TARTU ÜLIKOOL
LOODUS- JA TEHNOLOOGIA TEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Viktorija Kukuškina

Mereteo *Conus consors* geenide arvu ja struktuuri ennustamine bioinformaatiliste meetoditega.

Magistritöö

Juhendaja
Lauris Kaplinski

Tartu 2013

Sisukord

Kasutatud lühendid

5'/3' UTR	*5′/3′ untranslated region* – 5'/3' mittetransleeritav regioon
ADRC	ADP-ribosüül tsüklaas
ass	*acceptor splice site* - aktseptor splaisingu sait
ATP	adenosiintrifosfaat
BLAST	*Basic Local Alignment Search Tool* (programmi nimetus)
BLAT	*Blast Like Alignment Tool* (programmi nimetus)
bp	*basepare* – aluspaar
cDNA	*complementary DNA*- komplementaarne DNA
CDS	*coding DNA sequence* – kodeeriv DNA järjestus
DNA	*deoxyribonucleic acid* – Desoksüribonukleiinhape
dNTP	desoksüribonukleotiidtrifosfaat
dss	*donor splice site* - doonor splaisingu sait
ELH	*egg-laying hormone* – munemist reguleeriv hormoon
EST	*expressed sequence tag*- ekspresseeritud järjestusmärgis
gb	*GenBank* (formaat)
Gb	*gigabase* – miljard aluspaari
gff	*general feature format* (formaat)
GHMM	*Generalized Hidden Markov Models* – generaliseeritud varjatud markovi mudelid
GMAP	*Genomic Mapping and Alignment Program* (programmi nimetus)
HMM	*Hidden Markov Models* – varjatud markovi mudelid
Kb	*kilobase* – tuhat aluspaari
Mb	*megabase* – miljon aluspaari
mm	millimeeter
mRNA	*messenger RNA* – informatsiooni-RNA
mtDNA	mitokondriaalne DNA
PASA	*Program to Assemble Spliced Alignments* (programmi nimetus)
PE	pair-end - paaris ots, kasutatakse Illumina lugemite jaoks
PERL	*Practical Extraction and Reporting Language* (programmeerimis keel)

pg	pikogramm
psl	*Property Specification Language* (formaat)
RNA	*ribonucleic acid* – ribonukleiin hape
SE	*single-end/read* - üksik ots/lugem, kasutatakse Illumina lugemite jaoks
SHARP	*Short Read Assembly Protocol*- lühikeste lugemite kokkupanemise protokoll
SRS	*Short Read Sequencing* – lühikeste lugemite sekveneerimine
WWAM	*Windowed Weight Array Model*
YGAP	*Yeast Genome Annotation Pipeline* (programmi nimetus)

Sissejuhatus

Teise põlvkonna sekveneerimistehnoloogiate kasutusele võtmisega kaasnes uute liikide genoomide sekveneerimise oluline lihtsustumine ja andmehulkade kiire kasv. Kuna ühe genoomi sekveneerimise hind langes mitmeid suurusjärke, muutus sekveneerimine paljudele laboritele kättesaadavaks, ning aina kiiremini hakati sekveneerima nii varem uuritud kui ka uusi seni uurimata liike. Senikasutatud Sangeri tehnoloogia oli täpsem, aga ka oluliselt kallim. Seetõttu olid varasemad genoomsed järjestused enamasti täpsemad, neid oli vähem ja iga genoomi kohta oli proportsionaalselt rohkem valgulist ja transkriptoomset informatsiooni. See võimaldas esimesi genoome käsitsi annoteerida. Praeguste andmehulkade juures pole käsitsi annoteerimine enam mõeldav, sest selleks tuleks kaasata liiga palju inimesi. Seetõttu tekkis vajadus protsessi automatiseerimiseks ja lihtsustamiseks, nii hakati looma programme, mis aitaksid teadlastel annoteerida geene uute sekveneeritud liikide madala kvaliteediga genoome. Need programmid baseeruvad põhiliselt erinevatel statistilistel mudelitel, eelkõige Markovi mudelil või selle teisendustel, nagu HMM või GHMM. Paljud geniannotatsiooniprogrammid on spetsialiseeritud mingi konkreetse eesmärgi, organismi või organismide rühma jaoks.

Hiljuti sekveneeriti kalatoidulise mereteo *Conus consors* genoom. Antud organismi uurimise motivaatoriks oli farmakoloogiline huvi. Nimelt kasutab *C. Consors* oma saakloomade paralüseerimiseks peamiselt neuroaktiivsetest konopeptiididest koosnevat mürki. Teise põlvkonna meetoditega sekveneeritud genoomidele tüüpiliselt on sekveneeritud järjestus madala kvaliteediga ning koosneb suurest hulgast lühikestest fragmentidest. Samas on aga molluskite hõimkonna esindajate genoome sekveneeritud vähe ja nende geenide struktuurist ei ole palju teada. Seetõttu tekkis huvi analüüsida, kui edukalt on võimalik kasutada automaatseid geeniennustusalgoritme selleks, et analüüsida uudse ja vähe uuritud organismi madala kvaliteediga sekveneeritud genoomi, sealhulgas kirjeldada tema geenide struktuuri ja hinnata geenide koguarvu.

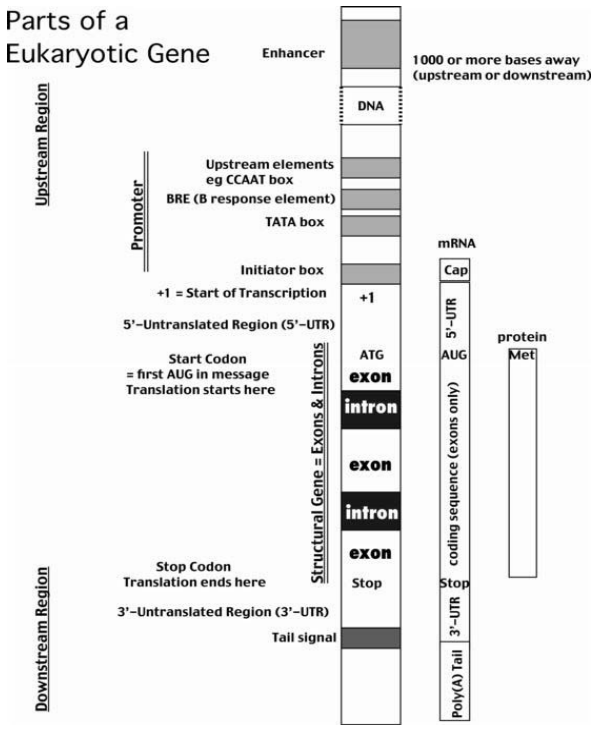
Suur tänu kõikidele töökaaslastele, eriti Laurisele positiivse suhtumise ja innustava juhendamise eest.

Eriline tänu abikaasale igakülgse abi ja toetuse eest.

Kirjanduse ülevaade

I. Eukarüootide geenistruktuur

Eukarüootide geenistruktuur on palju keerulisem, kui prokarüootide oma. Tüüpilises eukarüoodis ei ole valku kodeeriv DNA regioon tavaliselt pidev vaid koosneb vaheldumisi paiknevatest eksonitest ja intronitest. Transkriptsiooni käigus transkribeeritakse nii eksonid kui intronid nende lineaarses järjekorras RNAks. Järgneb protsess, mida nimetatakse splaisinguks (*splicing*) ja mille käigus RNAsse kuuluvad intronid lõigatakse sellest välja. Eksonid liidetakse omavahel formeerides nii küpse mRNA. Tüüpiline multieksonilise geeni struktuur on toodud joonisel 1. Geeni promootoralast 1000 või rohkem aluspaari ülesvoolu (5' suunas) võib asuda võimendav *(enhancer)* järjestus, mis soodustab polümeraasi seondumist promootorpiirkonda. Geen algab promootorpiirkonnaga, millele järgneb lähte-ekson *(initial exon)*. See algab transkribeeritava, kuid valku mittetransleeritava alaga – 5' *untranslated region (5'UTR)* millele järgneb ATG startkoodon. Lähte-eksonile järgneb seeria vaheldumisi paiknevaid introneid ja sise-eksoneid (*internal exons*), millele järgneb termineeriv ekson (*terminal exon*), mis sisaldab *stop*-koodonit. Sellele omakorda järgneb teine mittetransleeritav ala, mida nimetatakse 3' UTR. Eukarüootse mRNA kõige lõpus asub polüadenülatsiooni signaal. See nimetus tuleneb sellest, et sinna liideakse *polyA* saba (*polyA tail*), mis koosneb järjestukustest adeniinidest. Eksonite ja intronite piiri määravad splaisingsaidid (*splice sites*), millel on spetsiifiline 2-4 aluspaari *(base pairs - bp)* pikkune järjestus. 5' introni ja 3' eksoni otsa nimetatakse doonor-saidiks *(donor site)* ja 3' introni 5' eksoni otsa nimetatakse aktseptor-saidiks *(acceptor site*) (http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec07/node8.html).



