BIOLOOGIA- GEOGRAAFIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Kadiliina Kuusik

Automaatne PCR praimerite disain kõikidele hiire geenidele

Keskastme lõputöö

Juhendanud: Maido Remm Prof., PhD

TARTU 2003

Sisukord

Sisukord 2
Lühendid ja mõisted 3
I Teoreetiline osa 4
Sissejuhatus 4
1. Genoomide annoteerimine ja andmebaasid 4
2. Hiire ja inimese genoomi võrdlus 6
3. Kordusjärjestused ja genoomi maskeerimine 9
3.1 DUST 12
3.2 RepeatMasker 12
3.3 GenomeMasker 13
4. PCR praimerite disain 14
4.1 Olulised PCR parameetrid 14
4.2 Praimerite seostumiskohtade kontroll 15
4.3 Praimerite disaini programmid 16
II Praktiline osa 18
Töö eesmärgid 18
1. Meetodid 18
1.1 Andmete päritolu ja struktuur 18
1.2 Kasutatud programmide kirjeldus ja tööpõhimõte 19
1.3 Vajalik mälumaht ja tööaeg 21
2. Tulemused 22
2.1 Tulemuste kirjeldus 22
2.2.Andmebaasi loomine 28
2.3 Kasutajaliides 29
Arutelu 31
Kokkuvõte 34
Summary 35
Kasutatud kirjandus 36

Lühendid ja mõisted

Amplikon	ühe PCR–i produktina amplifitseeritav DNA järjestus, töös kasutatud sünonüümina blokkide mõistet
Bp	aluspaar, nukleotiidialuste paar
Flanking piirkond	amplikoni mõlemalt poolt külgnev järjestus praimerite seostumiseks
Gb	gigabait, 10^9 bp
kb	kilobait, 10^3 bp
Mb	megabait, 10^6 bp
PCR	polümeraasi ahelreaktsioon *(Polymerase Chain Reaction)*
Sensitiivsus	suurus, mis kajastab algoritmi võimet leida "õigeid" positiivseid (Sn, *sensitivity*)
"shotgun" sekveneerimis-meetod	DNA lõhutakse paljudeks väikesteks tükkideks ehk tihti nimetatud kontiigideks, superkontiigideks, ultrakontiigideks
Tm	sulamistemperatuur, temperatuur, mille juures pool DNA ahelatest on üheahelalised ja pool kaheahelalised
Transpositsioon	DNA segmendi ülekandumine uude asukohta kromosoomis või plasmiidis
Transposoonsed elemendid	heterogeensed elemendid, mis võivad liikuda ühest kromosoomi alast teise
UTR	eksoni mittetransleeritavad alad (*untranslated region*)

I Teoreetiline osa

Sissejuhatus

Tänapäeval on saadaval suur hulk erinevate genoomide järjestusandmeid. Me saame vaadelda organismide genoome tervikuna ja püstitada kogu genoomi hõlmavaid küsimusi. Pidevalt täiustuv kaasaaegne tehnoloogia võimaldab läbi viia suuremahulisi eksperimente. Kuna tegemist on tõesti suurte andme hulkadega, siis efektiivse järjestusanalüüsi tarbeks on meil vaja automaatseid vahendeid.

Käesoleva töö eesmärgiks on luua terviklik lahendus hiire geenide PCR praimerite automaatseks disainiks, et meil oleks võimalus huvipakkuvad geenid genoomselt DNA–lt üles amplifitseerida kiirelt ja mugavalt.

Kuna inimese ja hiire genoomid on peaaegu sama suured, jagades ka ligikaudu samasugust geenikomplekti, on hiire genoomi järjestus võtmevahendiks inimese genoomi tõlgendamiseks, lahtimõtestamiseks, ja mis peamine, inimese geeni funktsioonide identifitseerimiseks ja uurimiseks.

1. Genoomide annoteerimine ja andmebaasid

Hiire ja inimese genoomijärjestuse võrdlemisel on täheldatud väga kõrget sarnasust. Nimelt on tõestatud, et 40% nukleotiidsel tasandil inimese genoomi järjestusest saab joondada hiire genoomi omale (Waterston *et al*.2002). Samuti on praeguseks teada üle ühe miljoni hiire inimese vahelise konserveerunud elemendi (Couronne *et al*. 2003). Et üldse saaks toimuda hiire ja inimese genoomide võrdlus, on meil vaja, et genoomid oleksid annoteeritud.

Genoomide annoteerimine saab toimuda kahel erineval viisil: manuaalselt ja automaatselt. Hoolimata sellest, et manuaalsel annoteerimisel saadud tulemused on kahtlemata kõige usaldusväärsemad, on sel moel, arvestades ajakulu, peaaegu et võimatu ka lähima viiekümne aasta jooksul saada terviklikku pilti elusloodusest ja selle evolutsioonist. Tänapäevane kiire arvutitehnika areng on võimalikuks teinud

automaatse annoteerimise, mille tarbeks on koostatud palju programme ja nende poolt kasutatavaid algoritme. Viimaste töötamise korrektsus on väga oluline, sest genoomijärjestuse väärtus on üks-üheses seoses annoteerimiskvaliteedi tasemega (Reese *et al.*, 2000).

1990. aastal loodi HGP (Human Genome Project) , kuhu kaasati ka hiir kui üks viiest mudelorganismist. Eesmärgiks oli luua hiire liini C57BL6/J nii geneetiline-, füüsiline-, kui ka genoomijärjestuse kaart. 2000. aasta oktoobris võttis hiire genoomi täieliku annoteerimise enda kanda MGSC (*Mouse Genome Sequencing Consortium*), mis algselt koosnes kolmest suuremast annoteerimiskeskusest: MIT (*the Whitehead/Massachusetts Institute of Technology*), Center of Genome Research ja Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/). Nendega teeb koostööd rahvusvaheline andmebaas ENSEMBL. Praeguseks on hiire genoomi annoteerimisega liitunud veel paljud teadlased ja institutsioonid üle kogu maailma. MSC (*Mouse Sequencing Cosortium*) sekveneerimine põhineb "shotgun" meetoditel. Järjestuse andmed on lühikestes 500-700 bp fragmentides. Praeguse seisuga arvatakse, et annoteeritud on 96% hiire eukromatiidsest DNA järjestusest. Hiire genoomi järjestuse kokkukogutud andmed on koondatud suurteks superkontiigideks, mis on joondatud kromosoomide peale vastavalt hiire geneetilisele kaardile.

Hiire genoomi järjestus on vabalt kättesaadav NCBI (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov/) ja ENSEMBL (http://www.ensembl.org/) kodulehel. Mina kasutan oma töös ENSEMBL andmebaasi, mis pakub erinevate järjestuste mitmekesist analüüsi ja järjestuste ning andmete erinevaid visuaalseid vaateid. ENSEMBL andmebaasi laienemine toimub pidevalt ja seda ka erinevate võrdlevate genoomijärjestuste vaadete lisamisega. Lisaks hiirele on andmebaasis vabalt kättesaadavad veel seitsme organismi esialgsed genoomijärjestused (rott, sebrakala, fugu, sääsk, äädikakärbes, varbussid C. elegans ja C.briggsae). Programmid MEGABLAST ja SSAHA võimaldavad teostada kahe genoomi vahelisi sarnasuseotsinguid.

Annoteerimisprotsess on pideva täiustamise teel, kuna selle missiooniks on järjestusele omandada võimalikult palju infot, mis puudutab näiteks geeni asukoha ja struktuuri määramist, erinevate markerite ja regulatoorsete üksuste identifitseerimist jpm. Järjestuse määramise üheks esmärgiks on kindlasti identifitseeritud geenide funktsionaalne klassifitseerimine, millele aitavad kaasa eelkõige just homoloogiaotsingud, kus praegu on kõige enam levinumaks inimese genoomi

vastandamine hiire omale. Annoteerimisprotsessi kaugemaks eesmärgiks nimetaksin mina kogu elusloodusе annoteeritud genoomse DNA järjestuste andmebaase, mille kaudu, pole kahtlust, muutuks meie elu palju lihtsamaks.

2. Hiire ja inimese genoomi võrdlus

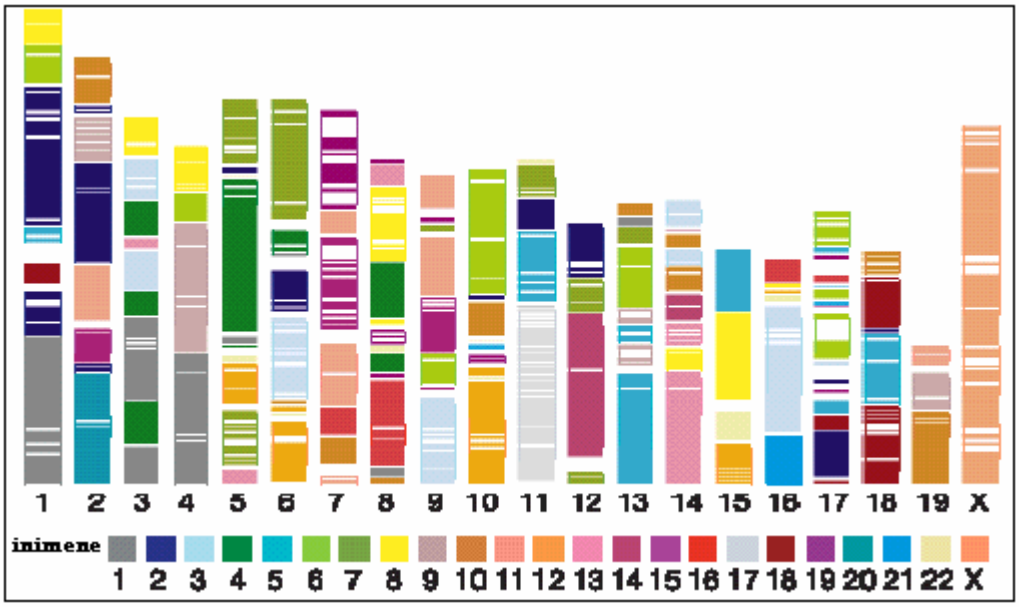
Üha enam kogub jõudu võrdleva genoomika suund, mis kahe organismi vahel paralleele tõmmates annab meile terviklikuma pildi bioloogilisest evolutsioonist kui sellisest üleüldse. Kuid miks vastandatakse inimese genoomile just hiire oma?

Esimeseks põhjuseks on see, et meil on olemas hiire genoomi järjestus. Teiseks, hiired on geneetiliselt modifitseeritavad. Samuti on hiired inimese genoomi uurimiseks parimad mudelorganismid tänu madalatele ülalpidamiskuludele ja kõrgele kasvukiirusele. Eelnevalt püstitatud küsimusele võiks veel vastata-, et tegu on kahe imetaja vaheliste väga suurte patoloogiliste mehhanismide sarnasusega, mis kiirendab biomeditsiinilisi uuringuid veelgi.

Järgnevalt annaksin ülevaate inimese ja hiire genoomi erinevustest. Vaatluse all on kahe organismi vahelised konserveerunud alad, sünteenilised segmendid ja teised genoomile omased osad.

