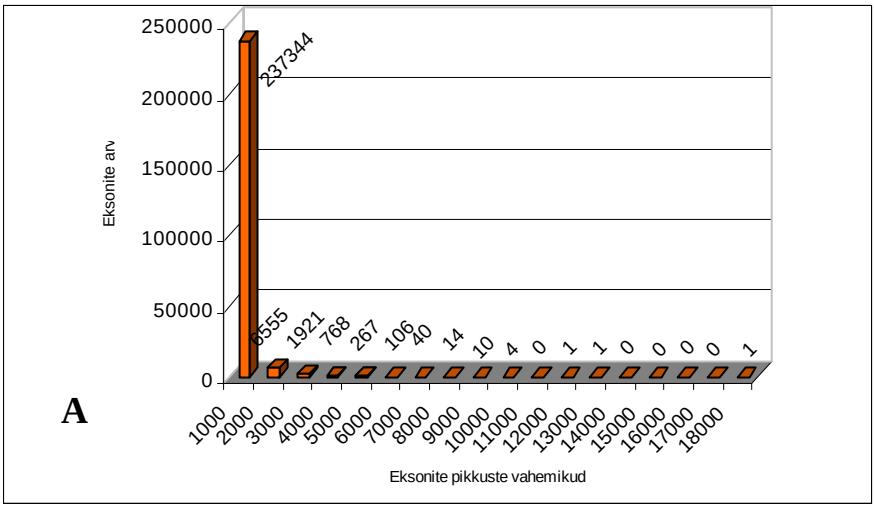
Geenile praimerite disainil osalevate programmide nimed	tööaeg, h	Üle genoomi praimerite disainil osalevate programmide nimed	tööaeg, h
blokid_gene.pl	<0,09	blokid.pl	<0,09
gene_primers.pl	2,10	exon_primers_31.pl	2,20
gtest_g.pl	4,50	gtest_31.pl	4,50
gene_primers100.pl	5,00	exon_primers100_31.pl	4,30
gtest100_g.pl	5,20	gtest100_31.pl	6,00
TOTAL	16,89	TOTAL	17,09

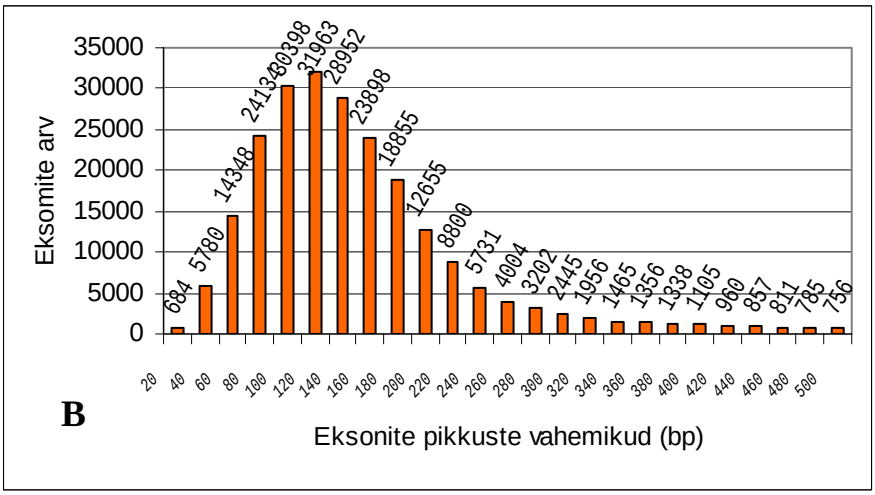
Tulemused

1. Eksonite kirjeldus

Enamustes inimese geenides leidub rohkem kui üks ekson. Valdavalt on eksonid geenis oluliselt väiksemad kui intronid; enamus eksoni pikkuseid jääb alla 1,000 aluspaari (Joonis 1, A) ja keskmine eksonite suurus jääb vahemikku [100,200] aluspaari (Joonis 1, B), samal ajal kui keskmine introni pikkus on 1,000 aluspaari ringis. Leitud on ka 100,000 aluse pikkuseid introneid. Pikim teadaolev ekson on ca 17,000 aluspaari pikk. Kuigi eksonite arv geenis on väga varieeruv suurus, võib tinglikult keskmiseks eksonite arvuks ühes geenis lugeda 10.

Käesolevas töös kasutasime 248,449 ENSEMBL-i eksonit, mis igaüks omas unikaalset identifikaatorit (*Exon ID*) ühise mustriga ENSE0000xxxxxxx ja lisaks veel Sangeri Keskuse poolt annoteeritud 20. ja 22. kromosoomi eksoneid ID-ga OTTHUME0002xxxxxxx. Praimerite disainil ei eristatud praimerite 3' ja 5' otsi.