Joonis 1. Eukarüootse geeni stuktuur. 1000 bp ülesvoolu asub võimendaja *(enhancer)*, geen algab promootor piirkonnaga, millele järgneb lähte-exon (*initial exon*) koos mittetransleeritava alaga – 5'*UTR* ja start koodoniga, ning sellele vaheldumisi järgnevad intronid ja sise-eksonid (*internal exons*), viimasena jääb termineeriv ekson, mis sisaldab stop koodonit. Selle järel on veel 3'*UTR* ning mRNA puhul veel *polyA* saba. [http://www.micro.siu.edu/micr302/Gene.html]

Geenide tuvastamine eukarüootide genoomist on nende geenistruktuuri varieeruvuse tõttu väga keeruline. Näiteks on keskmine selgroogse geen 30 tuhat aluspaari (*Kb*) pikk, millest kodeeriv järjestus moodustab ainult 1Kb. Kodeeriv regioon sisaldab endas keskmiselt umbes 6 eksonit, igaüks umbes 150 bp pikk. Samas kõrvalekalded keskmisest on suured ja neid on palju. Näiteks võib tuua düstrofiini geeni, mis on 2.4 miljonit aluspaari (*Mb*) pikk. Vere koagulatsioonifaktoril VIII, on 26 eksonit, selle eksonite pikkused varieeruvad alates 69 bp kuni 3106 bp (http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec07/node8.html).

Kõige paremini läbi uuritud eukarüootse geenistruktuuri näiteks on inimese (*Homo sapiens*) genoom ja geenid. Genoomi pikkuseks on 3.2 miljardit aluspaari (*Gb*) ja see jaguneb 23 kromosoomi vahel. Kõige väiksema kromosoomi (21) pikkus on 50 Mb ja suurima kromosoomi (1) pikkus on 263 Mb. Enamik genoomist, ligi 50%, koosneb mittekodeerivatest unikaalsetest järjestustest ja kordusjärjestustest. Valku kodeerib umbes 2% genoomist. Ülejäänud 48% on niinimetatud unikaalne DNA, millest enamik arvatavasti koosneb divergeerunutest mobiilsetest elementidest. Täpse geenide hulga kohta puudub ühene konsensus, hinnangud kõiguvad tavaliselt 20000 kuni 30000 geeni haploidse genoomi kohta. Enamikus inimese geenides katkestavad kodeerivat järjestust mittekodeerivad alad – eksonid on vaheldumisi intronitega. Nagu ka teistes eukarüootides jagunevad eksonid 5'UTR eksoniteks, 3'UTR eksoniteks, kodeerivateks eksoniteks ja nende kolme kombinatsioonideks, sealhulgas üksikeksoniteks mis katavad terve mRNA. Inimese eksonid on suhteliselt lühikesed, mediaanpikkusega 167 bp ja keskmise pikkusega 216 bp. Lühim ekson on 12 bp ja pikim 6609 bp. Mediaan ja keskmised pikkused erinevate eksonite rühmade jaoks on: 5'UTR eksonid 118 ja 191 nukleotiidi, CDS (*coding DNA sequence*) eksonid 1191 ja 1424 nukleotiidi ning 3'UTR eksonid 534 ja 576 nukleotiidi pikkad. Geene transkribeeritakse ühelt või teiselt ahelalt, mõned geenid võivad asuda ka teiste geenide intronites. Umbes 100 geenipaari on 3' otsas ülekattes, ehk siis nad kasutavad sama 3'UTR piirkonda, ehkki erinevatel ahelatel (Makalowski, W. 2001).

II. Molluskite (limuste) genoomid

Molluskid on üks mitmekesisemaid loomarühmi. Nende liigid varieeruvad morfoloogiliselt ja suuruse poolest, mikroskoopilistest kuni 1 m suurusteni ja asustavad nii mere, magevee kui maismaa elupaiku. Molluskite hulgas on olulisi vesiviljelusliike, keskkonnavalvureid, kahjureid ja haiguste vektoreid. Hõimkond *Mollusca* sisaldab endas *ca* 100000 kirjeldatud kaasajal elavat liiki. Molluskite hulka kuulub 6 või 7 liini, mis põhiosas kattuvad traditsiooniliste klassidega. Kolm suurimat neist on *Bivalvia* (karbid), *Gastropoda* (teod) ja *Cephalopoda* (peajalgsed) (Takeuchi, T., jt 2012). Molluskite genoomi pikkus varieerub umbes 0,4 Gb (*Lottia gigantea*) kuni 5,9 Gb *(Neobiccium eatoni)*. Mitokondriaalse DNA pikkus varieerub 10 kuni 42 kb ja see on suhteliselt AT rikas (Simpson, W.B., Boore, J.L., 2010). Võrdluseks võib öelda, et imetajate mtDNA suurus on umbes 16 kb (Ferris, S. D., jt 1983) ning äädikakärbse mtDNA suurus on umbes 14.9 kb (Montooth, K. L., jt 2009).

Animal genome size andmebaasis on olemas andmed 263 molluskiliigi genoomi kohta. Käesoleva töö kirjutamise ajaks oli täisgenoom sekveneeritud järgmistel molluskiliikidel: *Conus Consors* (meie uuritav liik - tigu), *Crassostera gigas* (karp), *Pinctada fucata* (karp), *Lottia gigantea* (tigu), *Aplysia californica* (tigu) ja *Conus bullatus* (tigu).

1. Karbid (*Bivalvia*)

Bivalvia klassi kuulub *ca* 20000 kaasajal elavat liiki. Karpe iseloomustab see, et nende koda koosneb kahest poolmest. Poolmed tekkisid evolutsiooni käigus nii, et eellasliigi seljal kantav paaritu koda jagunes kaheks pooleks (Takeuchi, T., jt 2012). *Animal genome size* andmebaasi järgi varieerub karpide genoomide suurus 1200 kuni 2100 Mb.

Joonis 2. *Pinctada fucata,* autor Didier Descouens.

Praeguseks on sekveneeritud kahe karbi genoomid. Aastal 2012 sekveneeriti pärlikarp *Pinctada fucata* (Joonis 2.) (Takeuchi, T., jt 2012). Pärlikarbi genoomi uurimiseks oli kaks peamist põhjust. Esiteks on pärlikarp majanduslikult oluline ja laialdaselt kultiveeritav organism. Tema bioloogiliste protsesside parem mõistmine on oluline pärlitööstuse jaoks ja võib avada uusi võimalusi seda täiustada. Teiseks sooviti saada täpsemaid andmeid karpide ja laiemalt molluskite bioloogiast, eriti nende evolutsioonilisest põlvnemisest *lophotrochozoa* ülemhõimkonnas. Pärlikarbi

ca 1150 Mb genoom sekveneeriti 40-kordse katvusega Roche 454 GS-FLX ja Illumina GAIIx tehnoloogia abil. Genoomi suuruse tuvastamiseks kasutati haploidset spermi tuuma. Võrreldes teiste karpide genoomidega on *P. fucata* genoom väike. Tal on 28 kromosoomi ja GC sisaldus ei ületa 34%, seega tal on AT rikas genoom. Transposoonid, retrotransposoonid ja tandeemsed kordusjärjestused moodustavad 9.8% kogu genoomist. *Ab initio* geeniennustusprogrammi AUGUSTUS abil leiti 43760 geenimudelit, nendest 23257 täielikku (start ja stop koodonitega). Nendest 23257 geenimudelist 70% jaoks leidusid ka ekspresseeritud järjestusmärgised *(EST, expressed sequence tags)*. *P.fucata* geenide keskmine pikkus on 6700 bp ja eksonite hulk geeni kohta 3,2. Keskmine eksoni pikkus on 589 bp, intronilist järjestust geeni kohta on kokku 4815 bp ja valgud keskmiselt koosnevad 274 aminohappest (Takeuchi, T., jt 2012).

Teine karp, mille täisgenoom on sekveneeritud on Vaikse Ookeani auster *Crassostera gigas* (Joonis 3.). See liik pakub huvi arengubioloogidele, kuna tema areng toimub mosaiikselt, tüüpiliste molluskite arengustaadiumitega. Lisaks pakkus huvi tema kohastumine mereloomade jaoks ekstreemsete keskkonnatingimustega. *C. gigas* elab tõusu-mõõna vööndis, ta suudab taluda mitu ööpäeva kestvat kuivale jäämist ja kestvat ülekuumenemist päikese käes. Tema 559 Mb suurune genoom sekveneeriti 155-kordse katvusega. Kombineeritud meetoditega leiti sealt kokku 28027 geeni (Zhang, G., jt 2012).

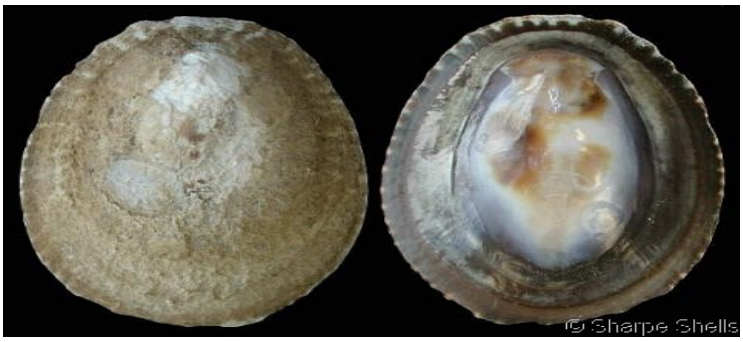
Joonis 3. *Crossastera gigas,* Thunberg, 1793

2. Teod *(Gastropoda)*

Klass *Gastropoda* sisaldab endas 60000 kuni 80000 kaasajal elavat teoliiki. Tigude globaalne levila on väga lai, ulatudes Antarktika ja Arktika servadest troopikani. Nad on hästi kohastunud mitmesugusteks elutingimusteks, ning asustavad kõiki loodusvööndeid. Kuigi maismaa- ja mageveeteod on paremini uuritud, elab kaks kolmandikku kirjeldatud teoliikidest meres. Nad asustavad lisaks ookeanide ja merede rannavöötmele ka süva- ja avamerd (Zenkevitš, L. 1969). Ehkki teod on olulise tähtsusega ning huvitava bioloogiaga, on nende genoome vähe uuritud.