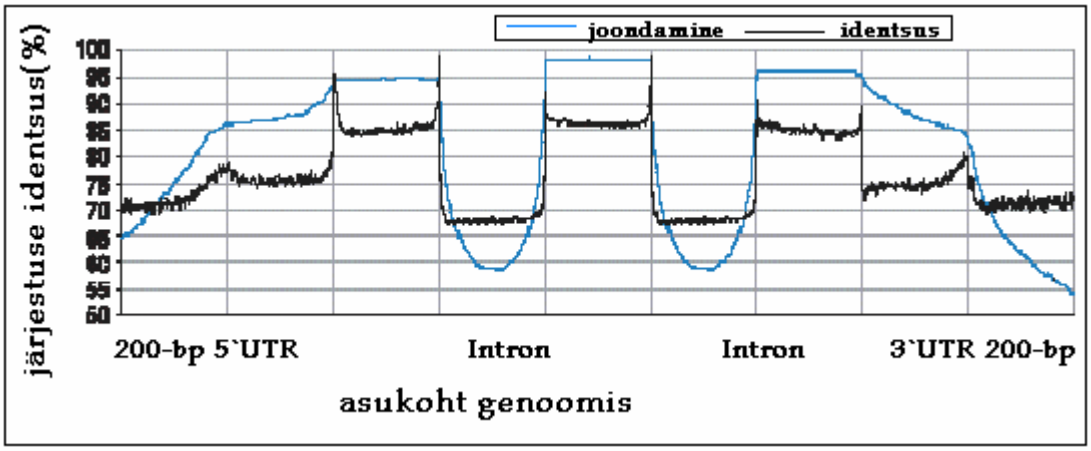
Hiire genoom sisaldab 2.5 miljardit aluspaari. Inimese genoom on 22% suurem, sisaldades 3.2 miljardit aluspaari.

90.2% inimese ja 93.3% hiire genoomist asetseb n.n. konserveerunud sünteenilistes segmentides. Liikide vahel on sünteenia (*synteny*) mitte ainult tänu ortoloogsete geenide olemasolule, vaid ka tänu nende samasugusele asetsemisele genoomis (Waterston *et al*. 2002). See on väga tugevaks eelduseks viitamaks ühisele eellasele *(common ancestor).* See tähendab seda, et me ei näe mitte ainult homoloogiat liikide vahel, vaid ka seda, et geenid asuvad kromosoomi peal samas järjekorras. Eksisteerivad samad sünteenilised segmendid, mis on aga eri kromosoomide vahel ümber organiseeritud . Sünteenilisi segmente on inimese ja hiire vahel ~340 (Hudson et al. 2001; Nadeau ja Taylor 1984). Ühe sünteenilise segmendi pikkus on keskmiselt 6.9 Mb. Seda illustreerib joonis 1.



Joonis 1. Näitab, kus samad sünteenilised segmendid on hiire ja inimese eri kromosoomide vahel ümber organiseeritud. Erinevad värvid sümboliseerivad erinevaid kromosoome inimese genoomis. 314 sünteenilist blokki on eraldatud valge joonega. Sünteenilisi segmente on kokku 217. Näiteks, osa hiire 2. kromosoomist moodustab täielikult inimese 20. Inimese 17. kromosoom moodustab osa hiire 11.ndast kromosoomist jne (Waterston *et al*., 2002).

96% hiire ja inimese geenidest asuvad samades sünteenilistes kohtades. Joonis 2 näitab konserveerumistaset geeni eksonites ja intronites.



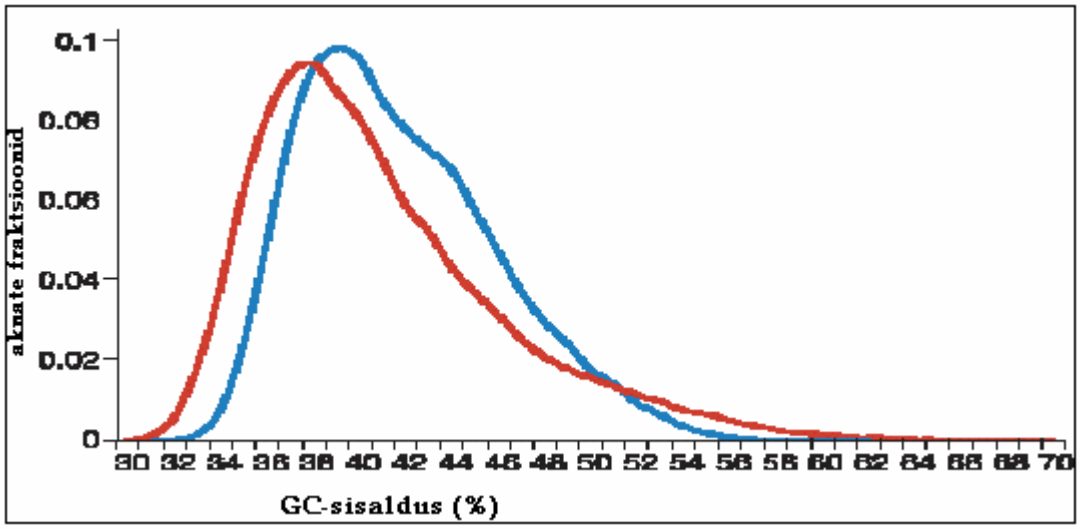
Joonis 2. Näitab, kui sarnased on hiire ja inimese sama geeni järjestused joondamisel, mida funktsionaalsem piirkond (eksonid), seda kõrgemalt on ta konserveerunud (Waterston *et al*., 2002).

Ortoloogid on geenid, mis on sarnased ja mis on pärinenud samast eellasest. Hiire ja inimese ortoloogsete geeni paaride vahel on ka geeni struktuur väga kõrgelt konserveerunud, nii eksonite arv kui pikkused. Kui ortoloogsetel eksoni paaridel on erinevad pikkused, siis on need kodeerivast raamist tulenevalt just kolmekordsete erinevustega. Ortoloogsed intronid ei oma tavaliselt identseid pikkusi, küll aga eksisteerib korrelatsioon genoomi ja introni suuruse vahel ehk siis hiirel on väiksem genoom, lühemad intronid. Vähem kui 1% hiire geenidest (118) ei oma homoloogi inimeses, küll aga teistes liikides. 80% hiire geenidest omavad üks-ühest ortoloogi inimese genoomis ja ülejäänud 20% ei ole samuti unikaalsed ainult hiire genoomile (Waterston *et al.* 2002).

tRNA geene on nii hiires kui inimeses peaaegu samapalju, vastavalt 335 ja 345. Hiire puhul on tõeliste tRNA-de kindlaks määramine raske, kuna genoom koosneb tRNA-dest tulenevatest aktiivsetest SINE elementidest.

Väga huvitavad tulemused on ilmnenud inimese ja hiire genoomide võrdlemisel. Nendes genoomides on leitud väga palju konserveerunud alasid, mille funktsioon ei ole teada (Dermitzakis *et al.*2002). Väidetavalt moodustab selline konserveerunud osa inimese ja hiire genoomist 5%, millest ainult 1.5 % on kodeeriv. See tähendab seda, et 3.5 % moodustab mittekodeeriv osa, mis viitab konserveerumisele ka UTR-des, geeniekspressiooni kontrollelementides, RNA geenides, kromosoomi struktuuri mõjutavates elementides ja mujal (Dermitzakis *et al.*, 2002).

GC nukleotiidide sisaldus inimese ja hiire genoomide vahel oluliselt ei erine (hiire 42% inimese 41% vastu), küll aga selle jaotus. Hiirel on GC nukleotiidide jaotus ühtlasem, seega ei esine genoomi eri piirkondades suuri kõikumisi. Nii hiirel kui inimesel asuvad 75-80% geenidest genoomi GC-rikkamas pooles. GC-sisalduse jaotust näitab joonis 3.

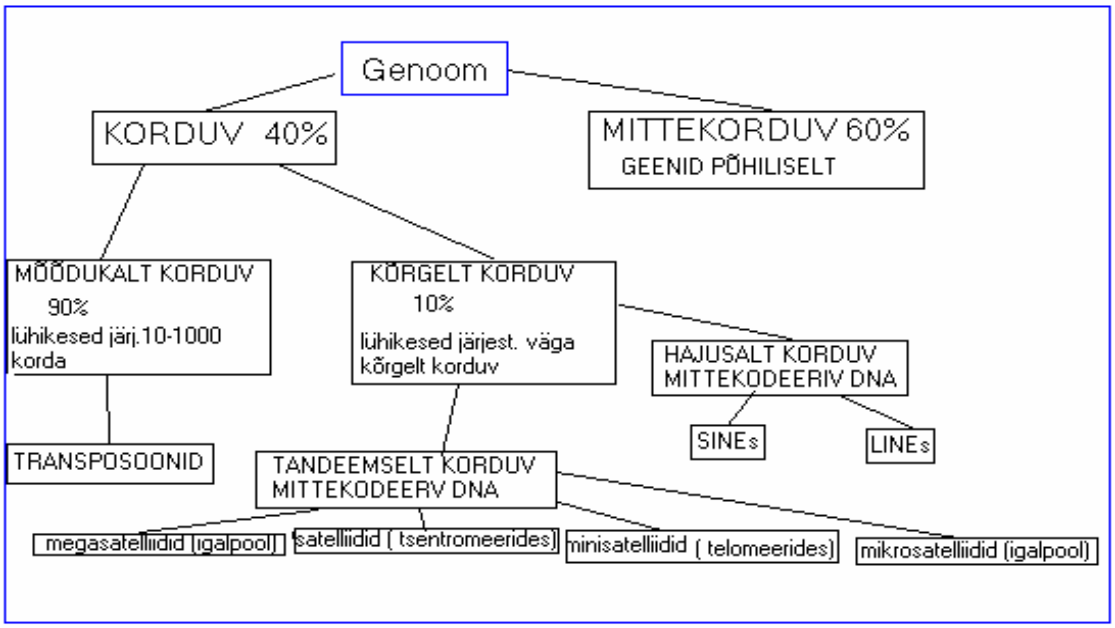


Joonis 3. X telje peal on GC-sisalduse %, Y telje peal 20 kb aknad, mis on asetatud suvalistele kohtadele genoomis. Punane joon iseloomustab inimese, sinine hiire GC-sisaldust. Nagu jooniselt näha on inimesel see jaotus laiem.

3. Kordusjärjestused ja genoomi maskeerimine

Üldjoontes on kordusjärjestuste sisaldus imetajate genoomides sarnane. Kõikidel imetajatel on neli põhilist klassi transposoonseid elemente: LINE *(long interspersed nuclear element)*, SINE *(short interspersed nuclear elements)*, LTR *(retrovirus-like elements with long terminal repeat)* ja DNA transposoonid. Erinevused hiire ja inimese genoomi vahel näitavad meile transposoonsete elementide dünaamilist iseloomu, sarnasused aga liikide bioloogilist evolutsiooni.

Hiire/ imetajate genoomis on umbes 60% mittekorduvat, unikaalset DNA-d. Ülejäänud järjestustest on umbes 30% mõõdukalt korduvat ja 10% kõrgelt korduvat DNA-d. Kordusjärjestused asuvad genoomi ekstrageenses alas ja neid võib liigitada mõõdukalt- ja palju-korduvaks DNA piirkondadeks. Mõõdukalt korduvat DNA-d võib veel jagada genoomis rühmiti e. tandeemseteks kordusteks (*tandem repeats*) ja hajusalt laiali asetsevatele e. hajuskordusteks (*interspersed repeats*). Jaotust näitab joonis 4.