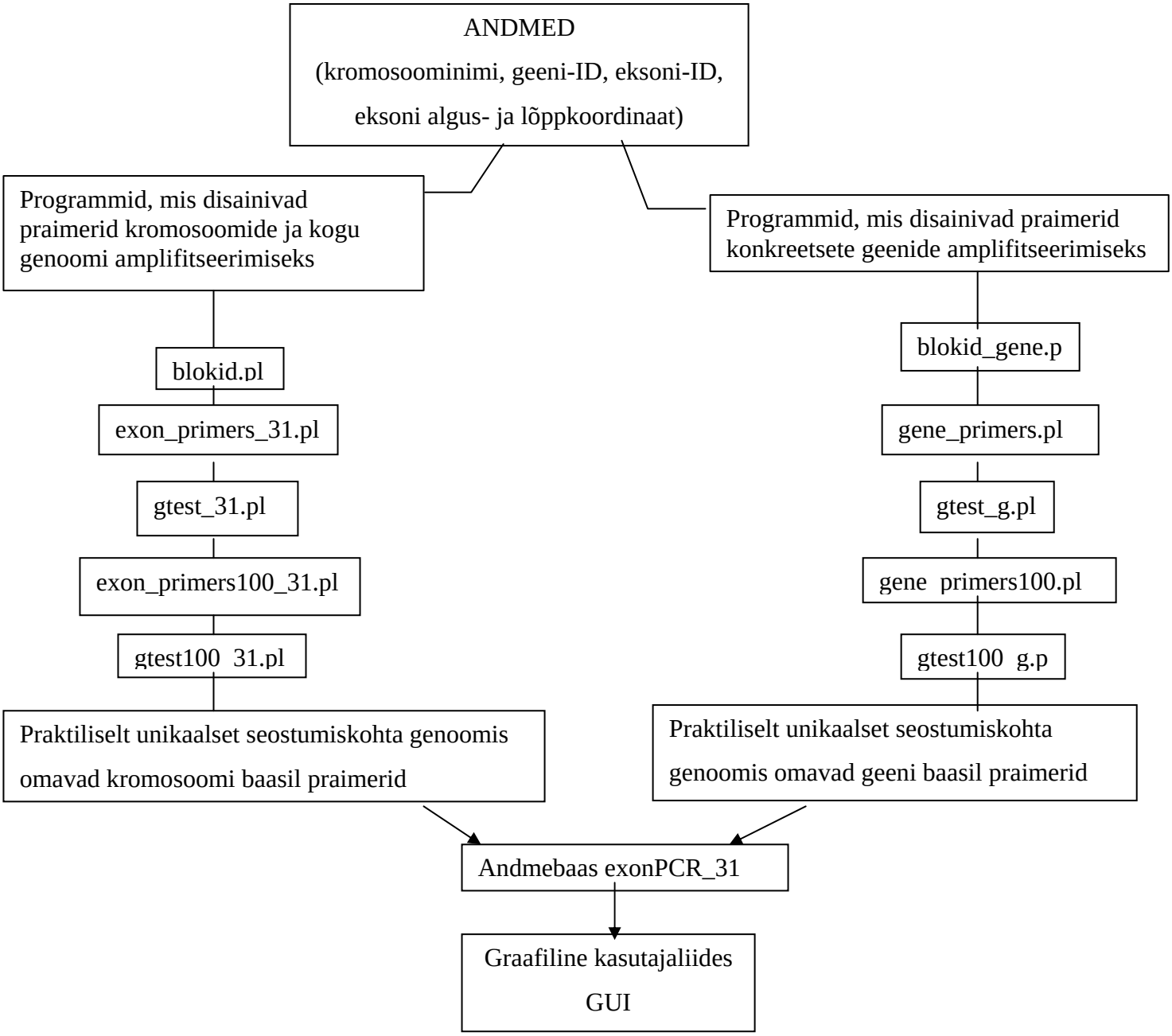
Joonis1. Eksonite pikkuste jaotus

A - Eksonite pikkuste jaotus üle genoomi. Enamus eksoneid on alla 1000 aluspaari pikad.

B - Alla 500 aluspaariste inimese eksonite pikkuste jaotus. Keskmine eksonite pikkus inimese genoomis on 200 aluspaari.

2. Programmide tööpõhimõte

Üldjoontes võib kirjutatud programmid jagada kaheks – ühed, mis disainivad praimerid spetsiifiliselt geenidele ja teised, mis teevad praimerid kogu genoomile ega arvesta konkreetsete geenide piire. Üldine programmide tööpõhimõte on toodud Joonisel 2. Edasi vaatleme tähtsamaid momente, mida kõne all oleva töö praktiline osa kätkeb.



Joonis 2. Olulised momendid praimerite disainimisel.

blokid.pl ja blokid_gene.pl genereerivad amplikonid, millele praimereid disainitakse; exon_primers_31.pl ja gene_primers.pl tõmbavad vastavast andmebaasist *flanking* piirkondadega amplikonide koordinaatide järgi järjestused ja disainivad igale amplikonile ühe praimeripaari; gtest_31.pl ja gtest_g.pl testivad saadud praimerite seostumiskohti üle genoomi; exon_primers100_31.pl ja gene_primers100.pl disainivad uued praimerid (100 tk) neile amplikonidele, millele eelnevalt disainitud praimerid genoomitestil kaduma läksid; gtest100_31.pl ja gtest100_g.pl testivad eelmise programmi poolt disainitud primerite seostumiskohti üle genoomi; saadud praimerid, mis on unikaalse seostumiskohaga genoomis, viiakse andmebaasi ja sealt vastavate päringute abil graafilise kasutajaliidese vahendusel saab kasutaja endale

2.1. Amplikonide konstrueerimine

Kuna eksonite keskmine pikkus genoomis on vahemikus [150,200] ja traditsiooniline PCR tehnoloogia võimaldab amplifitseerida kvaliteetseid DNA fragmente pikkusega ca 3,000 nukleotiidi, siis käesolevas töös esimeseks etapiks priamerite disainimisel on lühikeste eksonite liitmine pikemaks DNA järjestuseks maksimaalse pikkusega kuni 3,000 nukleotiidi. Neid ca 3,000 nukleotiidilisi DNA fragmente e. amplikone konstrueerivad eksonite algus- ja lõppkoordinaatide järgi programmid nimega blokid.pl ja blokid_gene.pl, mis on vastavalt kas genoomi- või geenipõhiste praimerite disainimiseks. Eelpool nimetatud programmid erinevad vaid selle poolest, et esimene konstrueerib sellised amplikonid, kuhu võivad kuuluda eksonid ka erinevatest geenidest, samal ajal kui teine programm ei luba ühte amplikoni rohkema kui ühe geeni eksonite kuulumist. Samas annavad amplikone genereerivad programmid ka igale amplikonile ühe iga kromosoomi kohta unikaalse identifikaatori ehk aID.