2007 aastal sekveneeriti meretigu *Lottia gigantea* (Joonis 4.). See liik valiti sekveneerimiseks, kuna ta on populaarsust koguv evolutsiooni ja arengu ning lisaks ka ökoloogia ja keskkonnakaitse

mudelorganism. Võrreldes teiste molluskitega on *L. gigantea* genoom väike, koosnedes 359.5 Mb. Genoom sekveneeriti 8.87-kordse katvusega, saades 20,146 kontiigi (lühikest assembleeritud järjestust) ja 4,475 skaffoldi (kontiigidest kokku pandud pikemat järjestust). *L. gigantea* keskmine geeni pikkus on 5234 bp, transkripti pikkus 1287 bp, keskmine eksoni pikkus 213 bp ja introni pikkus 787 bp. Kokku tuvastati 23800 geenimudelit milles oli keskmiselt 6 eksonit geeni kohta (Simakov, O., jt 2012).



Joonis 4. *Lottia gigantea,* autor Sharpe Shells.

Merijänes *Aplysia californica* (Joonis 5.) on levinud katseloom neurobioloogilistes uuringutes. Tema genoomi suurus on 1.8 Gb ja see sekveneeriti 8-kordse katvusega.

Joonis 5. *Aplysia californica,* Berkeley Science Review, issue 12

Meretigu *Conus bullatus* koonuskodalaste sugukonnast (Joonis 6.) on röövloom, kelle peamiseks toiduks on kalad. Ta kuulub koos *C.cervus*, *C.dusaveli* ja *C.consors*'iga koonustigude *Textilia* klaadi. *C.bullatus* elutseb India ja Vaikses ookeanis Havai saartest kuni Lõuna-Aafrikani. Tigu sekveneeriti osaliselt konopeptiidide uurimiseks. Tema genoomi suurus on *ca* 2.56 Gb ja sekveneerimise katvus oli 3-kordne. Genoomi GC sisalduseks hinnati 42.88%. Väga suure osa genoomist moodustavad kordusjärjestused – samas väga kõrge koopiarvuga korduste osa genoomist oli võrreldes inimesega väiksem. Kokku leiti *C. Bullatus*'e genoomist 2410 oletatavat konopeptiidide kontiigi (Hu, H., jt 2011).

Joonis 6. *Conus bullatus*, author Joël Orempuller.

III. Molluskite geenide struktuur

Molluskite genoome on väga vähe uuritud ja nende geenide struktuuri kohta ei ole palju informatsiooni. Põhiliselt on eraldi analüüsitud üksikud huvipakkuvaid geene. Näiteks on põhjalikult uuritud gastropoodi *Biomphalaria glabrata* globiini geen. Geenis paiknevad kolm intronit, mis on 1116, 1008, ja 582 bp pikad. Globiini initsiatsiooni koodoni ees on mittetäielik 45bp UTR. Puudub liiderjärjestus, avatud lugemisraam ulatub 148 koodonini, ning sellele järgneb erakordselt pikk 970 bp UTR. On olemas normaalne polüadenülatsiooni sait (Dewilde, S., jt 1998). *Cephalopoda* liinist *Octopus vulgaris'*e tsefalototsiini *(cephalotocin)* ja oktopressiini *(octopressin)* geenid ei sisalda valgu kodeerivas regioonis üldse introneid ja koosnevad ühest eksonist, samas kui enamiku selgroogsete ja ka mudakuke (*Lymnea stagnalis)* Lys-oktopressiin sisaldab 3 intronit ja 2 eksonit (Kanda,A., jt 2003). *Aplysia californica* ELH (*egg-laying hormone*) geen koosneb kolmest eksonist, mis on eraldatud üksiku vahepealse järjestusega. TATA ja CAAT boksidega homoloogsed järjestused asuvad esimesest eksonist ülesvoolu. Geeni 3' otsas asub polyA sait (Mahon, A. C., jt 1985). *Aplysia kurodai* ADRC *(ADP-ribosyl cyclase)* geen on umbes 7 kb pikk ja koosneb kaheksast eksonist ja seitsmest intronist (Nata, K., jt. 1995). Nagu ülalpool juba mainitud, sisaldavad *Pinctada fucata* geenid keskmiselt 3,2 eksonit keskmise pikkusega 589 bp. *Lottia gigantea* keskmine eksoni pikkus on 213 bp ja keskmine introni pikkus 787 bp.

IV. Imetaja (inimene) ja molluskite võrdlus genoomi tasemel

Nii imetajad kui molluskid kuuluvad loomade riiki. Imetajate liike on palju uuritud ja nendest on olemas mitmed kõrge kvaliteediga genoomijärjestused, samas kui molluskite genoome on uuritud vähe. Eriti vähe on teada meres elavate molluskite liikide kohta. Teise põlvkonna sekveneerimistehnoloogia tulekuga on nüüd ilmunud esimesed sekveneeritud molluskite genoomid.

Molluskite genoomi suurus varieerub palju rohkem, kui imetajate oma, mis on ka loomulik arvestades seda, et molluskid moodustavad terve hõimkonna, samas kui imetajad on suhteliselt homogeenne klass keelikloomade hõimkonnast. Imetajate genoomi suurus varieerub 1.73 pg (pikogrammi) *Miniopterus schreibersi*'l kuni 8.40 pg *Tympanoctomys barrerae*'l (http://www.genomesize.com/statistics.php?stats=mammals), molluskite genoomi suurus vastavalt 0.43 pg *Lottia gigantea* kuni 7.85 pg *Diplommatina kiiensis kiiensis* (http://www.genomesize.com/statistics.php?stats=molluscs). Paljudel molluskitel on leitud polüploidsust (Gregory, T. R., 2005), samas kui imetajatel on teada ainult üks juhtum (Gallardo, M.H., jt 2002). On teada, et gastropoodide veeliikidel on genoom väiksem, kui maismaa liikidel (Vinogradov, A.E. 2000). Imetajate keskmine genoomi suurus 3.37 pg ± 0.04 (http://www.genomesize.com/statistics.php?stats=mammals) on suurem molluskite keskmisest genoomi suurusest 2.10 pg ± 0.08 (http://www.genomesize.com/statistics.php?stats=molluscs).

Võrreldes kahte konkreetset liiki - inimest (*Homo sapiens)* ja pärlikarpi (*Pinctada fucata),* võime välja tuua järgmised erinevused: Inimese genoom on 3.2 Gb ja koosneb 46 kromosoomist (http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml#genomesize), pärlikarbi genoom on väiksem, 1150 Mb ja sisaldab 28 kromosoomi (Takeuchi, T., jt 2012). Inimese genoom on GC-rikkam, tema keskmine GC sisaldus on 41% (International Human Genome Sequencing Consortium 2001), pärlikarbil on see 34% (Takeuchi, T., jt 2012). Kuigi inimese genoomi on väga põhjalikult uuritud, ei ole tema täpne geenide hulk teada. Encode projekt hindab inimese geenide arvuks on 20697 valku kodeerivat geeni (Pennisi, E. 2012). Pärlikarbi genoomist leiti 23257 täielikku geenimudelit (Takeuchi, T., jt 2012). Pärlikarbiarbi keskmine eksonite arv geeni kohta on kirjanduse andmetel 3.2 , keskmine eksoni pikkus on 589 bp, valkude keskmine pikkus 274 aminohapet (Takeuchi, T., jt 2012). Inimesel on need numbrid karbi omadest suuremad, keskmine eksonite hulk geeni kohta on 8.8, keskmine eksoni pikkus 145 ja valgu keskmine pikkus on 447 aminohapet (International Human Genome Sequencing Consortium 2001). Samas oli meie hinnangul pärlikarbi keskmine eksonite arv geeni kohta palju suurem: 9.7 eksonit geeni kohta.

V. Koonustigu *Conus consors*

Meie töögrupis uuritud teoliik *Conus consors* (Joonis 7.) kuulub mürgiste gastropoodide sugukonda *Conidae* koos veel *ca* 700 teise liigiga (http://www.conco.eu/cone_snails.html). Koonusteod on röövloomad, kes toituvad põhiliselt ussidest, teistest molluskitest või kaladest.

Joonis 7. *Conus consors,* autor Jan Delsing, 2011.

C.consors koja suurus varieerub keskmisest kuni suureni (50 – 118 mm). Vastsetel on kojal tavaliselt 3 keeret ja maksimaalne läbimõõt 0.8 mm, täiskasvanutel 9-11 keeret ja läbimõõt 50 – 90 mm. Koonusekujulise (millest ka sugukonna nimi) koja põhivärv on valge kuni pruun, joonte värvus varieerub kollakaspruunist lillani või tumepruunini. Jalg on alt helepruun ja äärtest tumepruun. *C. Consors* elab India ja Vaikses ookeanis kuni Marshalli saarteni, Malaneesia ja Queenslandi rannikul kuni 200 m sügavusel liivas või mudas. Erinevad alamliigid on kohastunud eluks erinevatel sügavustel. (http://biology.burke.washington.edu/conus/recordview/record.php?ID=717l66l10129713421llll&tabs=21101011&frms=0&pglimit=&offset=&res=gengrp&srt=&sql2=)

C. Consors kuulub kalatoiduliste koonustigude hulka keda iseloomustab väga omapärane jahipidamisviis. Need röövloomad on varustatud keerulise jahiparaadiga mille abil nad tulistavad saaklooma kehasse tugevatoimelist mürgisegu mis halvab või tapab selle.