Joonis 4. Imetajate genoomi jaotus korduvaks ja mittekorduvaks osaks.

Eristatakse veel tandeemselt korduvat mittekodeerivat DNA-d ja tandeemselt korduvat kodeerivat DNA-d (tandeemsed geenid, nagu näit. ribosomaalsed geenid). Esimesena mainitud klassi võib jaotada nelja klassi satelliitideks: mega-, ja mikrosatelliidid, mis lokaliseeruvad erinevates genoomi piirkondades; minisatelliidid, mis on detekteeritavad kromosoomi telomeersetes alades; ja satelliit- DNA, mille funktsioon võib olla seotud tsentromeeride ja homoloogiliste kromosoomide ühinemisalade ehk kinetohooride stabiliseerimisega.

Hajuskordusjärjestuste alamklassist LINE on nii inimesel kui hiirel esindatud Kpn-perekond ehk L1 (pikad hajuskordused). Põhiline erinevus kordusjärjestuste esinemises hiire ja inimese genoomide vahel seisneb selles, et hiire puhul võib täheldada väga paljude liigispetsiifiliste kordusjärjestuste esinemist. Inimesel on ainsaks aktiivseks SINE-elemendiks Alu kordusjärjestus (hiirel 1.4 miljonit, inimesel 1.1 miljonit). Näiteks SINE perekond on hiires palju rohkem diversifitseerunud (B1, B2, ID, B4). Hajuskordused jaotuvad ühtlaselt üle genoomi. Eksisteerib n.ö. SINE/LINE muster, mida iseloomustab järgnev tabel 1.

Tabel 1. Hajuskorduste lokalisatsioon kromosoomides ja iseloomustus.

	SINE	LINE
Lokalisatsioon	R-vöödi alas	G-vöödi alas
aluspaarid	G+C rikkad	A+T rikkad
CpG saared	sisaldavad CpG saari	ei sisalda CpG saari
replikatsioon	varases S-faasis	hilises S-faasis
geenid	housekeeping geenid	koespetsiifilised geenid

DNA transposoonide diversiteet on omakorda hiire genoomis väiksem kui inimese genoomis. Näiteks on hiire genoomis näriliste-spetsiifilisi DNA transposoone ainult 4 klassi (MMAR1, URR1, RMER30, RChar1), inimese 14 vastu. Siin arvatakse, et kuna tegu on DNA transposoonide genoomide vahelise horisontaalse ülekandega viiruste ja teiste sisemiste parasiitide vektorite kaudu, siis on imetajate genoom üldse suhteliselt raskesti läbitav, näriliste oma veidi rohkemgi. Konserveerunud kordusjärjestuste esinemise protsent jääb aga kõvasti alla inimese omale. Seda illustreerib tabel 2.

Tabel 2.

	hiir (%genoomist)	inimene(%genoomist)
konserveerunud kordusjärjestused	5%	22%
liigi spetsiifilised kordusjärjestused	32%	24%

Tabelis näidatud tulemused on seletatavad hiire umbes kahekordselt kõrgema muteerumiskiirusega (vastavalt $4.5*10^{-9}$ ja $2.2*10^{-9}$ asendust iga nukleotiidi kohta aastas). Suur mutatsioonikiirus on omakorda seletatav peamiselt hiirte lühikese generatsiooniajaga. Põlvkondade arvu suurenedes kasvab võimalus meiootilise duplikatsiooni käigus erinevate vigade tekkimiseks. Teiseks põhjuseks võib ka olla transpositsiooni muutumise kiirus, mis hiire puhul on läbi aegade jäänud enam-vähem konstantseks, inimesel tõusnud märgatavalt kuni 40 miljonit aastat tagasi ja siis järsult pidurdunud (Waterston *et al*., 2002).

Imetajate genoomist moodustavad kordusjärjestused väga suure osa, jäädes enamus ektrageensesse genoomi osasse. Need piirkonnad on ebainformatiivsed ja seetõttu on oluline need piirkonnad kas geenide ennustamisel või parimerite disainimisel ära maskeerida. Maskeerimisel kasutatavatest programmidest tuleb juttu järgnevalt.

3.1 DUST

DUST on kordusjärjestuste maskeerimise programm, mille abil saame järjestusest välja filtreerida bioloogiliselt ebaolulised kohad (ftp://ftp.ncbi.nih.gov/pub/tatusov/dust/version1/). DUST maskeerib DNA järjestuses vähese keerukusega alad (*low complexity DNA*), s.t. alad, mille nukleotiidne koostis ei ole piisavalt mitmekesine (lühikesed ühe-, kahe-, ja kolmenukleotiidsed kordused). Algoritm on vaikimisi kasutatav BLASTN programmi poolt teostavates DNA homoloogiaotsingutes. Programmi autoriteks on Roman Tatusov ja David Lipman.

3.2 RepeatMasker

RepeatMasker on programm, mis kontrollib FASTA formaadis DNA järjestust teadaolevate kordusjärjestuste andmebaasi vastu, tagastades maskeeritud järjestuse, mis on valmis erinevateks andmebaasi otsinguteks ja vastava maskeeritud regiooni annoteerimiseks (http://repeatmasker.genome.washington.edu/RM/RepeatMasker).

Programmi väljundiks on sisendjärjestuse detailne korduste annotatsioon ja järjestus, kus kõik annoteeritud kordused on maskeeritud ehk asendatud N-dega. Üheks võimalikuks väljundfailiks on otsinguks kasutatud järjestuse joondus kokkulangevate kordusjärjestustega andmebaasis. Kasutajal on võimalus valida, milliste korduste vastu tahab ta maskeerimist sooritada. Otsingu sooritamiseks on põhimõtteliselt kolm võimalust. Vaikimisi maskeeritakse nii hajuskordusjärjestused kui ka lihtsad tandeemsed kordused, polypuriini, AT- ja/või GC rikkad piirkonnad ja mikrosatelliidid. "*Mask only simple repeats*" optsiooni korral saadakse tulemus ainult polümorfsete lihtsate korduste kohta. Vastupidiselt, "*do not mask simple repeats*" valiku korral maskeeritakse ainult hajuskordusjärjestused. Järjestuste võrdlemine viiakse läbi programmi cross-match abil, mis on Smith-Waterman-Gotoh algoritmi täiendus.

Hajuskordused on liigispetsiifilised, sõltudes transposoonsete elementide aktiivsuse ajast, millest nad on pärinenud. Peaaegu pooled kordused langevad kõikidel imetajatel kokku. Hiire genoomis on aga palju liigispetsiifilisi, närilistele omaseid korduseid. RepeatMaskeril on eraldi optimiseeritud protokoll analüüsimaks näriliste ja primaatide genoome, teiste imetajate jaoks vastavat võimalust veel ei ole.

Programmi kiirus ja sisendjärjestuse pikkus on lineaarses seoses, ehk mida lühem järjestus seda lühem tööaeg. Programmi tundlikkust on võimalik suurendada või vähendada kiiruse arvelt. Tundlikum otsing võtab kolm korda kauem aega, aga maskeeritakse 0-5 % rohkem kordusi kui vaikimisi. Kiire otsing on küll 3-5 korda kiirem, kuid võib märkamata jätta 5-10 % vaikimisi maskeeritud kordustest. Programmi autoriks on Phil Green.

3.3 GenomeMasker

GenomeMasker on programmipakett, mis on mõeldud korduste leidmiseks ja vältimiseks PCR-i praimerite disainil. GenomeMasker-i programmipaketti kuuluvad kaks korduvate järjestuste maskeerimise ja nende leidmise programmi GMASKER ja GTESTER (http://www.ismb02.org/posters/poster/Remm_1.pdf). GMASKER maskeerib FASTA formaadis järjestusest üleesindatud sõnad vältimaks praimerite disainil praimeri sattumist kordusjärjestusi sisaldavatele piirkondadele. Programm ei kasuta järjestuse joondamise algoritme, ja seega on tema töökiirus küllalt kõrge. Selle asemel kasutab ta programmi BLACKLISTER poolt koostatud üleesindatud sõnade nimekirja (*black list*). Sõna pikkust on võimalik varieerida vahemikus [8, 16]. Kasutades sellist nimekirja maskeerib programm GMASKER kõik korduvad piirkonnad. GMASKER-i maskeerimisprotsessi põhimõte on sama mis RepeatMasker-il, kuid esimese puhul võib tuua välja mõningaid eeliseid. Esiteks, otsing teostatakse fikseeritud sõnapikkuste vastu, mis on täpsem, kui otsingu teostamine levinumate kordusjärjestuste vastu. Teiseks, maskeerib GMASKER ainult üleesindatud sõnade 3`otsa, mis võimaldab praimerite disaini ka siis, kui praimeri 5` osa asub osaliselt kordusjärjestuses. Kogu sõna maskeerimisel võib juhtuda, et maskeeritakse liiga suured alad, mis võivad siiski sisaldada meile huvipakkuvat informatsiooni. Kuna maskeeritakse ainult korduses olevate sõnade 3` ots, on GMASKER-i maskeerimine asümmeetriline ja maskeeritud nukleotiidid võivad kummalgi DNA ahelal olla erinevad.

Programmi GTESTER abiga saab leida PCR-i produktide arvu genoomis.

4. PCR praimerite disain

Annoteeritud genoomide järjestuste kättesaadavus tekitab üha rohkem vajadust eksperimentaalsete lähenemiste järele, mis tegelevad suure hulga geenide analüüsiga. Tänapäeval on üsna vähe selliseid DNA-uuringuid, kus ei kasutataks polümeraasset ahelreaktsiooni ehk PCR-analüüsi. Olulisel kohal PCR-i õnnestumises on sobivalt valitud praimeripaar ehk kaks praimerit, mis kinnituvad meid huvitava DNA piirkonna 5`ja 3` otstele ning millest algab DNA süntees. PCR õnnestumine on otseses sõltuvuses PCR kvaliteeti mõjutavatest parameetritest, milleks on kasutatud praimerite sulamistemperatuur, PCR produktide pikkus, praimerite GC-sisaldus ning praimerite omavahelised interaktsioonid ja seostumiskohad genoomis. Seega, et praimerid ei seostuks genoomsel DNA-l juhuslikku kohta, peavad nad olema väga spetsiifilised. Probleemi lahendamiseks on loodud palju programme ja algoritme, mis tegelevad automaatse PCR praimerite disainiga.

4.1 Olulised PCR parameetrid

Kõige olulisemad PCR praimerite disaini juures arvestatavad parameetrid on sulamistemperatuur (Tm) ja produkti pikkus. Sulamistemperatuuri ligikaudseks arvutamiseks kasutatakse valemit (Kämpke *et al*., 2001):

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

kus on näha, et sulamistemperatuur sõltub ainult nelja erineva nukleotiidi esinemissagedusest järjestusel. Vastavalt valemile, lisab iga A/T paar juurde 2 ºC, G/C paar aga 4 ºC. Seega, mida GC rikkama sisaldusega järjestus on, seda rohkem vesiniksidemeid on vaja lõhkuda ja seda kõrgem on sulamistemperatuur. DNA dupleksi sulamistemperatuuri arvutamise täpsemaks algoritmiks on *Nearest-Neighbor meetod* (Borer *et al*., 1974).

$$Tm^{primer} = \frac{dH}{dS + R\ln(c/4)} + 16.6\lg\frac{[K^+]}{1+0.7[K^+]} - 273.15$$

kus H- entalpia (kirjeldab protsessi soojussisaldust), S- entroopia (kirjeldab süsteemi korrapäratust), R- universaalne gaasikonstant (R = 8.31J/mol), c- oligonukleotiidide totaalne molaarne konsentratsioon (Rychlik *et al.*, 1990). Breslauer koos kaasautoritega on välja arvutanud kõigi võimalike nukleotiidipaaride seostumise kombinatsioonidel tekkivad energia väärtused (Breslauer *et al.*, 1986). Ideaalseks praimeri pikkuseks loetakse vahemikku [20, 25]. Optimaalne sulamistemperatuur langeb vahemikku [50, 70].