2.2. Praimerite konstrueerimine amplikonidele

Järgmine programm saab sisendiks amplikonide ID-d, amplikoni algus- ja lõppkoordinaadi ja vastavasse amplikoni kuuluvate eksonite ID-d ja spetsiifiliselt geenidele praimereid disainivad programmid saavad täiendavalt ka vastava geeni ID. Edasi pole kahe erinevat tulemust (kogu genoomi ja spetsiifiliselt geeni baasil praimerite disainimine) saavutava lähenemise vahel põhimõttelisi erinevusi, seetõttu allpool ei vaadelda kahte programmi eraldi. Sisendiks saadud amplikonide koordinaatidele lisatakse 100 nukleotiidi praimerite seondumiskohtadeks, neid amplikonide otstesse lisatud lühikesi praimerite seostumiseks mõeldud järjestusi nimetatakse nn *flanking* piirkondadeks. Saadud koordinaatide järgi tõmbavad programmid exon_primers_31.pl ja gene_primers.pl andmebaasist core_31 vastava DNA järjestuse. Seejärel teostatakse 15 nukleotiidi pikkuste lihtsate kordusjärjestuste maskeerimine kogu genoomis maskeerimisprogrammiga DUST, mis asendab järjestuses suurte tähtedena esinevad nukleotiidid (A,T,G,C) väikeste tähtedega (a, t, g, c). Keerulisemate kordusjärjestuste maskeerimist teostatakse programmiga GMASKER (kirjeldatud ülevalpool), mis kasutades inimese genoomis asuvaid erinevaid

kordusjärjestusi sisaldavat musta nimekirja (*black list*) kontrollib viimases sisalduvate kordusjärjestuste esinemist dustiga maskeeritud järjestuses. Saadud maskeeritud järjestus konverteeritakse praimereid disainiva programmi GM_PRIMER3 (kohandatud PRIMER3) sisendformaati, misjärel sooritatakse praimerite genereerimine programmiga GM_PRIMER3. Viimase sisendfail sisaldab PCR-i parameetrite lubatud varieerumiste vahemikke ja ka kõige optimaalsemaid PCR-i parameetrite väärtusi. Meie kasutasime järgnevaid PCR-i parameetrite väärtusi – produkti suurus võib olla vahemikus [100, 20000], kuid optimaalseks väärtuseks on 600 nukleotiidi; praimeri pikkused võivad kuuluda vahemikku [16,30], kusjuures optimaalseks pikkuseks arvestatakse 21 nukleotiidi; sulamistemperatuur võib kõikuda vahemikus [59,65] °C, kuid optimaalne väärtus on 62 °C ja sulamistemperatuuri erinevus kahe ühte praimeripaari kuuluva praimeri vahel on 4°C; GC nukleotiidide sisalduse protsendil lubatakse varieeruda piirides [20,80] ja optimaalne väärtus on 35%. Lisaks sellele saab määrata ka mitu PCR praimeripaari ühele amplikonile soovitakse vastava programmi töö tulemusena saada. Meie tahtsime saada ühte produkti, peale selle soovisime saada ka vastavat PCR produkti. Kuna GM_PRIMER3 ei suuda ühe sellise tsükli tulemusena kõigile amplikonidele häid praimereid leida, siis oli vajalik sellesse programmi tsükli lisamine. Kõige optimaalsemaks tsüklite arvuks kujunes 12, igas tsüklis suurendatakse amplikonidele liidetavate nn *flanking* piirkondade pikkust 100 nukleotiidi võrra ja lisaks sellele kasutab programm GMASKER alates 5ndast tsüklist väiksemat kordusjärjestuste musta nimekirja amplikonide maskeerimiseks. Sellise meetodiga leitakse praktiliselt kõigile amplikonidele praimeripaar.

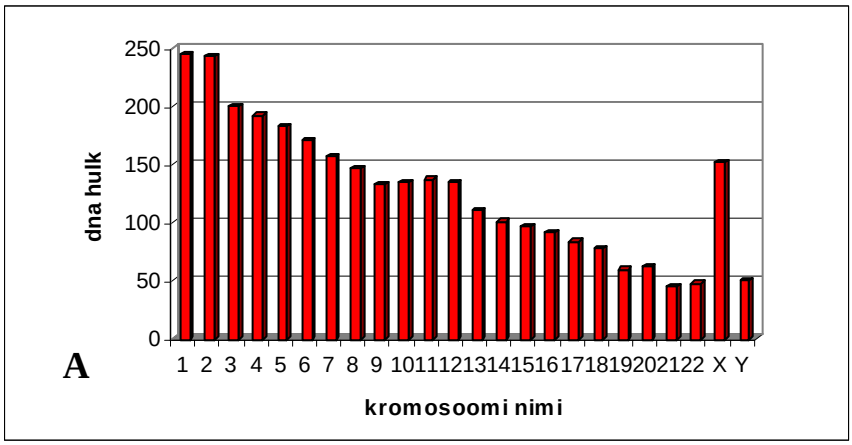
2.3. Praimerite seostumiskohtade unikaalsuse kontroll genoomis