Mürgijuha, mis on kuni kolm korda teo kojast pikem, ühes otsas asub mürginääre ja teises lihaseline paun, milles paiknevad kaltsiumkarbonaadist harpuunid. Harpuune sünteesitakse pidevalt juurde, sest kord välja tulistatud harpuuni tigu enam uuesti kasutada ei saa.

Kui tigu tajub enda läheduses saaklooma seab ta mürgiga kaetud harpuuni oma kehas laskeasendisse ning tulistab selle lihaste kontraktsiooni abil kehast välja. Konustigude mürk koosneb põhiliselt lühikestest peptiididest, mida nimetatakse konotoksiinideks. Need mõjuvad saaklooma närvisüsteemile, tekitades kas osalise või täieliku halvatuse. Kalatoidulised teod nagu *Conus consors* suudavad väiksema saaklooma oma mürgiga ka tappa. Kui saak on halvatud või surnud, neelab tigu selle alla. Koonusteod on võimelised neelama ka endast suuremat saaklooma (http://eol.org/pages/50322/details) (http://www.conco.eu/cone_snails.html).

Konopeptiidid on väiksed 15 kuni 40 aminohapet pikad peptiidid. Nad blokeerivad spetsiifilisi ioonkanaleid, retseptoreid ja transportereid. Konopeptiidid on tugeva struktuuriga, mida hoiavad sageli koos disulfiidsillad. Disulfiidsildade muster on üheks konopeptiidide klassifitseerimise kriteeriumiks. Erinevad konopeptiidid on väga varieeruva järejstusega, lisaks esineb posttranslatsioonilist modifitseeimist. Seetõttu interakteeruvad erinevad konopeptiid väga mitmesuguste sihtmärkmolekulidega. Tüüpiline konopeptiidi prekursorvalk koosneb kolmest regioonist: N-terminaalne signaaljärjestus, propeptiidi regioon, ja toksilise peptiidi regioon. Signaaljärjestus tagab konopeptiidi prekursori sekreteerimise rakust välja. Propetiid muudab prekursori inaktiivseks ja ta lõigatakse ära alles lõpliku toksiini moodustumisel mürgijuhas. Konopeptiide klassifitseeritakse superperekondadesse nende konserveerunud signaalpeptiidi järjestuse ja peptiidi tsüsteiinide mustrite alusel.

Nagu ka teised mürgised loomad, pakkuvad koonusteod suurt huvi farmaatsia ja meditsiinitööstusele. Konopeptiidid blokeerivad spetsiifiliselt mitmesuguseid ioonkanaleid ja nende baasil on juba loodud või loomisel esimesed ravimid ja kosmeetikatooted.

VI. II põlvkonna sekveneerimistehnoloogiad

Frederik Sanger avaldas 1977 aastal ahela lõpetamise meetodil põhineva DNA sekveneerimistehnoloogia, ning Walter Gilbert samal aastal teise meetodi, mis baseerus DNA keemilisel modifitseerimisel ja järgneval lõikamisel kindlate aluspaaride kohalt. Kõrge efektiivsuse tõttu ja sellepärast, et puudus vajadus radioaktiivsete isotoopide järele, sai Sangeri meetodist põhiline sekveneerimistehnoloogia akadeemilises ja kommertskasutuses. Pärast aastatepikkust arendustööd esitles Applied Biosystems 1987 aastal esimest automatiseeritud sekvenaatorit, mille nimeks oli AB370. AB370 suutis detekteerida 96 aluspaari korraga, 500 kb päevas. Ühe lugemi *(read)* pikkus võis olla kuni 600 aluspaari. Samal tehnoloogial põhinev kaasaegne sekvenaator AB3730xl suudab töödelda 2.88 Mb päevas, lugemi pikkus võib ulatuda 900 aluspaarini. Sangeri meetodil põhinevat sekveneerimistehnoloogiat nimetatakse esimese põlvkonna tehnoloogiaks, sest see oli esimene suure jõudlusega meetod DNA järjestuse määramiseks. Seda tehnoloogiat kasutatakse laialt ka kaasajal, sest ta võimaldab saada pikki ja kõrge kvaliteediga lugemeid. Samas on esimese põlvkonna sekveneerimismeetodi puuduseks reagentide suur kulu, aeglus ja kõrge hind (Liu, L., jt 2012).

Teise põlvkonna tehnoloogiad erinevad esimese põlvkonna tehnoloogiast massilise paralleliseerimise ja miniaturiseerimisega. See on suurendanud sekveneerimise jõudlust ja alandanud hinda mitmete suurusjärkude võrra. Samas on kõigi teise põlvkonna sekveneerimistehnoloogiate puuduseks lühikesed lugemid ja suur vigade arv. On olemas kolm suurimat ja enim kasutatavat paralleelse sekveneerimise süsteemi: Roche 454, AB SOLiD, Illumina GA/HiSeq (Liu, L., jt 2012).

Roche 454 oli esimene kaubanduslikult edukas teise põlvkonna süsteem. See sekvenaator kasutab pürosekveneerimise tehnoloogiat, mis põhineb nukleotiidi ahelasse lülitamisel vabastatud pürofosfaadi detekteerimisel. Spetsiifiliste adaptoritega DNA raamatukogud (*libraries*) denatureeritakse üheahelalisteks ja seotakse amplifikatsiooni kuulikestega PCR emulsiooniga. Pikotiiterplaadil liidetakse üks trinukleotiididest (dNTP) DNA polümeraasi poolt sünteesitavale ahelale. Reaktsioonisegus on lisaks ATP sülfurülaas, lutsiferaas, lutsiferiin ja adenosiin 5' fosfosulfaat. Reaktsiooni käigus vabaneb pürofosfaat, mille hulk võrdub ahelale lisatud nukleotiidide hulgaga. Pürofosfaadist tekitatud ATP muudab lutsiferiini oksülutsiferiiniks ja genereerib nähtavat valgust. Samal ajal paardumata alused degradeeritakse apüraasi toimel. Seejärel lisatakse järgmine kogus trinukleotiide ja pürosekveneerimise reaktsioon kordub. 2005 aastal võimaldas Roche 454 tehnoloogia lugemite pikkust 100-150 aluspaari ja genereeris kuni 20 Mb järjestusi ühe sekveneerimiskorraga. 2008

aastal suurendati lugemite pikkust 700 aluspaarini. Lugemite täpsuseks saadakse peale filtreerimist 99.9%. Täiustatud tehnoloogia võimaldab 24 tunni jooksul genereerida kuni 700 Mb järjestusi. 2009 aastal suurendati tehnoloogia jõudlust 14 miljardi aluspaarini 24 tunni jooksul. Suurim Roche 454 tehnoloogia eelis on tema kiirus, samuti on ka lugemite pikkus võrreldes teiste meetoditega suurem. Tehnoloogia puuduseks on eelkõige vead hompolümeeride pikkuse hindamisel. Selle tehnoloogia hinnaks on $12.56 * 1 Mb kohta (Liu, L., jt 2012).

2006 aastal tuli turule SOLiD tehnoloogia. See sekvenaator kasutab ligeerimisel põhinevat kahe aluse sekveneerimise tehnoloogiat. Raamatukogude sekveneerimine toimub 8 alus-proovi ligeerimisega. Ligeeritav fragment sisaldab ligeerimissaiti (esimene alus), lõikamissaiti (viies alus) ja nelja fluorestseeruvat värvi (seotud viimase alusega). Fluorestseeruv signaal detekteeritakse siis, kui proov on komplementaarne matriitsahelaga ja signaal kaob, kui viimased 3 alust lõigatakse proovilt ära. Fragmendi järjestuse saab määrata pärast viiendat sekveneerimisringi kasutades trepp-praimereid. Esialgu oli SOLiDi tehnoloogia lugemite pikkus 35 bp ja ta väljastas 3 Gb järjestusi ühe sekveneerimisoperatsiooni käigus. Lugemite täpsuseks oli peale filtreerimist 99.85%. 2010 aastal ilmus täiustatud versioon lugemi pikkusega 85 bp, täpsusega 99.99% ja ühe sekveneerimisoperatsiooni mahuga 30 Gb. Selle tehnoloogia hinnaks on $40 * 1 Gb kohta (Liu, L., jt 2012).