Erinevate PCR meetoditega saab üles amplifitseerida erineva pikkusega DNA fragmente. Produkti pikkus on tavaliselt vahemikus 100 kuni 3000 bp. Kuigi eksperimentaalselt on tõestatud, et lühemate PCR produktide korral on PCR-i tulemused paremad, on viimasel ajal arendatud välja mitmeid PCR-i meetodeid just pikkade, kuni 6kb pikkuste DNA fragmentide amplifitseerimiseks (Kämpke *et al.*, 2001).

Praimerite GC sisaldus peaks olema 50-55%. Väiksema GC sisaldusega praimereid tuleks pikendada vastavalt nii palju, et vajalik sulamistemperatuur oleks üle 50° C. Praimeri sptsiifiliseks seostumiseks DNA järjestusega on oluline tema 3`otsa GC nukleotiidide sisaldus. 3`otsa nukleotiidse sisalduse optimaalsuse printsiip ei ole veel täielikult välja töötatud, sest mõned teooriad väidavad, et õige oleks kasutada GC rikast 3`otsa, mõned jälle, et kui see on väga kõrge, siis võib põhjustada vale seostumise (Beasley *et al.*, 1999; Li *et al.*, 1997). On väidetud, et praimeri 3´otsas peaks selle stabiliseerimiseks olema G või C nukleotiid (Henegariu *et al.*, 1997). Oluline on ka sekundaarstruktuuride vältimine.

4.2 Praimerite seostumiskohtade kontroll

Unikaalse PCR-i produkti saamiseks tuleb disainitud praimereid testida, sest peale selle, et nad seostuvad uuritava genoomse DNA-ga, võivad praimerid hübridiseeruda iseendaga või vastaspraimeriga, samuti on võimalik praimeri alternatiivne seostumine genoomsel DNA-l. Rohkema kui ühe produkti amplifitseerimine võib anda vale-positiivseid signaale genotüpiseerimisel. Viimase põhjustajateks on põhiliselt praimerid, mis sisaldavad DNA kordusjärjestuste motiive. Seega, enne praimerite disaini on oluline kordusjärjestuste maskeerimine (vt. teema 3).

Praimerite seondumiskohtade arvu ja asukoha leidmiseks saab kasutada erinevaid programme, millest tuntumad on BLAST (*Basic Local Alignment Search Tool*) (Altschul et al., 1997), MEGABLAST (Zhang *et al*., 2000), SSAHA (*Sequence Search and Alignment by Hashing Algorithm)* (Ning *et al*., 2001) ja GTESTER GenomeMaskeri paketist (Reppo *et al*., unpublished). BLAST (http://www.ncbi.nih.gov/blast/) on mõeldud sarnasuse otsinguteks erinevatest järjestuse andmebaasidest. BLAST-i üheks tugevaks küljeks on mitmekülgsus, kuna tal on palju erinevaid alamprogramme erinevat tüüpi otsingute jaoks. Üks BLAST-i variantidest on MouseBLAST (http://mouseblast.informatics.jax.org/) , mis on mõeldud hiire-, roti- ja teiste näriliste andmebaasidest inimese järjestusega kokkulangevate järjestuste otsimiseks. Pikkade järjestuste uurimisel kasutatakse MEGABLAST-i. SSAHA (http://www.sanger.ac.uk/Software/analyses/) on BLAST tüüpi programmidest kiireim, paisates järjestusinformatsiooni paisktabeli (*hash)* kujulisse andmestruktuuri, kust edasi toimub efektiivne sarnasuste otsing. Suur kiirus aga saavutatakse mälumahu arvelt (programm vajab vähemalt 1GB muutmälu). Kõige sobivam PCR-i praimerite testimiseks on spetsiaalselt selleks otstarbeks kirjutatud programm GTESTER, mis loendab genoomis kõik vastava praimeripaari esinemiskohad ja testib mitu PCR produkti nad tekitavad. Praimeri seostumiskohtade leidmiseks genoomil võib kasutada MEGABLAST-i või SSAHA-d.

4.3 Praimerite disaini programmid

Tabel 3. Näidatud on valik praimerite disainimise programmidest. Enamus praimeri disaini programme on kommertsiaalsed. Tärniga on märgitud tasuta saadaval olevad programmid akadeemilisele kasutajale. Eraldi on jaotatud programmid suuremahuliste projektide jaoks ja väiksemate koguste praimerite valmistamiseks.

valik PCR praimerite disainimise programmidest	
nimi	aadress
suuremahulised projektid	
PRIMER3 *	http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
PRIDE	http://www.dkfz-heidelberg.de/tbi/services/Pride/prideform
GST- PRIME	leister@mpiz-koeln.mpg.de
väiksemad projektid	
OLIGO	http://www.oligo.net/
PRIME	http://www.accelrys.com/products/gcg_wisconsin_package/program_list.html#Primer

PRIMERSELECT	http://www.dnastar.com/products/PrimerSelect.html

Suuremamahuliste projektide jaoks on mõeldud PRIMER3 (Rozen *et al*.,1998), PRIDE (Haas *et al*., 1998), GST-PRIME (Varotto *et al*., 2001) jt. Väiksemate praimerikoguste disainimiseks on programmid- OLIGO (Rychlik *et al*., 1989), PRIME, PRIMERSELECT jt. Teine programmide klassifitseerimise võimalus on nende automatiseeritus ehk kui palju peab praimeri disaini juures vahele sekkuma kasutaja. Heaks näiteks on siinkohal PRIDE, kus ei eemaldata ühtegi praimeri kandidaati enne, kui on välja arvutatud kõik parameetrid kõikidele praimeritele (Haas *et al*., 1998). GST-PRIME-i teeb eelistatuks see, et disainimist saab alustada juba valgu järjestuse tasemel. PRIMERSELECT pakub palju võimalusi nii praimerite disainimiseks kui kasutaja enda olemasolevate praimerite analüüsimiseks. Vajaduse korral saab kasutaja PRIMERSELECT-i poolt vaikimisi valitud termodünaamilistele parameetritele valida erinevate alternatiivide vahel. Tulemuste analüüsimise hõlpsustamiseks on illustratsioonid, graafikud, statistilised aruanded. OLIGO on samuti laialt kasutatav kommertsiaalne programm, mis on väga mitmekülgse kasutajaliidesega ja mida on täiustatud pidevalt juba aastast 1989.

II Praktiline osa

Töö eesmärgid

Käesoleva keskastme lõputöö eesmärgiks oli luua terviklik lahendus kõikide hiire geenide automaatseks praimerite disainiks. Töö sisaldas olemasolevate praimerite disainimise ja testimise programmide kombineerimist nii, et praimerid leitakse võimalikult suurele osale eksonitest, tulemuste viimist andmebaasi ning webilahenduse ehk kasutajaliidese loomist. Kaugemaks eesmärgiks on programmi tulemuste õigsuse kontrollimine praktikas ja selle pidev täiustamine.

1. Meetodid

1.1 Andmete päritolu ja struktuur

Geenide asukoha andmed antud töö tarbeks tõmbasin annotatsioonikeskusest ENSEMBL, kasutades EnsMart webiliidest MartView. Töö teostamise hetkel oli kättesaadav hiire andmebaasi versioon 17.30.1. Tabel 4 näitab antud töö tegemise käigus kehtivaid andmeid hiire genoomi andmebaasi kohta, võrdlevalt inimese genoomi andmetega.

Tabel 4. Näidatud on antud töö käigus kehtivad andmed hiire ja inimese geenide, eksonite, transkriptide, kontiigide ja kloonide arvu kohta. Nii genoomi suuruse, geenide, eksonite ja transkriptide arvu osas suuri erinevusi ei esine.

	hiire genoom versioon 17.30.1	inimese genoom versioon 18.33.1
Ensembl geeniennustus	24 948	24 261
Genescan geeniennustus	94 075	185 691
Ensembl eksonid	217 228	226 669
Ensembl transkriptid	32 911	32 997
kontiigid	136 483	26 679
kloonid	136 481	46 981
bp	2 497 038 443	3 232 141 503

Kõikide kromosoomide kohta jäädvustati järgnev info geenide asukoha kohta: geeni ID, eksoni ID, eksoni algus- ja lõppkoordinaat, kromosoomi nimi. Iga kromosoomi andmed jäädvustati eraldi faili ja sorteeriti asukoha järgi.

1.2 Kasutatud programmide kirjeldus ja tööpõhimõte

Minu poolt koostatud programmid võib jagada üldjoontes kaheks vastavalt erinevale arvutusalgoritmile. Ühed programmid teostavad praimerite leidmise üle kogu genoomi või valitud genoomi piirkonna. Teine programmipakett genereerib geenispetsiifilisi praimereid etteantud genoomi piirkonnast.

Tabel 5. Näidatud on programmipakettides olevate programmide nimed. Kui programmi nimes on laiend _chr, siis on tegemist praimerite disainimisega piirkonnale, mis ei pea tingimata olema pärit ühest geenist; kui nimes laiendit ei ole või on laiend _gene, siis on tegemist piirkonnaga ühest geenist, mida tahetakse üles amplifitseerida.

üle kogu genoomi	Geenispetsiifiliselt
blocks_chr.pl	blocks_gene.pl
primers_chr.pl	primers_gene.pl
gtest_chr.pl	gtest.pl
primers_chr100.pl	primers_gene100.pl
gtest100_chr.pl	gtest100.pl

Esimeseks sammuks pärast geenide ja eksonite asukoha andmete tõmbamist on nii-nimetatud blokkide või amplikonide moodustamine (teostatakse programmide blocks_chr ja blocks_gene poolt). Amplikonide moodustamine toimub eksonite algus- ja lõppkoordinaatide järgi. Kuna paljud eksonid on väga lühikesed ja PCR tehnoloogia lubab kvaliteetselt üles amplifitseerida ka palju suuremaid DNA fragmente, siis eksonite liitmise teel saadakse amplikonid maksimaalse pikkusega kuni 3000 bp. Üksikud eksonid on lubatud vahemikust pikemad, seega tuleb nad jaotada mitme amplikoni vahel. Seejärel disainitakse vastavale regioonile praimerid (kasutades programme primers_chr.pl ja primers_gene.pl). Need programmid kasutavad kõigepealt programmi DUST, et maskeerida lihtsad kordusjärjestused ja siis programmi GMASKER, mis maskeerib mustas nimekirjas esinevaid 16-nukleotiidi pikkusi kordusi. Seejärel konverteeritakse saadud järjestus programmi GM_PRIMER3 (kohandatud PRIMER3) sisendfailiks. GM_PRIMER3 sisendfail

sisaldab nii PCR-i parameetrite lubatud varieeruvuse vahemikke kui ka optimaalseid väärtusi.