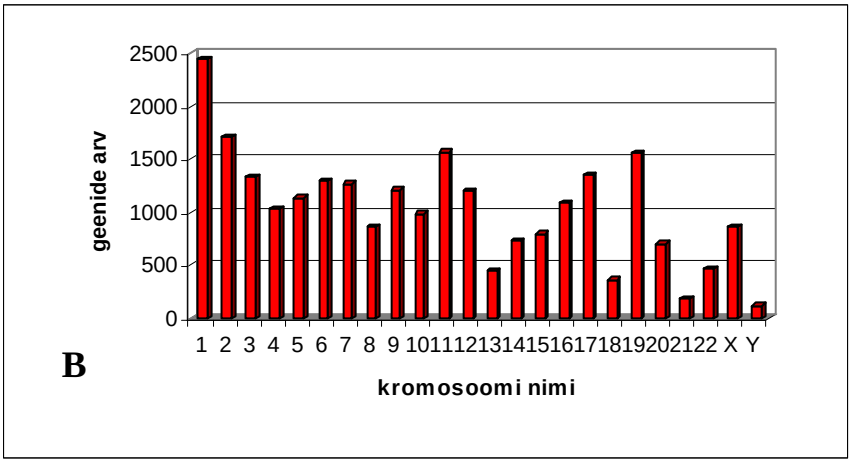
Programmid gtest_31.pl ja gtest_g.pl testivad eelnevalt disainitud praimerite seostumist genoomile kasutades selleks programmi GTESTER, mida kirjeldati käesoleva töö teoreetilises osas. Kui praimeril on seostumiskohti genoomis rohkem kui üks, siis tehakse vastavatele amplikonidele uued praimerid programmidega exon_primers100_31.pl ja gene_primers100.pl, mis on vastavalt genoomi (kombinatsioon "_31" programminimes) ja spetsiifiliselt geeni baasil (sõna "gene" programminimes) konstrueeritud amplikonidele vastavad programmid. Arv 100 tähendab, et nüüd konstrueeritakse ühe praimeripaari asemel ühele amplikonile 100

erinevat praimeripaari, et tagada kõigile amplikonidele parimate praimeripaaride konstrueerimine, põhimõtteliselt teistes kohtades ei erine nimetatud programmid eelpool kirjeldatud 12 tsüklit sisaldavatest programmidest. Saadud saja praimeripaari võimalikke seostumiskohti testitakse jällegi eelpool kirjeldatud programmidega analoogseid seondumiskohtade testimist läbi viivaid programme gtest100_31.pl ja gtest100_g.pl. Selliste järjestikuste protsessidega tagatakse praktiliselt kõikidele konstrueeritud amplikonidele genoomis unikaalset seostumiskohta omavate praimeripaaride disainimine.

3. Tulemuste kirjeldus

Inimese tuuma DNA hulk on määratud kromosoomide arvu ja suuruse poolt. Inimese genoomi suurus on 3000 Mbp. DNA hulk kromosoomide lõikes on toodud Joonisel 3A. Nagu on näha Joonisel 3B pole kromosoomis sisalduva DNA hulga ja geenide arvu vahel üks-ühest seost, nt kui 18. ja 19. kromosoom on nendes sisalduva DNA hulga suhtes võrdväärse suurusega, siis võrreldes nendes sisalduvat geenide arvu võib näha teravaid erinevusi.