Illumina GA *(Genome Analyzer)* tuli turule 2007 aastal. See sekvenaator kasutab sünteesil põhinevat sekveneerimistehnoloogiat. Fikseeritud adaptoritega raamatukogu denatureeritakse eraldi üksikahelateks ja viiakse küvetti, kus teostatakse sild-amplifikatsioon *(bridge amplification)*. Selle käigus moodustuvad kloon-DNA fragmente sisaldavad klastrid. Enne sekveneerimist lõigatakse raamatukogu linearisatsiooni *(linearisation)* ensüümi vahendusel üksikahelateks. Seejärel sünteesitakse komplementaarsed ahelad, kasutades nukleotiide millega on seotud üks nejast erinevast fluorestseeruvast värvist ja keemiliselt eemaldatav sünteesi jätkumist blokeeriv rühm. GA väljastas algselt 1Gb järjestusi ühe sekveneerimisoperatsiooni kohta. Praeguseks on tehnoloogia jõudlust suurendatud 50 Gb ühe sekveneerimiskorra kohta. Samal põhimõttel töötav HiSeq tehnoloogia väljastab kuni 600 Gb ühe sekveneerimisoperatsiooni käigus. Lugemi pikkuseks on 50SE, 50PE või 101PE, täpsus 98% (100PE). See on hetkel kõige odavam sekveneerimise tehnoloogia $0.02 * 1 Mb kohta (Liu, L., jt 2012).

Teise põlvkonna sekveneerimise üheks peamiseks puuduseks on lühikesed lugemid. Täispika genoomi kokku panemine lühikestest lugemistest on keeruline bioinformaatiline ülesanne, mis sõltuvalt

katvusest, sekveneerimise kvaliteedist ning genoomi iseloomust võib olla lahendamatu. Pika genoomiga liikide, nagu *C.elegans*, sekveneerimisel jääb suur osa genoomist katmata. 50 aluspaariste lugemitega on seni suudetud ainult 51% genoomist panna kokku 10000 aluspaaristeks või pikemateks kontiigideks. Kvaliteetsema tulemuse saamiseks on kaks võimalust. Esiteks kui sekveneeritakse liiki, millele lähedase liigi genoom on juba kvaliteetselt kokku pandud, saab olemasolevat genoomi kasutada uue genoomi kokkupanemisel abimaterjalina. Teiseks võib suurendada katvust (lugemite arvu) ja kasutada efektiivsemaid programme lugemitest genoomijärejstuse kokku panemiseks (assembleerimiseks) (Pop, M., Salzberg, S. L., 2008).

See, missugust programmi sekveneerimisandmete kokkupanekuks kasutada sõltub sellest, milliste andmetega on tegemist ja missugust eesmärki tahetakse saavutada. Näiteks resekveneeritud genoomi jaoks sobivad ka tavalised BLAST või BLAT (*Blast Like Alignment Tool*), kuna referentsjärjestus, millele uus genoom kaardistada on juba olemas, ning resekveneerimise eesmärgiks on eelkõige tühjade kohtade täitmine. Kui tegemist on *de novo* sekveneeritud genoomiga, siis on vaja kasutada keerulisemaid algoritme. Kui probleemi tekitavad kordusjärjestused, siis tuleks kasutada hierarhilise sekveneerimise strateegiat SHARP (*Short Read Assembly Protocol*), kus genoomi jaotatakse suurte fragmentide kaupa teekidesse, ning igat fragmenti sekveneeritakse SRS meetodil (*Short Read Sequencing*). Lugemeid kasutatakse selleks, et määrata BAC-kloonide jaotus pikki genoomi. Assembleerimine põhineb jaotuse lokaalsetest regioonidest pärit lugemite kokku panemisel. Individuaalsed kokkupandud lugemid kombineeritakse kokku BAC-kloonide jaotuse põhjal. Kvaliteetsete kokkupandud lugemite saamiseks peab lugemite pikkus olema 200 bp või rohkem (Pop, M., Salzberg, S. L., 2008).

SRS andmestiku kokku panemiseks on loodud mitmeid assembleerimisprogramme (assemblereid). Näiteks Newbler, mis kuulub paketti 454 Life Science instrumentidega ja mida on edukalt kasutatud bakterite genoomi kokku panemiseks. Väga lühikeste 30 – 40 bp lugemite jaoks on olemas programmid SSAKE, VCAKE ja SHARCGS, mis kasutavat sarnast meetodit. Nimelt valitakse välja lugemid millest tehakse “seemneid” kontiigide formeerimise jaoks. Iga selline seeme pikendatakse identifitseerides lugemeid, mis on seemnega ülekattes kas 3' või 5' otsast. Laiendamise protsess kasvatab kontiige tsükliliselt senikaua kuni on üheseid laiendeid - see tähendab, kuni lugemite järjestuste vahel, mis kattuvad kasvava kontiigi otsaga, ei ole erinevusi. Kuigi selline protseduur suudab vältida fragmentide vale kokkupanekut kordusjärjestuste kohalt, on tema puuduseks lühikesed kontiigid. *De novo* genoomi kokku panemiseks on olemas veel üks strateegia, mis kasutab hübriidi SRS ja Sanger assembleritest.

See võimaldab kahandada sekveneerimise hinda Sangeri meetodiga võrreldes, kattes samal ajal kloonimisvigade tõttu puuduvad lüngad. Selle strateegia käigus kasutatakse Newbler'it kontiigide kokku panemiseks SRS andmetest. Seejärel lõigatakse kokku pandud kontiigid Sangeri tehnoloogiale vastava suurusega fragmentideks. Need fragmedid omakorda pannakse kokku Celera assembleriga. Ülal toodud assemblerid kasutavad standardset lähenemist, mis genoomi kokkupanemise käigus käsitleb igat lugemit kui diskreetset ühikut (Pop, M., Salzberg, S. L., 2008).

On olemas ka teine perspektiivne lähenemine, nimelt assembler, mis kasutab deBruijn graafi. Assembler alustab lugemite hulga jagamisest lühikeste DNA lõikude (k-meeride) hulgaks. Graaf konstrueeritakse nii, et lõigus moodustavad sõlmed ja kaks lõiku on omavahel seotud siis, kui nad on ühes originaalses lugemis kõrvuti. Õige genoomi kokku panek on esitatav kui tee läbi selle graafi, mis läbib kõiki servi. Kuna algsed lugemid jagatakse siin väikesteks lõikudeks töötab see meetod lühikeste lugemite korral paremini. Lisaks võimaldab see erinevate pikkustega lugemeid lihtsa mehhanismiga kokku panna (Pop, M., Salzberg, S. L., 2008). See, milline meetod on konkreetse genoomi kokku panemiseks parim sõltub eelkõige sekveneerimistehnoloogiast ja genoomi katvusest. Lisaks on aga olulised ka genoomi enda omadused, eelkõige suurus, kordusjärjestuste hulk ja nende iseloom.

VII. AUGUSTUS

Geeniennususprogramme jaotatakse tavaliselt kolmeks rühmaks: *ab initio* programmid, sarnasusel põhinevad (*similarity based*) programmid ja kombineerivad programmid. Esimesed kaks kasutavad bioloogiliste signaalide matemaatilisi mudeleid. *Ab initio* programmid kasutavad treeningandmestikke teada oleva geenistruktuuriga selleks, et n.ö. treenida bioloogiliste signaalimudelite parameetreid. Sarnasustel põhinevad programmid kasutavad lisaks teadaolevale DNA järjestusele veel välist informatsiooni, nagu näiteks DNA järjestuse homoloogiat valguga või mõne teise DNA järjestusega (Stanke, M., Waack, S., 2003).

Mõlemal ülaltoodud programmitüübil on omad plussid ja miinused. *Ab initio* programmid ei ole nii täpsed kui sarnasusel põhinevad programmid. Sarnasusel põhinevad programmid seevastu ei suuda leida geene mittehomoloogilistel piirkondadel (Stanke, M., Waack, S., 2003). *Ab initio* programmide täpsust on tavaliselt hinnatud lühikestel järjestustel, mis sisaldavad ainult ühte geeni ja lühikest külgnevat DNA järjestust. Sellistel järjestustel on parimad *ab initio* programmid saavutanud väga häid tulemusi.

Kombineerivad programmid on kõige täpsemad ja usaldusväärsemad, kuna nad kombineerivad mitme teise ennustusprogrammi tulemusi. Näiteks võib kombineerida *ab initio* ja mõne sarnasusel põhineva programmi tulemusi. Mida rohkem erinevate ennustusprogrammide tulemusi kasutatakse, seda parem on tavaliselt ennustus (Stanke, M., Waack, S., 2003).

Viimasel ajal on hakatud looma liigispetsiifilisi geeniennustusprogramme, mida ei ole võimalik treenida erinevate liikide jaoks. Samas on nad tihti väga edukad ühe konkreetse liigi geenide ennustamisel. Selliseks programmiks on näiteks YGAP (*Yeast Genome Annotation Pipeline*), mis on mõeldud uute pärmide, mille kohta puuduvad transkriptsiooniandmed, genoomilt geenide ennustamiseks (Proux-Wéra, E., 2012). Programm teeb automaatselt *de novo* annotatsiooni, kasutades homoloogiaid teiste pärmidega ja sünteenilist informatsiooni Yeast Gene Order Browser andmebaasist. Võrreldes YGAP ja AUGUSTUSe geeniennustusi *Saccharomyces cerevisiae* genoomil leiti, et YGAP oli AUGUSTUSest parem. YGAP leidis täpselt 5119 geeni koordinaadid, samas kui AUGUSTUS leidis täpselt ainult 4938 geeni koordinaadid. YGAPi ennustuste hulgas oli valepositiivseid 99 ja valenegatiivseid 44, AUGUSTUSe ennustuste hulgas vastavalt 117 ja 172. Vale start või stop koordinaadiga geene oli YGAP ennustanud 376 ja AUGUSTUS 483. Autorid ise rõhutavad, et YGAP programm on väga pärmispetsiifiline ja on loodud eelkõige selleks, et võtta arvesse pärmispetsiifilisi

eripärasid (näiteks haruldased intronid) ja teadaolevaid genoomi omadusi (näiteks geenide järjekorra konserveerumine pikkadel distantsidel), ning see ei sobi teiste seeneliikide jaoks (Proux-Wéra, E., 2012).