Tabel 6. Näidatud on meie poolt kasutatud PCR–i parameetrite väärtused. PRIMER_MAX_DIFF_TM väärtus 4 tähendab, et maksimaalne praimeripaari praimerite sulamistemperatuuride erinevus võib olla 4 °C. PRIMER_MAX_POLY_ X väärtus 4 tähendab, et vastavas piirkonnas võivad järjestikku olla maskimaalselt 4 ühesugust nukleotiidi. Tabeli viimane rida näitab, et igale amplikonile tahame leida ainult ühte praimeripaari.

Põhilised PCR-i parameetrite väärtused		
	SIZE_RANGE	SIZE_OPT
PRIMER_PRODUCT_SIZE	100- 20000	600
PRIMER_SIZE	16- 30	21
PRIMER_TM	59- 65	62
PRIMER_MAX_DIFF_TM	4	
PRIMER_GC_PERCENT	20- 80	35
PRIMER_MAX_POLY_X	4	
PRIMER_NUM_RETURN	1	

Paraku ei suuda GM_PRIMER3 ühe sellise tsükli tulemusena kõigile amplikonidele häid praimereid leida, seega korratakse protsessi tsükliliselt. Iga uue tsükli käikuminekuga pikendatakse külgnevate piirkondade pikkusi 100 nukleotiidi võrra. Pärast 4. tsüklit toimub kordusjärjestuste maskeerimine vähem range kordusjärjestuste nimekirjaga. Sobiv praimeripaar püütakse leida kõikidele amplikonidele. Selliseid tsükleid korratakse kuni 12 korda.

Järgmiseks sammuks on saadud praimeripaaride unikaalsuse kontrollimine üle kogu genoomi (kasutades programme gtest_chr.pl ja gtest.pl). Võimalik, et praimeripaaril on rohkem kui üks seondusmiskoht. Viimased eemaldatakse ja nende asemele disainitakse 100 täiendavat praimeripaari samale amplikonile (kasutades programme primers_chr100.pl ja gtest100.pl). 100 erineva praimeripaari disainimise ja tsükliliselt pikendatavate külgnevate piirkondade abil püütakse leida optimaalseid praimeripaare kõikidele amplikonidele.

1.3 Vajalik mälumaht ja tööaeg

Kui räägime kogu loodud programmipaketist kui tervikust, siis pakub huvi palju on minimaalne selle kasutamiseks vajalik mälumaht. Seega, oluline on, kui palju võtab mälu kõige mälumahukam etapp. Praimeri disainimisega tegeleval programmil on selleks GMASKER, mille töötamiseks vajalik arvuti mälumaht on umbes 80-90 Mb. Praimeripaaride seostumise unikaalsuse kontrollimisega tegeleval programmil aga 400-450Mb ja seda põhiliselt tänu programmile GTESTER.

Programmide tööajad on toodud tabelis 7. Võib näha, et GTESTER on lisaks mälumahule ka tööaja suhtes kõige nõudlikum ja teda kasutavatel programmidel on kõige pikem tööaeg.

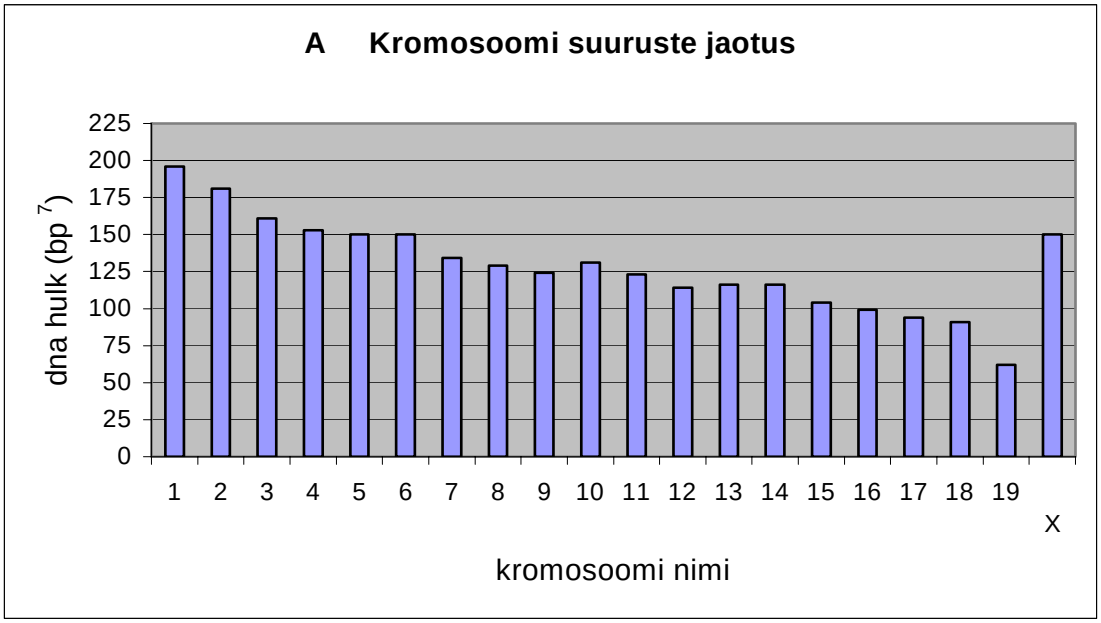
Tabel 7. Tööajad tundides.

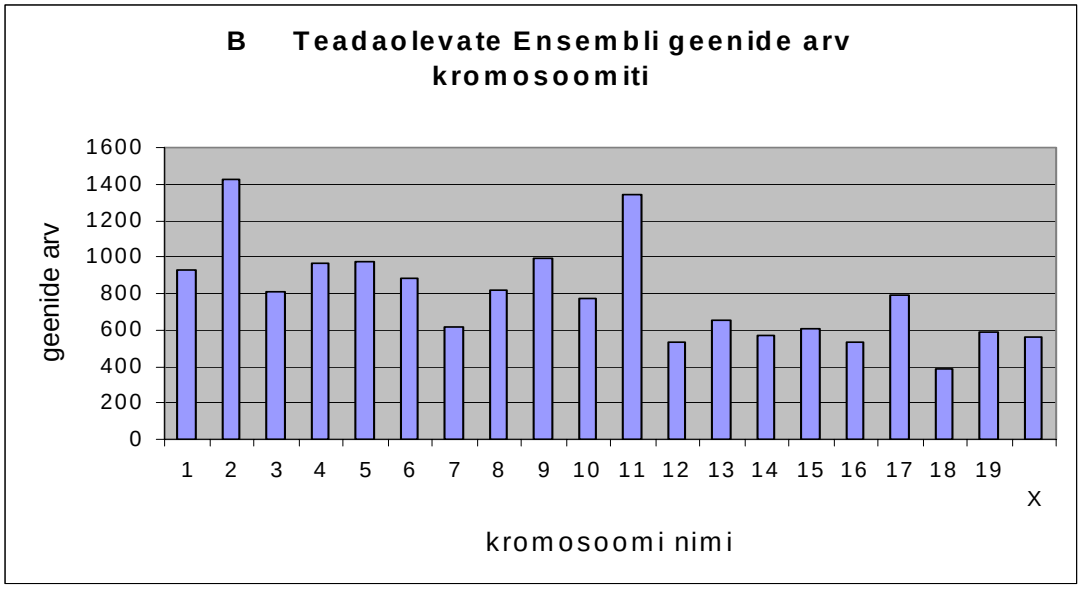
Programmide	umbkaudne tööaeg tundides		
blocks_chr.pl	0,08	blocks_gene.pl	0,08
primers_chr.pl	7	primers_gene.pl	8
gtest_chr.pl	10	gtest.pl	11
primers_chr100.pl	2,5	primers_gene100.pl	2
gtest100_chr.pl	12	gtest100.pl	10
TOTAL	31,58		31,08

2. Tulemused

2.1 Tulemuste kirjeldus

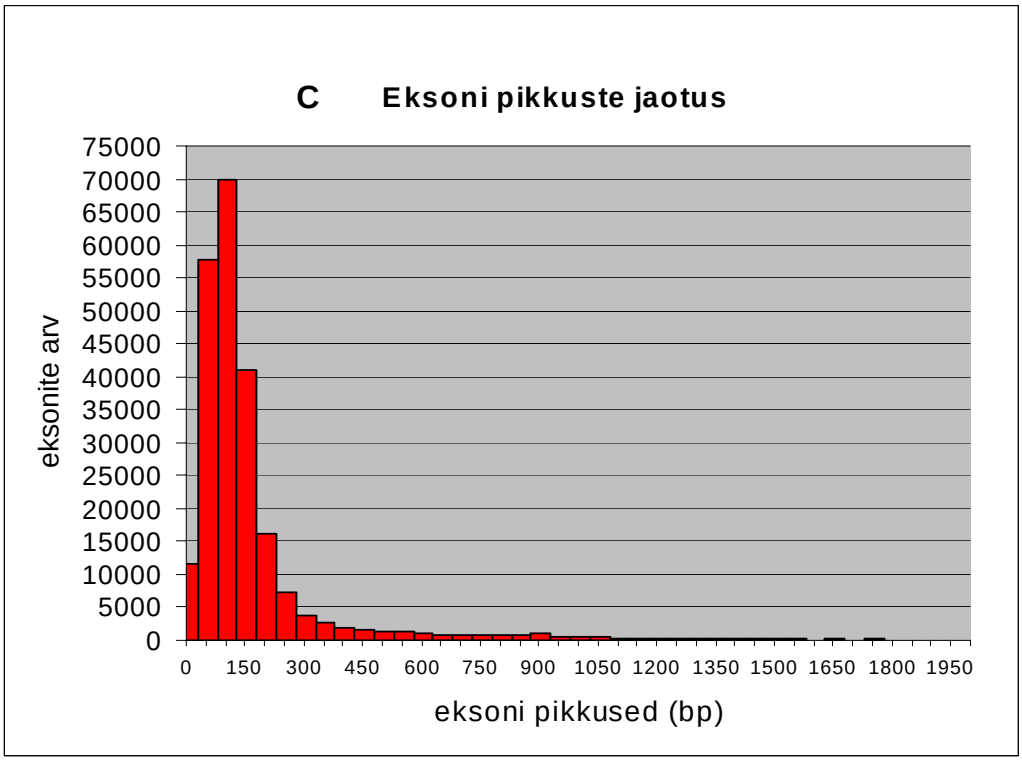
Hiire genoomil on 19 kromosoomi pluss X ja Y kromosoom. Y kromosoom on nii vähe sekveneeritud, et töö käiku polnud mõtet teda kaasata. Järgnevad joonised näitavad, et mingit korrelatsiooni kromosoomi suuruste ja nendes sisalduva geenide hulga vahel ei ole. Võrreldes jooniseid A ja B võime näha näiteks, et kõige suurem kromosoom on küll 1., kuid geenikomplekti hulga poolest jääb ta kõvasti alla 2. ja 11. kromosoomile.