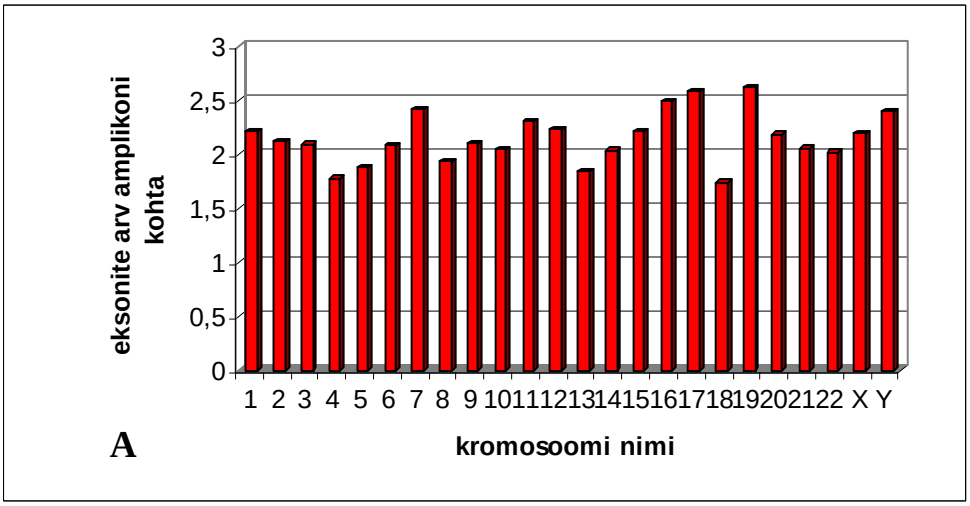


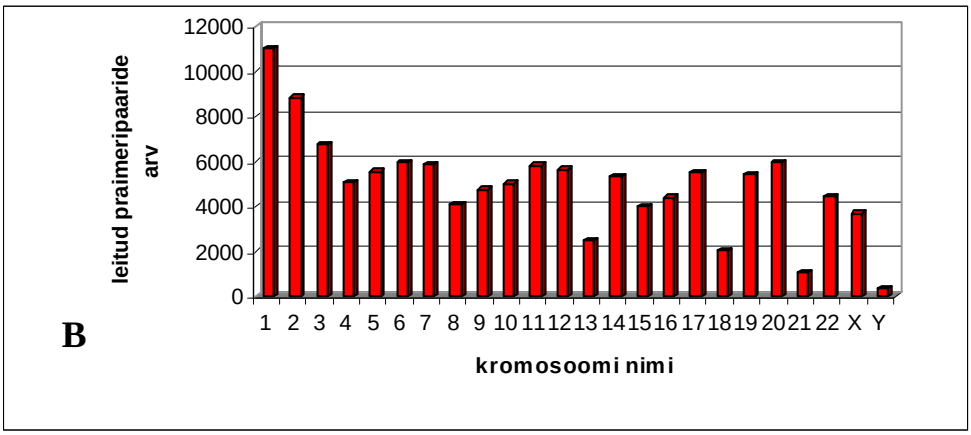


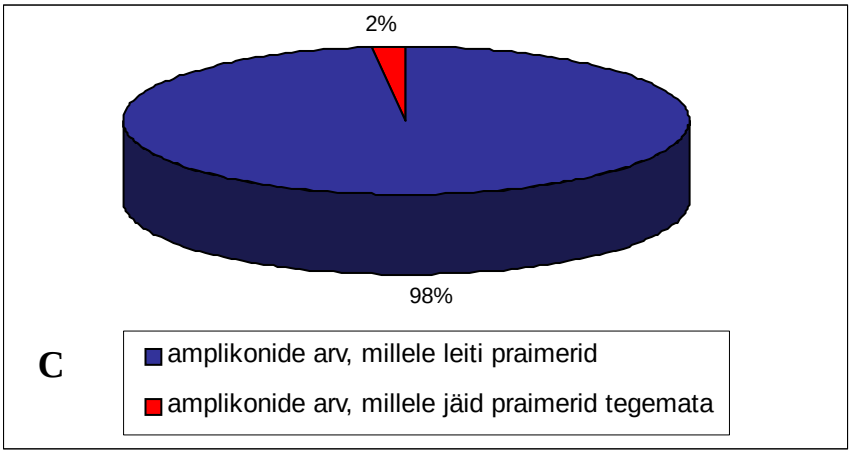
Joonis 3.
A - kromosoomide suurus (Mb)
B – geenide arv kromosoomi

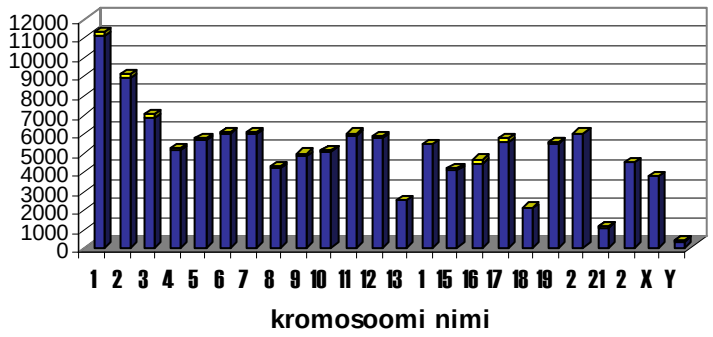
Kuigi geenide ja seega ka eksonite arv erineb kromosoomiti, siis keskmised eksonite ja intronite pikkused on eri kromosoomides ligikaudu samasugused. Seda näitab ka konstrueeritud amplikonidesse kuuluvate eksonite keskmine arv amplikoni kohta (Joonis 4A), mis on umbkaudu 2 igas kromosoomis. Joonisel 4B võib näha amplikonide (ühtlasi ka praimeripaaride) arvu kromosoomi lõikes, millele suudeti konstrueerida praimerid. Amplikonide, millele leiti unikaalne praimeripaar, protsentuaalset suhet kõikidesse konstrueeritud amplikonidesse kirjeldab Joonisel 4C olev diagramm. 2-le protsendile kõigist konstrueeritud amplikonidest ei suudetud unikaalset praimeripaari disainida. Kõikidest inimese geenidest moodustavad 8% sellised geenid, millel on puudu vähemalt ühe eksoni amplifitseerimiseks vajalikud praimereid, st vastaval geenil leidub ekson, mis kuulub sellisesse amplikoni, millele praimereid ei suudetud disainida (Joonis 4F). Totaalne konstrueeritud amplikonide arv on kasutatud andmete (eelpool kirjeldatud) korral 121,376, millest 118,994 leiti kirjeldatud kriteeriumitele vastavad praimerid, seega praimerid jäid tegemata 2,382 amplikonile, mis keskmiselt sisaldasid igaüks 2 eksonit. Kasutatud 248,500 eksonist 4,800 eksonit ei sisaldu amplikonides, millele leiti praimerid, so ligikaudu 2 protsenti totaalsest eksonite arvust. Disainitud praimeripaaridega amplikonide arvu suhet amplikonide arvu, millele ei leitud praimereid, illustreerib Joonis 4D. Suhteliselt väike oli üle 3,000 nukleotiidi pikkuste eksonite osakaal (Joonis 4E). Mäletatavasti tuli

efektiivsemaks amplifitsserimiseks eksonid pikkusega üle 3,000 nukleotiidi osadeks teha, nii et üks ekson kuulus korraga rohkemasse kui ühte amplikoni.