Geeniennustusprogramm AUGUSTUS avaldati esmakordselt aastal 2003. Ta baseerub varjatud markovi mudelil (HMM). AUGUSTUSe eeliseks on see, et ta kasutas palju erinevaid mudeleid – sealhulgas traditisioonilisi, nagu oligonukleotiidide sagedused ja uusi, nagu doonor splaisingsaitide mudel (Stanke, M., Waack, S., 2003).

AUGUSTUSe jaoks loodi uus HMM mudel DNA oleku kirjeldamiseks. HMM on tõenäosuslik mudel, mis koosneb bioloogilisele tähendusele vastavatest seisunditest (nt. intron, ekson, splaisingsait) ja nendevahelistest üleminekutest. Üleminekutel on võimalik arvestada bioloogilisi seaduspärasid (näiteks peab ekson algama ja lõppema splaisingsaidiga). Mudel määrab geeni oletatava struktuurse tõenäosusjaotuse DNA järjestustel. AUGUSTUSes kasutatava HMM mudeli iga olek võib potentsiaalselt vastata suvalise järjestuse ja juhusliku pikkusega DNA fragmendile. Ahelate jaotused eri olekute vahel ja nendevaheliste üleminekute tõenäosused määratakse kindlaks kasutades treeningandmestikku, mis koosneb uuritava liigi juba annoteeritud geenijärjestustest. Selleks, et määrata olekute jaotust kasutatakse olemasolevaid mudeleid nagu Markovi ahel, kõrgemat järku *windowed weight array* mudel *(WWAM)*, interpoleeritud Markovi Mudel, ning sarnasusel põhinevat järjestusmustrite kaalumist (Stanke, M., Waack, S., 2003).

AUGUSTUS kasutab lisaks uut meetodit, mis võimaldab täpsemalt modelleerida intronite pikkusi ning mida saab kasutada ka teistes HMM mudelitel põhinevates geeniennustusprogrammides. Lühikeste intronite jaoks modelleeritakse võimalikult täpne pikkuste jaotus, pikkade intronite jaoks kasutatakse geomeetrilist jaotust. Splaisingsaitide mudel põhineb sellel, et empiirilist jaotust kasutatakse, kui tõenäosusliku mudelit. Doonor splaisingsaidi puhul see empiiriline jaotus silutakse võttes arvesse, et mustrid, mis on sarnased splaisingu saitide mustritele, osutuvad suure tõenäosusega splaisingsaitideks.

AUGUSTUSes on kasutusel järgmised alammudelid:

- translatsiooni initsiatsiooni motiiv *(translation initiation motif)*– kolmanda järgu WWAM, akna suurusega *(window size)* 5' piirkonnas 20 alust enne translatsiooni alguspunkti.
- start koodon *(start codon)* – ATG tõenäosus on 1, kõigil teistel koodonitel 0.
- algmuster *(initial pattern)* – muster *p* maksimaalse pikkusega kuni neli alust, koos tõenäosusega, mis on antud selle mustri suhtelise sagedusega treeningsetis olevate kodeerivate järjestuste vastavas lugemisraamis. Mustri pikkuseks on 4 seni, kuni eksoni pikkus ei hakka

lubama ainult lühikesi mustreid (alla 4 nukleotiidi).

- algsisu mudel *(initial content model)* - interpoleeritud 3-perioodine neljandat järku Markovi mudel. Kui eksoni pikkus seda võimaldab, siis on väljastava järjestuse pikkus 15 nukleotiidi.
- eksoni sisu mudel *(exon content model)* – interpoleeritud neljandat järku Markovi mudel, mis on treenitud kõikidel treeningandmestiku geenijärjestustel.
- dss mudel *(dss model – donor (5')splice site model)* – arvesse võetakse ainult kanoonilisi splaising saite, mis vastavad GT-AG reeglile. See reegel kehtib 99% imetajaliikide jaoks. Mudel väljastab eksoni 3 viimast nukleotiidi, selle järel konsensus dinukleotiidi GT ja 4 introni nukleotiidi.
- hargnemispunkti mudel *(branch point model)* – kolmanda järgu WWAM akna suurusega 7, väljastab 32 nukleotiidi.
- ass mudel *(ass model – acceptor (3')site model)*- mudel väljastab kolm introni nukleotiidi, mis asuvad enne AG dinukleotiidi konsensust, siis AG nukleotiidi konsensuse ja seejärel eksoni esimese nukleotiidi. 4 vaba nukleotiidi (mis ei kuulu splaisingsaidi konsensusesse) muster määrab tõenäosuse vastavalt sama mustri esinemissagedusele treeningseti vastavates positsioonides.
- Sisemine 3' sisu mudel *(internal 3' content model)* – interpoleeritud 3-perioodine neljandat järku Markovi mudel. See on treenitud viiel nukleotiidil positsioonides -8 kuni -4, vastavalt doonor splaisingsaidile, kasutades kõiki sisemisi *(internal)* eksoneid treeningsetis.
- Stop koodon *(stop codon)* – väljastab TAG, TGA või TAA tõenäosustega 24%, 48% ja 28%.
- geenisisene regioon *(intergenic regioon)* – neljandat järku Markovi mudel, mis on treenitud kõikidel mittekodeerivatel järjestustel treeningsetist. Väljastab ainult ühe nukleotiidi korraga.

2006 aastal ilmus AUGUSTUSe uuendatud versioon AUGUSTUS+, mis suudab kasutada niinimetatud "vihjeid" *(hints)*. „Vihjed" leitakse uuritava genoomse järjestuse vastavustest EST (*expressed sequence tag*) ja valkude vahel. Seega AUGUSTUS+ kombineerib nii sisemist, kui ka välist informatsiooni geenide kohta. Väline informatsioon on täiendav tõend DNA järjestuse *s* geenistruktuuri kohta, mis pärineb teistest allikatest kui sisemine mudel. Tavaliselt saab sellist informatsiooni genereerida võrreldes järjestust *s* teiste järjestustega, nagu ESTid või teise lähedase liigi DNA. Põhimõtteliselt võivad ka eksperdi teadmised olla tõendiks geenistruktuuri kohta. „Vihjeid" erinevatest infoallikatest saab kasutada üheaegselt, võttes samuti arvesse nende usaldusväärsuse astet. Mudel võtab arvesse neid „vihjeid", mis näitavad, et järjestuse *s* mingi regioon on osa suurest eksonist, või et see on terve ekson.

„Vihjeid“ on kuut tüüpi: translatsioonisaidi, stop koodoni, doonor splaisingsaidi, aktseptor splaisingsaidi, kodeeriva regiooni ja eksoni „vihjed“. Iga „vihje“ võib viidata kas kodeerivale või mittekodeerivale DNA ahelale. „Vihjetel“ on võimalikud neli erinevat hinnet: käsitsi koostatud, ESTide vastu joondatud, valgu vastu joondatud ja kombineeritud (ESTide ja valgu vastu joondatud). Selleks laiendati esialgses AUGUSTUSes kasutatud GHMM (*generalized hidden markov model*) täiustatud GHMMiks, mis arvestab geeniennustamisel ka „vihjeid“. Geenistruktuuri tõenäosused on samad, mis olid esialgses AUGUSTUSes. AUGUSTUS+ arvestab ainult neid „vihjeid“, mis on kokkusobivad analüüsitava DNA järjestusega, mitte kokkusobivaid „vihjeid“ arvesse ei võeta. Juhul, kui samale regioonile on olemas erinevat tüüpi „vihjed“, jäetakse alles kõige usaldusväärsemad. Need potentsiaalsed geenistruktuurid, millel on olemas „vihjeid“, saavad plusspunkte ja need, millel ei ole „vihjeid“ miinuspunkte.

Start saidi „vihje“ on järjestusega kokkusobiv ainult siis, kui ATG asub õiges positsioonis, sõltuvalt „vihje“ olemasolust ühel või teisel ahelal. Stop saidi „vihje“ nõuab järjestusel stop koodoni olemasolu ja *ass* ning *dss* „vihjed“ nõuavad intronite dinukleotiidset konsensusjärjestust 'GU-AG'. Eksoni osa (kodeeriva regiooni) „vihje“ on kokkusobiv DNA järjestusega ainult siis, kui regioon ei sisalda stoppkoodoneid vastava ahela vastavas lugemisraamis. Viimast tüüpi eksoni „vihje“ on kokkusobiv DNA järjestusega siis, kui regioon määratud ahelal ei sisalda *in-frame* stop koodonit ja regioon on seotud splaisingu saidiga, start või stop koodoniga. AUGUSTUS+ ignoreerib BLASTi joonduse p arvu (näitab tõenäosust, et sündmus võis toimuda juhuslikult) ja e arvu (p arvu korrektsioon mitmekordse testimise järel), kuna nende kasutamine ei parandanud ennustustulemusi. Kuna EST „vihjed“ ei anna informatsiooni selle kohta, missugused järjestused on valku kodeerivad ja missugused mitte, siis see meetod leiab üles ka mittekodeerivaid eksoneid (Stanke, M., jt 2006).