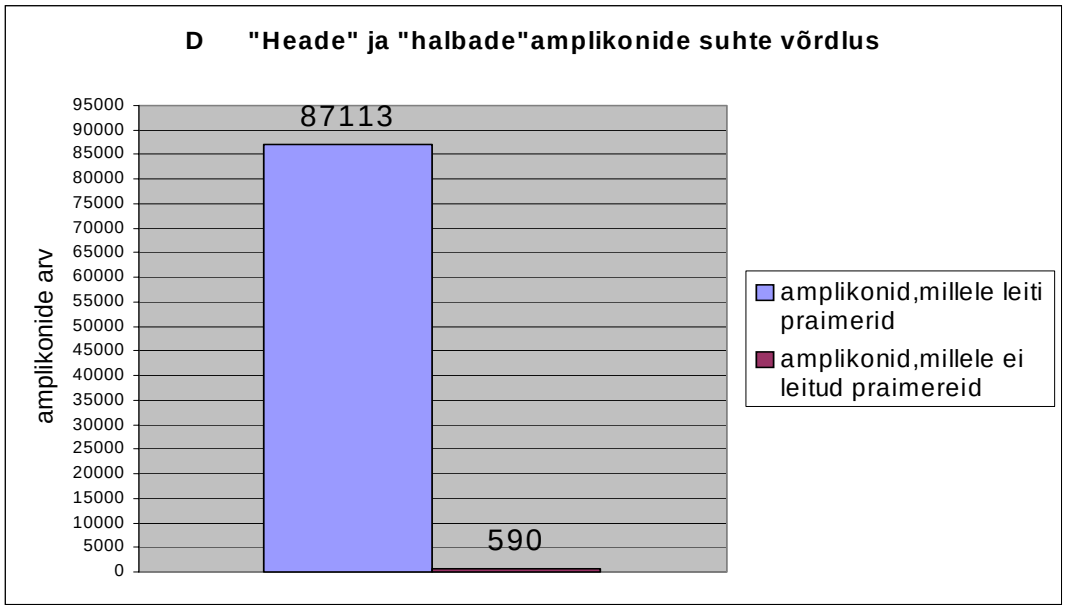
Joonised A, kromosoomi suuruste jaotus ja joonis B, teadaolevate ENSEMBL geenide arv kromosoomiti.

Geenid koosnevad mitte-kodeerivatest intronitest ja kodeerivatest eksonitest. Eksonite arv, struktuur ja tegelikult ka pikkused on eri orgasnismide vahel kõrgelt konserveerunud. Seda teemat puudutas ka antud töö teoreetilise osa 2. peatükk. Inimese ja hiire geeni struktuuri konserveeritus on väga kõrge. 86% ortoloogseid geene sisaldavad sama palju eksoneid, nendest 46%-l on ka eksonitel identne pikkus. Eksonite arvuks hiire ühes geenis võib lugeda 9, kuid see on väga varieeruv. Järgnev joonis näitab eksoni pikkuste jaotust vahemikus [0, 2000].



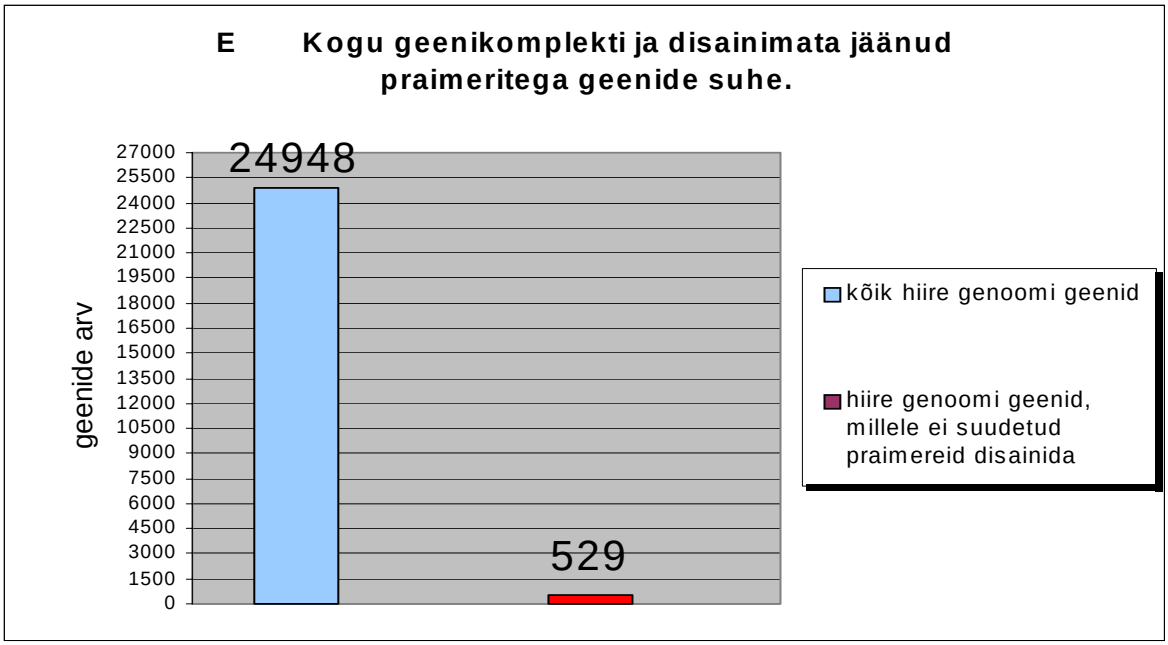
Joonis C. On näha, et suur enamus eksoneid on väga lühikesed, vahemikus [0, 250] bp.

Nagu näidatud joonisel C, on enamus eksoneid lühikesed, kuid on ka üksikuid selliseid, mis on pikemad kui 3000 bp. Maksimaalne lubatud pikkus ühe amplikoni jaoks on antud töös 3000 bp. Sellest pikemate eksonite puhul tuleb nad jaotada mitme amplikoni vahel. Enamus amplikone on aga siiski ainult natuke pikemad kui eksonid. Ühe amplikoni keskmine eksonisisaldus on kaks, mis on samuti väga varieeruv.

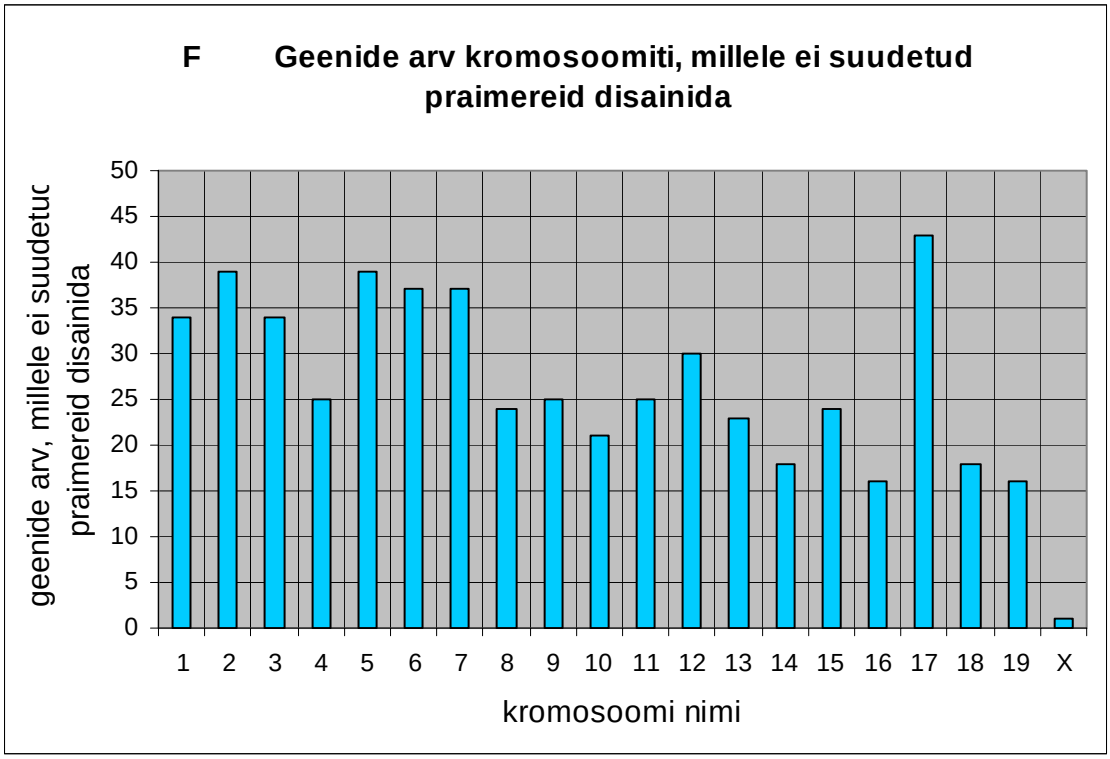


Joonis D, “heade” ja “halbade” amplikonide suhte võrdlus. “Head” amplikonid on sellised, millele õnnestus praimerid kontrueerida, “halvad” sellised, milles sisaldus siis vähemalt üks selline ekson, mille üles amplifitseerimiseks praimereid saada ei õnnestunud. Kokku oli üldse 87 703 amplikoni, millest “halvad” moodustasid 0.7%. Ja kui arvestada seda, et eksoneid on kokku 217 228, siis järelikult eksoneid, millele ei õnnestunud praimereid saada jäi kusagil 1.7%.

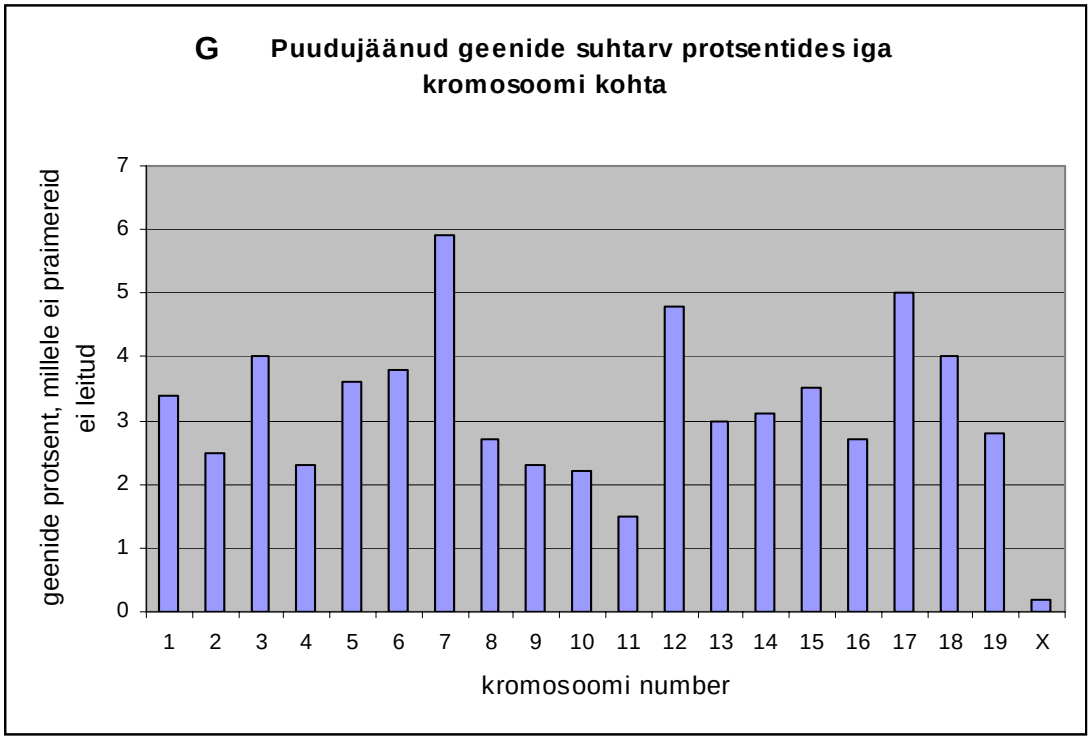
Tähtsam on tulemuste tõlgendamine geeni tasemel.