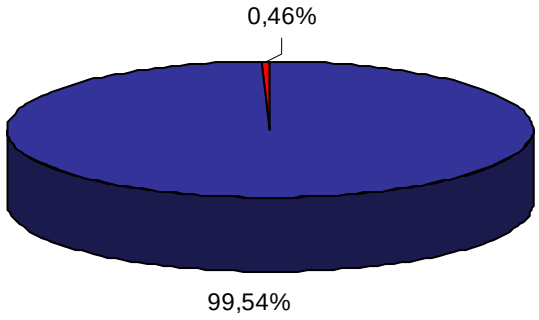






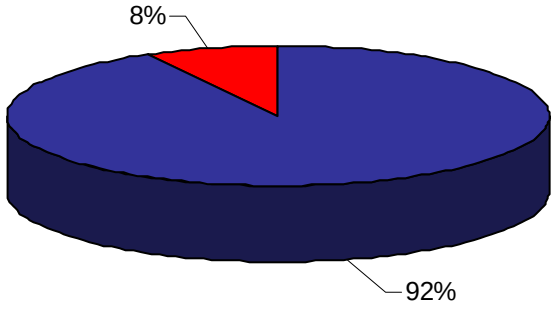
D

amplikonid, millele ei leitud praimereid
amplikonid, millele leiti praimerid



E

eksonite arv, mis on lühemad kui 3000 nukleotiidi
eksonite arv, mis on pikemad kui 3000 nukleotiidi



F

Geenid, mille kõigi eksonite amplifitseerimiseks vajalikud praimerid suudeti disainida

Geenid, millel leidub vähemalt üks ekson, mis kuulub sellisesse amplikoni, millele praimereid ei suudetud disainida

Joonis 4.

A – keskmine eksonite arv amplikoni kohta kromosoomide lõikes

B – amplikonide arv, millele leiti praimerid kromosoomi kohta

C - amplikonidele leitud ja leidmata praimerite arvu suhe totaalsesse konstrueeritud amplikonide arvu

D – amplikonide arv, millele leiti praimerid vs amplikonide arv, millele jäid praimerid tegemata kromosoomi lõikes

E – vähemalt kahte amplikoni kuuluvate eksonite arvu ja rangelt ühte amplikoni kuuluvate eksonite osakaal totaalse eksonite arvu suhtes

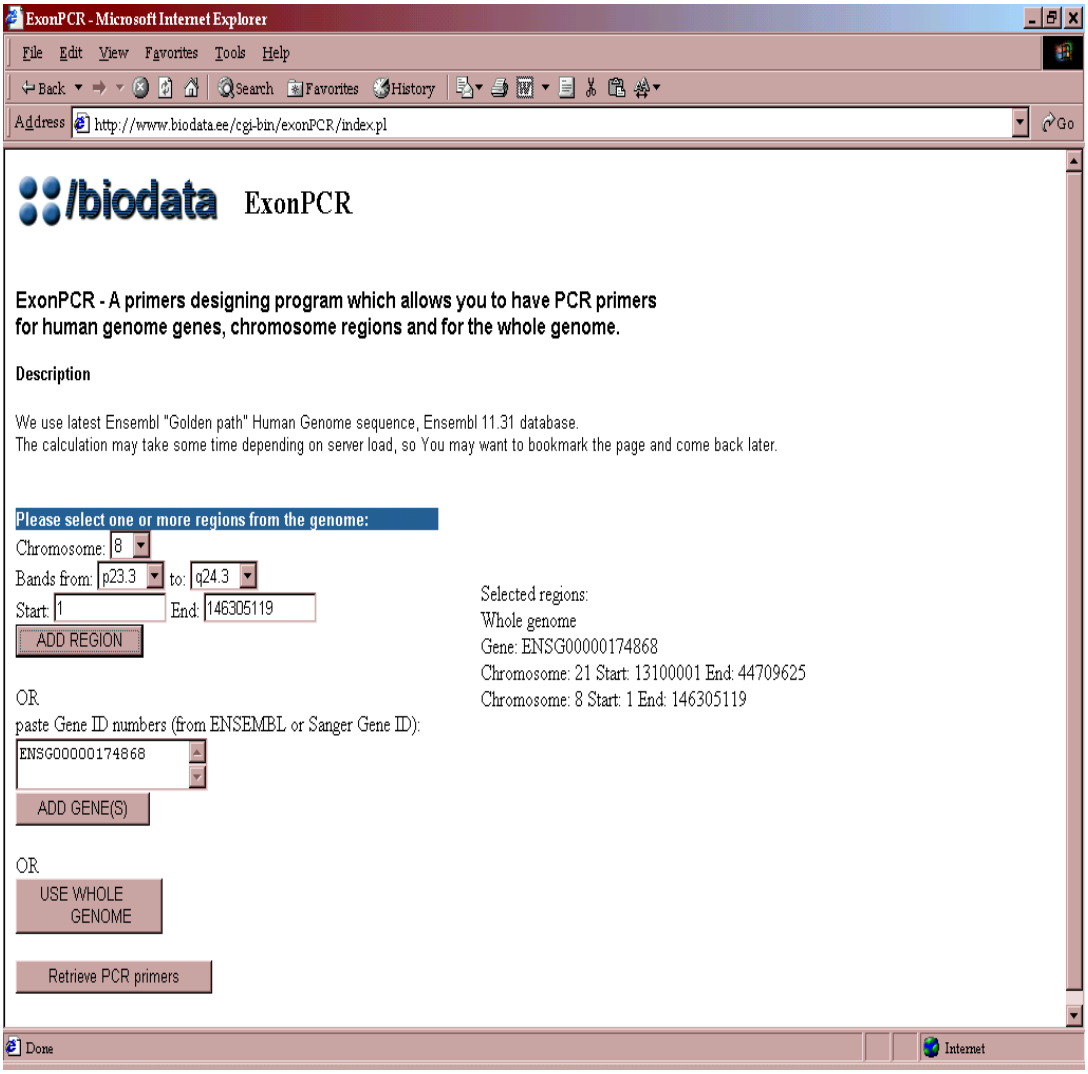
F – geenide arv, mille kõik eksonid on võimalik disainitud praimerite abil amplifitseerida versus geenide arv, millel on puudu vähemalt ühe eksoni amplifitseerimiseks vajaliku amplikoni praimeripaar