Kolmas AUGUSTUSe uuendatud versioon avaldati jaanuaris 2011 aastal – AUGUSTUS PPX (*protein profile extension*). Uus programm kasutab blokkprofiile, mis on moodustatud mitme järjestuste joondustest, ja kujutavad endast ennustuse valgulist tõestust. Blokkprofiil iseenesest on positsioonispetsiifiliste sageduste maatriksite kogum, mis kirjeldab aminohappe jaotust ühes blokis ja sarnaneb profiil-HMM'ile. Samas on, vastupidiselt profiil-HMMile, blokkidesse pandud järjestuse motiividel kindel pikkus ning insertsioonid ja deletsioonid ei ole blokisiseselt lubatud. Kuigi profiil-HMM on üldjuhul täielikum järjestuse mudel, valiti AUGUSTUSe valgulise signatuuri kirjeldamiseks blokkprofiilid, kuna itegratsioon GHMM'i vajab madalamat kompleksust. Selle asemel, et kasutada välisinformatsiooni allikana teiste programmide väljundit, nagu seda oli tehtud varasemas

versioonis „vihjete“ puhul, blokkprofiili kaardistamine (*mapping*) sihtmärkjärjestusele toimub paralleelselt *ab initio* ennustamisega arvestades ühtlasi teise ahela järjestust. Mitme valgujärjestuse joonduse kasutamine annab hea täpsuse kodeerivate alade leidmiseks, kuid ei sisalda informatsiooni intronite kohta (Keller, O., jt 2011).

Mida rohkem andmeid on kasutusel, seda täpsemaks muutub ennustamine. Geenide ennustamiseks piisab ka ainult treeningandmestiku olemasolust. „Vihjete“ lisamine tõstab ennustuse täpsust ning valguandmete lisamine muudab selle veelgi täpsemaks.

Võib arvata, et paremini uuritud liikide puhul kasutatakse tulevikus rohkem liigispetsiifilisi geeniennustusprogramme. Universaalsed treenitavad programmid nagu AUGUSTUS on vajalikud eelkõige uute ja halvasti kirjeldatud genoomide annoteerimiseks. Kuivõrd aga uusi genoome sekveneeritakse juurde järjest kiirenevas tempos, siis jäävad universaalsed geeniennustusprogrammid pikaks ajaks oluliseks etapiks uute geenide annoteerimisel.

Eksperimentaalne osa

I. Töö eesmärk

Meie töörühma poolt sekveneeriti hiljuti kalatoidulise mereteo *Conus consors* genoom. Teise põlvkonna meetoditega sekveneeritud genoomidele tüüpiliselt on see madala kvaliteediga ning teadaolev järjestus koosneb suurest hulgast lühikestest fragmentidest. Samas on aga molluskite hõimkonna esindajate genoome sekveneeritud vähe ja nende geenide struktuurist ei ole palju teada. Kuigi töörühma eesmärgiks oli eelkõige uudsete konopeptiidide leidmine ja uurimine, pakkus meile ka huvi saada kätte maksimaalselt palju informatsiooni genoomi, geenide koguhulga, struktuuri, kordusjärjestuste ja valgustruktuuri kohta.

Minu praktilise töö eesmärgiks oli analüüsida, kui edukalt on võimalik kasutada automaatseid geeniennustusprogramme selleks, et analüüsida uudsete organismide genoome, sealhulgas kirjeldada nende geenide struktuure ja hinnata geenide koguarvu. Täpsemad ülesanded olid:

1. Käivitada ja liigispetsiifiliselt treenida geeniennustusprogramm AUGUSTUS, vajadusel kasutada täpsust parandavaid andmeid.
2. Hinnata AUGUSTUSe poolt väljastatud tulemuste usaldusväärsust.
3. Välja selgitada kas teiste molluskite geene saab usaldusväärselt ennustada kasutades fülogeneetiliselt lähema organismi (teo) peal treenitud mudelit või sobib selleks paremini väga hästi treenitud, kuid fülogeneetiliselt kauge inimese geenimudel.
4. Valguhomoloogia meetodi tulemusi kasutades välja selgitada geenide struktuur.
5. AUGUSTUSe ennustuste ja valguhomoloogia meetodil põhinevate tulemusi kombineerides hinnata teo geenide koguhulk.

II. Materjal ja metoodika

1. *C. consors*'i genoomi ja transkriptoomi kokku panemine.

C. consors'i sekveneeritud genoom pandi kokku kahes etapis – eeltöötlus ja assembleerimine. Kõigepealt lõigati maha Roche 454 ja Illumina lugemite madalakvaliteedilised 3' otsad programmiga *fastq_quality_trimmer* FASTX Toolkit v.0.0.13 pakettist. Kvaliteedi piiriks oli 30 ja lugemite miinimumpikkuseks 50 bp. Seejärel puhastati lugemid eemaldades need järjestused, mis andsid tugeva homoloogia bakteriaalsete või inimese järjestustega programmi DeconSeq 0.4.1 abil (Schmieder, R., Edwards, R., 2011). Lugemite testimisel inimese genoomi ja NCBI 2370 bakteritüve vastu kasutati identsuse ja katvuse piiridena 0.9. Vektorid ja teised sekveneerimisel kasutatavad järjestused eemaldati programmiga *seqclean*. Antud programmi kasutati vaikimisi parameetritega, muudeti ainult lugemite pikkust („-l 50“), ning lülitati välja polüA/T otste mahalõikamine ja madala keerukusega fragmentide eemaldamine („-A -L“). Lugemid testiti UniVec andmebaasi vastu.

Genoomi kokkupanek (*assembly*) koosnes kahest eraldi sammust. Esiteks loodi algne genoomi versioon Illumina „*pair-end/mate-pare*“ lugemitest kasutades programmi SOAPdenovo 1.05 (Li, R., jt 2010). Eesmärgiks oli luua Illumina lugemitest kokku pandud genoomist *pseudo* 454 lugemid, et need andmed saaks sisestada Roche GS De Novo Assembler'isse (Newbler). SOAPdenovo programmil lasti käsitleda kordusi lugemitena ja kasutati k-meeri pikkusega 37. Selle tulemusena tekkis palju skaffolde, mis sisaldasid määramata järjestusi (tähistatud „N“ tähtedega). Need skaffoldid lõigati 300 aluspaari pikkusteks tükkideks omavahelise ülekatvusega 200 aluspaari, kasutades EMBOSS *splitter*'it (Rice P, jt 2000), et vältida valet tühikute pikkuste hinnangut. Järgmise sammuna kasutati programmi Newbler 2.7, et panna lõplikuks järjestuseks kokku kolme tüüpi lugemeid: 454 pikkusega 1892 bp, *pseudo* 454 pikkusega 300 bp, ning Illumina pikkusega 145 bp. Kokku koosnes assembleeritud genoomijärjestus 4,513,486 kontiigist. Kvaliteediindikaator N50 (kontiigi pikkus, millega võrdsed või pikemad kontiigid katavad 50% kogu genoomsest järjestusest) oli 819 bp.

Selleks, et hinnata kokkupandut genoomi kvaliteeti kasutati *C.consors* täielikult sekveneeritud ja publitseeritud mitokondriaalset genoomi (Brauer A, jt 2012). BLASTn 2.2.26+ (Altschul, S. F., jt 1997) programmiga joondati selle järjestuse vastu kõik kokkupandud genoomi kontiigid. BLAST'i väljastatud joondused filtreeriti ja loeti kokku minimaalne kontiigide arv, mis kataks maksimaalselt mitokondriaalse genoomi.

Teise kvaliteedikriteeriumina kasutati eukarüootide 458 konserveerunud tuumikgeeni järjestusi CEGMA andmebaasist (Parra, G., jt 2007, Parra, G., jt 2009). BLASTX programmi abil leiti, millised nendest geenidest leidusid kontiigides ning millises ulatuses oli nende aminohappeline järjestus kaetud. BLASTX joonduste osas nõuti minimaalset joonduse pikkust 60 aminohapet, minimaalset sarnasust 50% ja lubati kordusjärjestusi. Iga tuumikgeeni jaoks loeti vasteks ainult parima joonduse andnud kontiigi.

2. Valguhomoloogia meetod

Selleks, et leida esialgne geenistruktuur ja hinnata geenide hulka kasutati valguhomoloogia metoodit. Kõigepealt leiti programmi BLASTX abil kõikide genoomi skaffoldite joondused UniProtKB/Swiss-Prot andmebaasi valkude vastu. Saadud vasted grupeeriti valgu funktsiooni ja BLASTX bitiskoori järgi, ning koostati skaffoldite mittekattuvate regioonide ja unikaalsete valgufunktsioonide vastastiku parimate lokaalsete joonduste nimekirja. Seejärel lisati igale parimale joondusele need vasted, mille puhul antud genoomne regioon oli vastava valgu jaoks parim joondus ja järjestati need valgule. Kõik valgul kõrvutiasetsevad (kuid mitte ülekattes olevad) joondused loeti oletatavateks eksoniteks. Kui kaks sellist oletatavat eksonit asetsesid samal kontiigil ning olid valgus kõrvuti, loeti nendevaheline ala kontiigil oletatavaks introniks.

Selle meetodiga loodud oletatavate eksonite ja intronite faili kasutati hiljem, et hinnata keskmist eksonite/intronite hulka geeni kohta. Samuti oli see fail aluseks AUGUSTUSe treeningandmestiku koostamiseks.