Joonis E, kogu hiire genoomi geenikomplekti ja disainimata jäänud praimeritega geenide suhe. Praimerid jäid disainimata ligikaudu 2%-le geenidest.



Joonis F, geenide arv kromosoomiti, millele ei suudetud praimereid disainida. Keskmiselt oli iga kromosoomi kohta kusagil 15 sellist geeni, seda võib vaadata korrelatsioonis kromosoomide geeni arvuga üleüldse, ja sellest selgub, et lineaarset seost, et mida rohkem geene, seda enam leidus ka selliseid, mille praimereid ei leitud, ei ole. Seda näitab järgnev joonis.



Joonis G. Näitab, et kromosoomiti oli geenide arv, millele praimereid ei leitud täiesti juhuslik.

2.2.Andmebaasi loomine

Käesoleva töö üheks eesmärgiks oli ka tagada kasutajale saadud tulemuste mugav ja lihtne kättesaadavus ehk graafilise kasutajaliidese loomine. Andmebaasi tegemisel kasutasime MySQL-i võimalusi. MySQL on andmebaasi haldamise programm, mis on võimeline talletama suurtes kogustes informatsiooni organiseeritud formaadis, olles kergesti kättesaadav erinevatele scriptimiskeeltele (http://www.mysql.com/documentation/).

Minu poolt koostatud töö andmebaasi nimeks on exonPCR_mouse_3, mis on organiseeritud 6-ks tabeliks. Andmebaasi tabelite struktuuri näitab tabel 8.

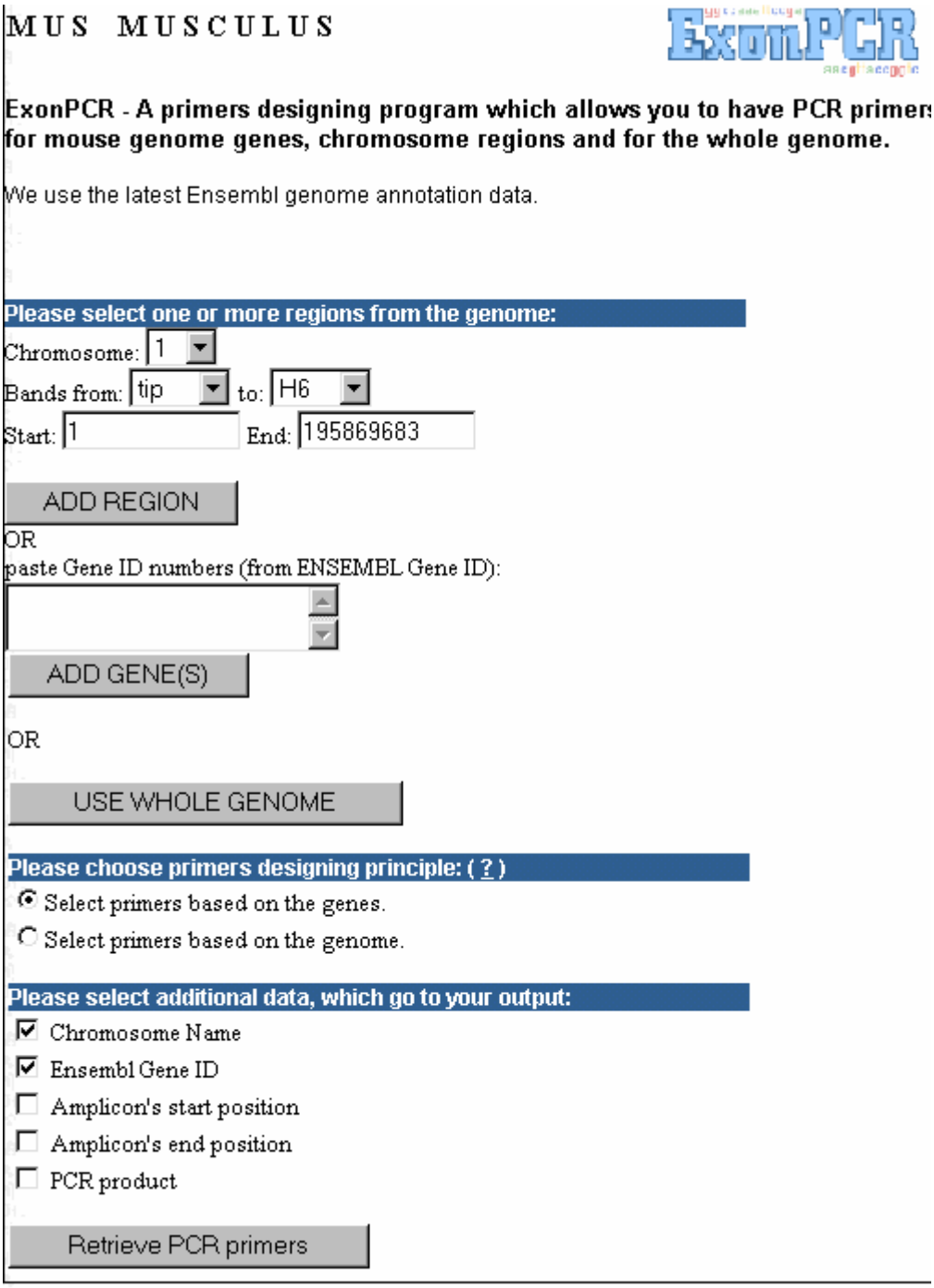
Tabel 8. 3 tabelit hoiavad andmeid amplikonidele leitud praimeripaaride kohta, mis on leitud n.n. “üle genoomi” põhimõtte järgi (laiendiga _chr), teised 3 selliste amplikonide kohta, mille eksonid on pärit ühest geenist (laiendiga _g).

Tables_in_exonPCR_mouse_3
Amplicons_chr
Amplicons_g
exons_chr
exons_g
primers_chr
primers_g

Amplikonide kaks tabelit talletavad endas infot amplikonide ID, amplikoni algus- ja lõppkoordinaatide ning kromosoomi nime kohta, kuhu vastav amplikon kuulub. Eksonite tabelite struktuur kätkeb endas kromosoomi nime, amplikoni ID–d ja sinna kuuluvate vastavate eksonite ID–sid. Ühes amplikonis ei saa olla erinevate ID–dega eksoneid, küll aga võib üks ekson, mille pikkus on üle 3000 bp, olla “hakitud” ja kuuluda korraga mitmesse amplikoni. Järelikult eksoni ID veerus ei leidu kahte rida, mis oleksid teineteisega identsed. Tabel primers_chr sisaldab endas kromosoomi nime, amplikoni ID-d, selle mõlemaid praimereid ja PCR–i produkti. Tabelite struktuurid on mõlemal põhimõttel peaaegu samad, ainult et tabelid laiendiga “_g” sisaldavad endas veel geeni ID-d (gid) garanteerimaks, et ühe amplikoni eksonid oleksid kindlasti pärit ühest geenist.

2.3 Kasutajaliides

Kasutajaliidese loomisel lähtusime eelkõige sellest, et on vajalik kasutaja poolt sisestatud andmete dünaamiline ja automaatne liikumine arvutiprogrammile ja programmi töö tulemusena genereeritud andmete visualiseerimine kasutajale. Programmide kodeerimisel oleme kasutanud programmeerimiskeelt Perl. Dünaamilise webilehe ehitamise vahendiks valisimegi CGI (*Common Gateway Interface*) tänu tema heale ühilduvusele Perl-iga. Samuti on CGI võimas vahend andmebaaside haldamiskeele MySQL-iga suhtlemiseks ja HTML formaadis teksti genereerimiseks. Näiteks saab tänu CGI võimalustele kasutaja piirata/laiendada väljundfailis genereeritavate andmete hulka jms. (http://hoohoo.ncsa.uiuc.edu/cgi/intro.html).



Joonis H. Graafiline kasutajaliides (http://www.biodata.ee/exonPCR_mouse/). Kasutajal on võimalik praimereid saada kolme erinevat võimalust kasutades: valida vastav kromosoom või selle piirkond, sisestada ise ENSEMBL geeni ID või saada tulemuseks kogu hiire genoomi piirkondade üles amplifitseerimiseks vajalikud praimerid.

Arutelu

Paraku on raske anda hinnangut, garantiist rääkimata millegile, mida praktikas pole kasutatud. Alati võib juhtuda nii, et teooria ja praktika sammuvad eri teid pidi. Seega, tulemusi saame üksnes analüüsida, kas need ka paika peavad, see on teine küsimus. Mis on aga kindel, on see, et tegemist on töökindla vahendiga, mis aitab meil tulemusteni jõuda kiirelt ja mugavalt.

Võttes veelkord kokku eelmist peatükki saame öelda, et üle genoomi suutsime leida praimerid peaaegu kõikide funktsionaalsete piirkondade üles amplifitseerimiseks. Amplikone, millele ei leitud praimereid jäi järgi 0.7% ja geene, mida ei ole võimalik üles amplifitseerida on kusagil 2%. Et neid arve maksimaalselt vähendada, peame me tehtud tööga läbi viima täiendavaid analüüse, manipuleerides erinevate võimalustega. Põhjus-tagajärg (PCR–i halvad tulemused) seoseid võib hakata otsima juba algandmetest, et kas andmebaas ENSEMBL on ikkagi piisavalt usaldusväärne. Teadlased võiksid anda rohkem tagasisidet (teatud tagasiside on), et kas ENSEMBL–i andmetega on saavutatud soovitud tulemusi. Ja samamoodi võiks olemas olla võrdlus teiste andmebaaside andmetega.

Võibolla oleks õigem meie poolt kasutatavate programmide asemel kasutada alternatiive. Kordusjärjestuste maskeerimiseks võiks kasutada programmi REPEATMASKER, kuid ka sellel on omad puudused, nagu aeglus ja see, et maskeerib korduses oleva ala täies pikkuses, hoolimata sellest, et see piirkond võis siiski sisaldada “häid” praimeri disaini regioone.

Üheks võimaluseks oleks kohe praimerite disainimise algetapis muuta GM_PRIMER3 sisendfailis PCR-i parameetrite lubatud varieeruvuse vahemikke. Näiteks vähendada GC-sisaldust vahemikku 20-50% ja piirata produktide pikkusi. Kuna praktikas on leidnud tõestamist see, et praimeri pikkus ei ole otseses korrelatsioonis PCR väljatulemise kvaliteediga (Haas *et al*., 1998), siis need väärtused jätaks esialgu puutumata. Pigem laseks lõdvemaks ja katsetaks, mis tulemused siis saame.