4. Andmebaasi loomine

Kuna iga bioinformaatilise analüüsi ja töö kasulikkus sõltub sellest kui hea on ligipääsetavus töötulemustele, siis kasutades MySQL-i võimalusi lõime konstrueeritud praimerite ja vastavate produktide jaoks andmebaasi, kust graafilise kasutajaliidese abil on võimalik vajalikele (genoomi piirkonna) geenidele vastavad praimeripaarid leida. Andmebaas nimega exonPCR_31 sisaldab 6 tabelit, kolm tabelit on mõeldud maksimaalselt ühe geeni eksoneid sisaldavate amplikonide praimerite jaoks ja kolm tabelit selliste amplikonide praimerite jaoks, mille korral amplikonide konstrueerimisel polnud sellist piirangut, et amplikoni võivad kuuluda ainult ühest geenist pärinevad eksonid. Tabelite struktuur on kahel juhul enam-vähem sama. Esimesed konstrueeritud tabelid on exon_g ja exon_chr, mis sisaldavad iga eksoni ID jaoks amplikoni ID-d ja kromosoomi nime, kuhu antud ekson kuulub, kusjuures üks ekson ei kuulu rohkemasse kui ühte antud kromosoomi konkreetsesse amplikoni, st et eksonite ID veerus ei leidu kahte sellist rida, mis oleksid teineteisega võrdsed. Teiseks moodustasime tabelid amplicons_g ja amplicons_chr, mis sisaldavad nagu eelmised tabelidki kromosoomi nime ja amplikoni ID-d, kuid lisaks sellele kuulub kõne all olevate tabelite struktuuri amplikoni algus- ning lõppkoordinaat ja amplicons_g koosseisu veel geeni ID. Kolmandaks lõime tabelid primers_g ja primers_chr, mis on struktuurilt identsed ja sisaldavad kromosoomi nime, amplikoni ID-d, praimereid ja vastavat PCR-i produkti.

5. Kasutajaliidese loomine

Kasutaja saab vajalikud praimerid kätte interaktiivse webilehekülje kaudu (Joonis 5, http://www.biodata.ee/exonPCR), kus päringute sooritamine on kiire ja mugav. Lehekülje tegemisel on kasutatud ühte paljudest võimalikest dünaamilise lehe loomise vahenditest, nimelt CGI-d. Lehekülg sisaldab (töö kirjutamise hetkel) kolme erinevat otsingumootorit. Esimese abil saab kasutaja valida kromosoomi nime ja vastava regiooni kromosoomil, kas kromosoomi bändide kaudu või ise kirjutada ennast huvitav regioon vastavatesse lahtritesse. Teise otsingumootori abil on võimalik sooritada päring ENSEMBLI või Sangeri geeni ID järgi ja kolmandaks saab endale ka kogu inimese genoomi geenide amplifitseerimiseks vajalikud praimerid tõmmata. Tulemused kuvatakse järgmise lehe kaudu, kust kasutaja saab endale päringu vastusena saabunud praimerifaili oma arvutisse kopeerida.



Joonis 5. Graafiline kasutajaliides. Võimaldab vastavalt kasutaja vajadusele sisestada ennast huvitava kromosoomi nime ja valida/kirjutada kromosoomi regiooni, geeni ID järgi leida vastavale geenile praimerid, ka on võiamlus kogu genoomi geenidele praimerid tõmmata.

Arutelu

Analüüsides töö tulemusi võib jõuda järeldusele, et väljatöötatud lahendus on töökindel ja seda saab kasutada automaatseks praimerite disainiks. Siiski jääb praeguse metoodikaga amplifitseerimata 2% amplikonidest. Puuduvaid eksoneid jääb ca 8% geenidest.

Seega - konstrueeritud programme tuleks maksimaalse tulemuse, so kõikidele amplikonidele praimerite disainimine, saavutamiseks edasi arendada ja täiustada. Viimast on võimalik teostada erinevatel viisidel. Üheks võimaluseks oleks PCR-i parameetrite ulatuslikum varieerimine ja rohkemate erinevate parameetrite kombinatsioonide läbi proovimine praimerite disainil. Kõige töömahukam lähenemine kõigile amplikonidele praimerite saamisel oleks olnud käsitsi analüüsida amplikonide *flanking* piirkondi, kuhu praimereid ei õnnestunud teha ja seejärel sobiv lahendus leida vastavalt sellele, milline oli *flanking* piirkonna nukleotiidne järjestus, näiteks võib lahenduseks olla maskeerimistingimuste muutmine vms. Samuti oleks võinud konstrueerida teistsuguste kriteeriumite alusel uued amplikonid nende asemele, millele ei õnnestunud praimereid disainida, nii et amplikone ümbritsev *flanking* piirkonna järjestus oleks olnud erinev. Kindlasti on võimalusi praimeridisaini programmide efektiivsemaks muutmiseks veelgi. Oluline on veel märkida, et erinevate praimeridisainis osalevate programmide erinevate parameetrite liigne varieerimine tagaks vaid näilise unikaalset seostumiskohta omavate praimerite disaini. Tõenäosus, et sellised praimerid ei oma rohkemat kui ühte seostumiskohta genoomis, on väike.

Kahtlemata on ka tähtis graafilise kasutajaliidese ja ühtlasi ka andmebaasi täiustamine. Kasutajal peaks olema võimalus saada rohkem infot järjestuste kohta, millele ta soovis praimereid. Näiteks, kui kasutaja sisestas kromosoomi regiooni bändide abil, siis võiks päring vastusena tuua ka, mis geenid selles regioonis sisalduvad, mida vastavad geenid kodeerivad, homoloogsed geenid teistest organismidest jne.