Kutsumuseks omaette oleks loomulikult just nende amplikonide edasi analüüsimine, millele jäid praimerid disainimata. Analüüsida nende asukohta genoomis, sest võimalik on, et nende ümber ei olnud meie poolt nõutavat 100 nukleotiidilist külgnevat ala või oli see ala liialt korduv. Kui produkti ümbert järjestus

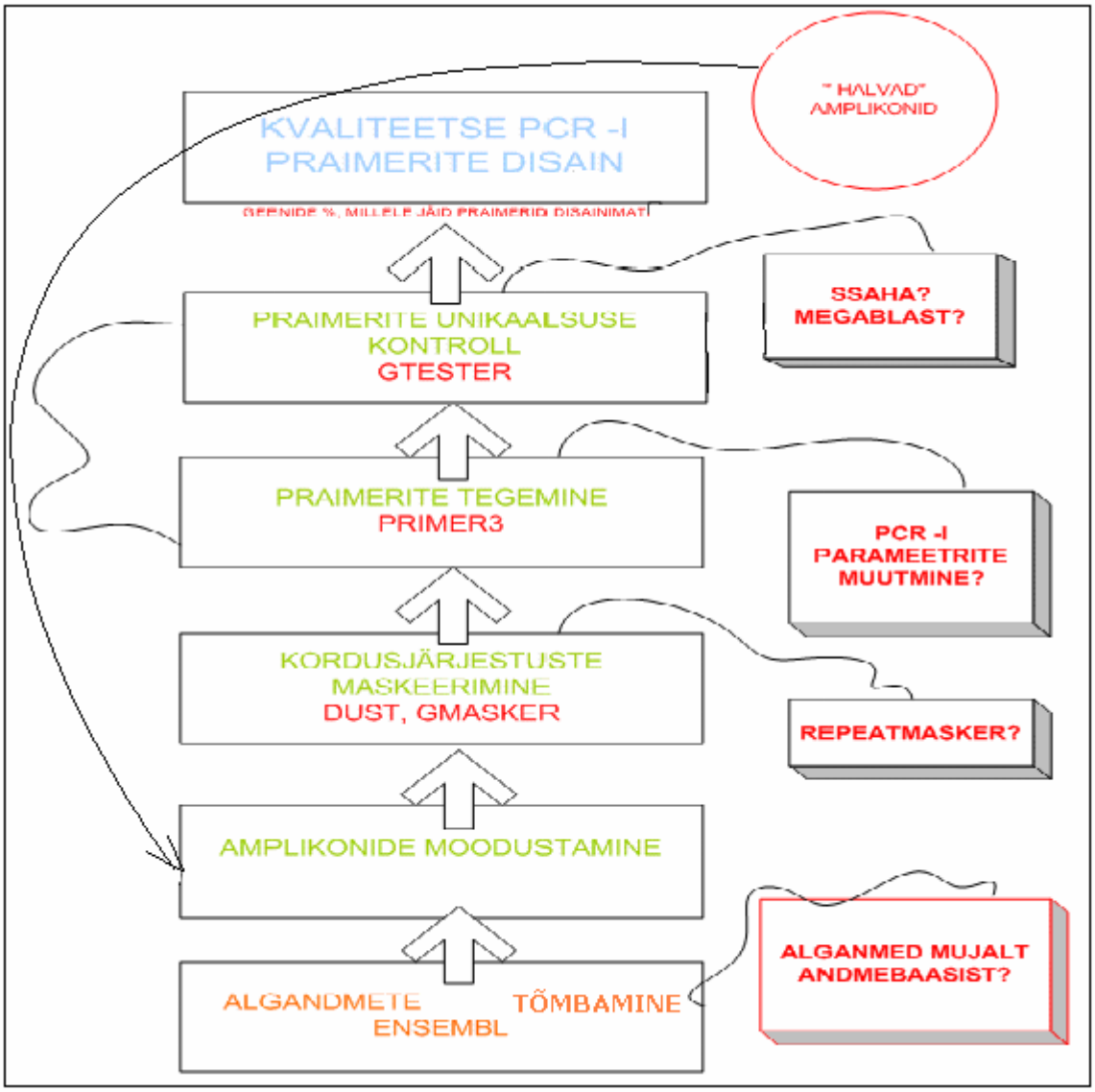
puudub, s.t. on katkendlik (ekson on katki ja teine pool eksonist sekveneerimata), siis meie poolt kasutatud põhimõtte järgi praimereid teha ei saa. Tegelikult on see võimalik, sest praimeri 5`ots võib jääda ka täpselt eksoni piiri peale. Kui järjestust on vähe, siis saab sinna "suruda" halva kvaliteediga praimer, mis oleks võibolla siiski parem kui mitte midagi. Kui külgnev ala on korduses, on raskem lahendust leida. Võibolla saaks rakendada 2-astmelist *nested* PCR-i, kus esmalt amplifitseeritakse pikem unikaalne lõik ja selle produkti pealt lühem lõik. Teisisõnu, nende amplikonide asemele, millele praimereid ei leitud, tuleks konstrueerida uued, teatud teise n.n "*flanking* piirkonna" põhimõtte alusel.

Mida kindlasti ei tohiks teha, on teaduslikult tõestatud parameeetrite, mis PCR–i tulemuste kvaliteeti kindlasti mõjutavad, liiga lai varieerimine, kuna see ei tagaks meile kindlasti tegelikku unikaalset seostumiskohta omavate praimerite disaini.

Kõiki asjaolusi arvestades, julgen ma öelda, et 2% geene, millele ei suudetud praimereid disainida ei ole nii suur hulk ja on esimese korra kohta küllaltki hea tulemus. Kõige esimeseks sammuks ei oleks kindlasti õige aega raisata töö meetodite muutmise, kui just nende tulemuste testimise peale, et kui head praimerid suudab see programm ikka leida ja alles siis otsustada, mida edasi teha.

Kahtlemata ei tohi ära unustada graafilise kasutajaliidese ja seega ka andmebaasi täiustamist, sest praegu ei saa kasutaja oma järjestuse kohta, millele praimereid soovib disainida, eriti lisa informatsiooni. Kasutaja praimerite otsing peaks saama olema soovi korral veel piiritletum.

Järgnev joonis iseloomustab etapiliselt töö käiku ja võimalikke kitsaskohti.



Joonis I. Näidatud on ära, kuidas läbi edutegurite saavutada meie töö eesmärk ja millised on võimalikud kitsaskohad. Joonisega käib kaasas ütlemine, et ahela tugevuse määrab ära tema kõige nõrgem lüli.

Kokkuvõte

Käesolevas töös disainiti hiire genoomi geenide üles amplifitseerimiseks vajalikud praimerid, kusjuures kogu töö käigus prooviti järgida optimaalsuse kriteeriume, sellest ka ehk hea töö tulemuslikkus.

Esimene etapp oli algandmete tõmbamine, kus andmebaasist ENSEMBL saime endale kõik hiire genoomi geenide asukohad, sealt edasi ka vastavad järjestused. Kuna PCR–i abil saab üles amplifitseerida kuni 3000 nukleotiidiseid juppe, siis moodustasimegi maksimaalselt 3000 bp amplikonid ehk PCR-i produktid (töös on kasutatud veel blokkide nime). Kogu töökäik käis kahte sarnast teed pidi, mis erinesid üksteisest ainult algselt, amplikonide moodustamise põhimõtte poolest. Ühed amplikonid moodustati selliselt, et nad kindlasti koosneks ühe geeni eksonitest. Edasi kirjutati programmid, mis hakkasid vastavaid teid pidi moodustatud amplikonidele praimereid disainima. Saadigi praimerid, millest osad aga pärast genoomis seostumiskohtade unikaalsuse kontrollimist jäid faili bad_primers, s.t. faili, kuhu suunati praimerid, mis omasid rohkem kui ühte seostumiskohta genoomis. Neile disainiti samade põhimõtete järgi uued praimerid, mitte aga üks, vaid 100 paari. Sel moel saadigi tulemuseks 99.3%, mis kajastab endas amplikonide osa, millele suudeti praimerid kontrueerida.

Töö muutmiseks täiuslikuks, viisime andmed (nii praimerid kui produktid ehk amplikonid) andmebaasi, kust edasi koostasime CGI ja MySql–i koostööna dünaamilise kasutajaliidesega webilehe praimerite leidmiseks kõikidele hiire genoomi geenidele.

Summary

In the present research, the primers necessary for amplifying the genes of the genome of a mouse were designed, trying to follow the criteria of optimum throughout the working process, which can most probably be seen in the good results of this work.

The first stage consisted of data mining where we downloaded all the sites of the genes of a mouse's genome from the ENSEMBL database. After that we got the corresponding alignments. As using PCR, it is possible to amplify 3000 nucleotide bits. Therefore, we formed maximum 3000 bp amplicons or PCR products (in this research the name of blocks is still used). The working process progressed in two similar courses. They differed from each other only initially in the principle of forming the amplicons. One part of these amplicons was formed in such a manner that they would definitely consist of the exons of one gene. Further on programs were written that started to design primers for the amplicons through corresponding manners. Primers were obtained. One part of primers remained in the file of bad primers after controlling the uniqueness of the connection points in the genome. It means that they remained in the file into which those primers were directed that had more than one connection point in the genome. In order to find a pair of primers for most of the amplicons, new primers were designed for them following the same principles but not only one pair but 100 pairs of them. In that way the result of 99,3% was obtained that expresses this part of the amplicons for which primers could be constructed.

In order to perfect the present research, we inserted the data (both the primers and the products or the amplicons) in a database and in cooperation between CGI and MySql composed a web page with a dynamic user interface to find primers for all the genes in a mouse's genome.

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. Bradley A.. Mining the mouse genome. Nature. 420(6915):512-4, 2002.
3. Borer P.N., Dengler B., Tinoco I.Jr. Stability of Ribonucleic acid Double-stranded helices. J.Mol. Biol. 86:843, 1974.
4. Breslauer J.K., Dengler R., Blocker H., Marky L. Predicting DNA duplex stability from the base sequence. Proc.Natl.Acad.Sci. 83:3746-3750, 1986. 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. Dermitzakis Emmanoull T., Alexandre Reymond, Robert Lyle, Nathalie Scamuffa, Catherine Ucla, Samuel Deutsch, Brian J. Stevenson, et. al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature. 420(6915):578-82, 2002.
6. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science. 302(5647):1033-5, 2003.
7. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP et al. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. Proc Natl Acad Sci U S A. 100(3):1140-5, 2003.
8. Haas s., Vingron M., Poustka A., Wiemann S. Primer design for large scale sequencing. Nucleic Acid Res 26(12):3006-3012, 1998.
9. 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.
10. Kämpke T., Kieninger M., Mecklenburg M. Efficient primer design algorithms. Bioinformatics 17(3):214-25, 2001.

11. 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 Automates Large-Scale DNA Sequencing. Genomics 40(3):476-85, 1997.
12. Mouse Genome Sequencing Consortium: Initial sequencing and comparative analyses of the mouse genome. Nature 420(6915):520-62, 2002.
13. Ning Z., Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res 11(10):1725-9, 2001.
14. Pennacchio Len A.. Insights from human/mouse genome comparisons. Mammalian Genome.14(7):429-36, 2003.
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. Reymond A, Marigo V, Yaylaoglu MB, Leoni A, Ucla C, Scamuffa N et al. Human chromosome 21 gene expression atlas in the mouse.Nature. 420(6915):582-6, 2002.
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. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol 7(1-2):203-14, 2000.
20. Tatusova T.A., Madden T.L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbio Lett 174(2):247-50, 1999.
21. 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.