Oluline oleks ka eksperimentaalsete katsete teostamine laborites programmide töö efektiivsuse ja õigsuse hindamiseks. Tuleks otsida võimalikult mitmekesine disainitud praimerite komplekt, mille funktsioneerimise täpsust laboris kontrollitakse PCR-i protsessi abil. Komplekt peaks sisaldama selliseid praimereid, mis konstrueeriti

neile amplikonidele, mis sisaldasid ühte (pikka) eksonit või, mis konstrueeriti neile amplikonidele, milles oli rohkesti erinevaid eksoneid.

Muidugi tuleb arvestada algandmete võimaliku ebaõigsusega. Suhteliselt kvaliteetseid annoteerimistulemusi pakuvad nii ENSEMBL-i kui ka NCBI. Seda, kumb neist täpsemaid andmeid pakub, on praegu veel raske hinnata. Tõenäoliselt saab täpsema hinnangu selle kohta anda siis, kui vastavaid andmeid on laialdasemalt kasutatud.

Kokkuvõte

Käesolevas töös disainiti praktiliselt kõik inimese geenide amplifitseerimiseks vajalikud PCR praimerid. Kogu inimese genoomi geenide amplifitseerimiseks vajalikud praimerid modelleeriti ühe programmide variandi poolt ja teine programmide pakett konstrueeris kasutaja poolt määratud geenide amplifitseerimiseks vajalikud praimerid. Programmid on analoogsed, erinavad vaid amplikonide konstrueerimise erineva põhimõtte poolest.

Tööd alustati ENSEMBL-i andmebaasist vajalike järjestusandmete tõmbamisega. Seejärel kirjutati programmid, mis tegid ühe PCR produktina amplifitseeritavad amplikonid, kuhu ühe variandi järgi kuulusid vaid ühes geenis sisalduvad eksonid, teisel juhul sellist kitsendust polnud. Edasi kirjutati praimereid konstrueerivad programmid. Saadud praimeritele teostati seostumise unikaalsuse kontroll üle genoomi. Kui leidus praimereid, mis sisaldasid genoomis rohkemat kui ühte kinnituskohta, siis disainiti vastavale amplikonile uued praimerid samal põhimõttel kui enne, kuid ühe praimeripaari asemel saadi programmi töötulemusena sel korral 100 praimeripaari, et garanteerida unikaalse praimeripaari leidmine. Ka viimasena nimetatud praimeritele teostati unikaalsuse kontroll üle genoomi. Sellise meetodiga saadi 98%-le konstrueeritud amplikonidest unikaalsed PCR praimerid.

Eelviimase etapina loodi disainitud praimerite ja nende produktide jaoks andmebaas ja lõpuks konstrueeriti ka efektiivseks töötulemuste kättesaamiseks kasutajale dünaamiline webilehekülg.

Kasutatud kirjandus

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389-402, 1997.
2. Burge C. B. Chipping away at the transcriptome. Nature Genetics 27:232-234, 2001.
3. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. Mol Biol 268(1):78-94, 1997.
4. Clamp M., Andrews D., Barker D., Bevan P., Cameron G., et al. Ensembl 2002: accommodating comparative genomics. Nucleic Acid Res 31(1):38-42, 2003.
5. Collins J.E., Goward M.E., Cole C.G., et al. Reevaluating Human Gene Annotation: A Second-Generation Analysis of Chromosome 22. Genome Res 13:27-36, 2003.
6. Fickett JW, Guigo R. Estimation of protein coding density in a corpus of DNA sequence data. Nucleic Acid Res 21(12):2837-44, 1993.
7. Haas S., Vingron M., Poustka A., Wiemann S. Primer design for large scale sequencing. Nucleic Acid Res 26(12):3006-3012, 1998.
8. Henegariu O., Heerema N.A., Dlouhy S.R., Vance G.H., Vogt P.H. Multiplex PCR : Critical Parameters and Step-by-Step Protocol. BioTechniques 23:504-511, 1997.
9. Hubbard T., Barker D., Birney E., Cameron G., Chen Y., et al. The Ensembl genome database project. Nucleic Acid Res 30(1):38-41, 2002.
10. Kämpke T., Kieninger M., Mecklenburg M. Efficient primer design algorithms. Bioinformatics 17(3):214-25, 2001.
11. Makalowski W. The human genome structure and organization. Acta Biochim Pol 48(3):587-98, 2001.
12. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res 11(10):1725-9, 2001.
13. Pruitt KD., Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acid Res 29(1):137-140, 2001.

14. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet 16(1):44-7, 2000)
15. Reese M.G., Hartzell G., Harris N.L., Ohler U., Abril J.F., Lewis S.E. Genome annotation assessment in Drosophila melanogaster. Genome Res 10(4):483-501, 2000.
16. Rogic S., Mackworth A.K., Ouellette F.B.F. Evaluation of Gene-Finding Programs on Mammalian Sequences. Genome Res 11:817-832, 2001
17. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-86, 2000.
18. Rychlik W., Spencer W.J., Rhoads R.E. Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acid Res 18(21): 6409-6412, 1990.
19. Solovyev VV, Salamov AA, Lawrence CB. Identification of human gene structure using linear discriminant functions and dynamic programming. Syst Mol Biol 3:367-75, 1995.
20. Varotto C., Richly E., Salamini F., Leister D. GST-PRIME: a genome wide primer design software for the generation of gene sequence tags. Nucleic Acid Res 29(21):4373-4377, 2001.
21. Yeh R. F., Lim L.P., Burge C.B. Computational Inference of Homologous Gene Structures in the Human Genome. Genome Res 11:803-816, 2001.
22. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol 7(1-2):203-14, 2000.
23. Zhang M.Q. Computational Prediction of Eukaryotic Protein-Coding Genes. Nature Genetics 3:698-709, 2002.