3. AUGUSTUS

Selleks, et leida geene mis UniProtKB/Swiss-Prot andmebaasis puudusid (liigisptsiifilised või haruldased valgud), et leida valkude vähekonserveerunud regioone ja et määrata täpsem geenide struktuur kasutati geeniennustusprogrammi AUGUSTUS. AUGUSTUS ver. 2.5.5. paigaldati töögrupi serverisse programmiga kaasas oleva juhendi järgi (http://augustus.gobics.de/binaries/README.TXT) ja loodi liigispetsiifiline kaust, kus programm hoiab liigispetsiifilist metaparameetrite ja parameetrite faile (http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#meta).

Metaparameetriteks nimetatakse AUGUSTUSE programmis analüüsi sisemist metoodikat muutvaid parameetreid nagu splaisingsaidi mudelite akna suurus ja Markovi mudelite järjekord. Parameetriteks nimetatakse kodeerivate ja mittekodeerivate regioonide k-meeride esinemissagedusi. Metaparameetrid

määravad ära selle, kuidas arvutatakse parameetrid. Liigispetsiifiline kaust sisaldab kümmekond metaparameetrit, ning tuhandeid parameetreid. Metaparameetrite fail sisaldab endas lisaks veel erinevaid väljundi ning programmi seadeid, mida saab ilma faili muutmata anda ka käsurealt. Käsurea sisendiks faili loomise jaoks oli:

new_species.pl –species=conus_consors

Siin ***new_species.pl*** on AUGUSTUSe skript, mis loob kausta koos metaparameetrite failiga, ning argument *--species=conus_consors* määrab ära kausta ja selles oleva faili nimetused – tavaliselt kasutatakse selleks uuritava liigi nime.

3.1. Treeningandmestiku loomine

Enne, kui programmiga hakati geene ennustama tuli luua treeningandmestik. Treeningandmestik koosneb järjestustest mille kohta on usaldusväärselt teada geenide struktuur (eksonite ja geenide alguse ja lõpu koordinaadid). Selleks piisab ka ainult kodeerivatest järjestustest (*CDS – coding DNA sequences*). Kuivõrd liik *C.consors* oli seni sekveneerimata ja andmebaasides selle kohta infot vähe loodi treeningandmestik kasutades assembleeritud genoomi (joonis 8) ja valguhomoloogia meetodil

```
>contig607426  length=993   numreads=431
ACACACACACACACACACACACACACACACACGCACGCACGCACGCACACACATCCTCTC
TCTTTCCTTTCCTGAAATAGAAACAGACGTACATCTCCATTCTCGTTCTTCTCCTCaTGT
TTCTTCCTTTTTACCCTCCTCCTCCTCCTCGTCGTTGTCGTCGTCGTCGTCATCATTATC
ACCATCGTCGTCGTCGTTGTCGTTGTCATTGTCGTCATCATCATCATCATCATCATCATC
ATTATCATCATCGTCGTCGCCGTCGTCATCGCCGTCGTCGTCGTTGTCGTCGTCATTGTC
GTCGTTGTAGTCGTCATCATCATCATAATCATCGTTGTTGTTGTTGTTGTTGTTGTTGTT
GTCGgTGGTGGTGGTGGTGGTGGTGGTGGTGTTAGTCTCGTCGTCATTGTCCACCAG
CTCCTCCTGATCACCAcCGTGTTCCTTGTAAACAGAAAGGAAATAAGAGAAAGGAGGAga
gAAGTGAGAAGAGGTATGAGGGAAATGAAGAACaGAAATGAAGGAATATCAAAGAAATGG
AAGTTTAGGTCATAAAGATTGTAGCATTGAATGATAAAGATATAACAAGATCAGATAACT
ATACAAAAAAAAaCCcGTCAAGAACGACAACCAGAACGACAACAGCTGCAAATAAAAAGA
AAAAGAACAACtAAAAGAAGCAAGAAGAACAGAAcATTTCAAcGTAAAGCGAACAAGAAG
GGAAACCATAAAGTTTGTCTCTTTTATTCTCCAATTTATCATCACATCACAACCAGGCTC
TAAGGgTGATTTTCTTCTTAAAAACaAGaTcTTTTTTTctaAAAGATGCTTTACCTCCTC
TCTCATAACTTATGTGACTCACACCTGATCTCATTTTTTGCTTGTAaGAAGGAGACAAGT
GcTGGTTAAAATGAAAATAGAGGTAGCGTTGCTGCTGAACAATTTTTCtCCTGCGAGTGT
GTGTGTGTGTGTGTGTGTGTGTGTGTGTG
```

Joonis 8. Andmed *fasta* formaadis. Formaat esindab nukleotiidset või peptiidset järjestust, kus nukleotiidid või aminohapped on esitatud ühetähelises koodis. Algab alati „>“ märgiga, mille järel on tavaliselt olulisemad andmed antud järjestuse kohta ning järgmisest reast algab järjestus ise.

leitud oletatavaid eksoneid (Joonis 9.).

```
Exon:   131     253     contig607748    HMCN2   4909    4949    67.4

Function:       NTRI    80.5
Match:scaffold00002     57110   3       -3      sp|Q99PJ0|NTRI_MOUSE    NTRI    344     33793   34695   34      310
Exon:   33793   34695   scaffold00002   NTRI    34      310     107

Function:       KITM    64.5
Match:scaffold00013     18230   0       3       sp|O00142|KITM_HUMAN    KITM    265     8532    9101    50      220
Exon:   8532    9101    scaffold00013   KITM    50      220     88.6

Function:       KTHY    97.6
Match:scaffold00020     17218   0       -2      sp|P97930|KTHY_MOUSE    KTHY    212     15709   16323   5       211
Exon:   15709   16323   scaffold00020   KTHY    5       211     191

Function:       KC1A    87.5
Match:scaffold00023     16693   0       -3      sp|P67963|KC1A_XENLA    KC1A    337     13743   14627   19      313
Exon:   13743   14627   scaffold00023   KC1A    19      313     564

Function:       PTBP1   9.7
Match:scaffold00026     16389   0       -3      sp|Q00438|PTBP1_RAT     PTBP1   555     9416    9571    239     292
Exon:   9416    9571    scaffold00026   PTBP1   239     292     91.7

Function:       LETM1   47.5
Match:scaffold00030     16199   2       1       sp|Q0VCA3|LETM1_BOVIN   LETM1   732     4444    4758    102     190
Match:scaffold00030     16199   1       1       sp|Q0VCA3|LETM1_BOVIN   LETM1   732     8509    8733    273     347
Match:scaffold00030     16199   0       1       sp|Q0VCA3|LETM1_BOVIN   LETM1   732     10132   10380   346     428
Exon:   4444    4758    scaffold00030   LETM1   102     190     79.3
MISSING_FRAGMENT:       4758    8509    scaffold00030   LETM1   190     273
Exon:   8509    8733    scaffold00030   LETM1   273     347     88.2
Intron: 8733    10132   scaffold00030   LETM1
Exon:   10132   10380   scaffold00030   LETM1   346     428     118
Match:scaffold03961     6392    1       -1      sp|Q5ZK33|LETM1_CHICK   LETM1   752     996     1169    195     250
Exon:   996     1169    scaffold03961   LETM1   195     250     79.7
Match:scaffold06040     5736    0       -1      sp|Q5ZK33|LETM1_CHICK   LETM1   752     2110    2253    658     704
Exon:   2110    2253    scaffold06040   LETM1   658     704     48.9
```

Joonis 9. Fail eksonite ja intronite koordinaatidega. Programmi jaoks on olulised read, kus on kirjas *"Exon:"*, kuna need sisaldavad olulist infot: algus ja lõpp koordinaadid, mis *scaffold*'i või *contig*'i peal ekson asub, funktsiooni nimetust.

Selleks kasutati kohapeal kirjutatud programmi, mis väljastas oletatavad eksonid koos järjestustega g*enbank (gb)* formaadis (Joonis 10.). Edaspidises metoodika kirjelduses kasutame treeningandmestiku failinimena *conus_genes.gb*. Programm luges eksonite-intronite failist eksonite lõpp- ja alguskoordinaadid, skaffoldi nimetuse millel eksonid paiknevad ning vastavalt nendele koordinaatidele arvutas välja, kui pikk järjestus tuli välja võtta genoomi *fasta* failist. Samuti arvutas ta a/c/g/t aluspaaride hulga ning koostas *gb* formaadis väljundfaili.

```
LOCUS       scaffold00073              1429 bp            DNA
FEATURES             Location/Qualifiers
     source          1..1429
     CDS             join(1000..1179)
                     /gene="GMPR2"
BASE COUNT           316 a   282 c   215 g   324 t
ORIGIN
        1 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
       61 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
      121 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
      181 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
      241 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnggagtttg
      301 tgcggtcaaa gctgtatttg tgtcctgctg tctgaagggg tggttcttgg agagacttct
      361 gcctgtgctc tccatcttgc ttccctgata gctccatcca ttcatgcaca cacacaagtg
      421 tgtgacaccg agtctccttt agtccacagc ttcaaacatc ctactatttc caaccttaaa
      481 cagctttaaa cgtcccagag tttcctttcc ctaatgactt cacacatgtg aaagttctgg
      541 ttttgaacgc tctatatttc cattcttaaa acagtttaaa gcatcacaaa ttttccattc
      601 tgaaatgact tttactacaa gtttctataa atggctttgt tttcgtgctc aaaaattcaa
      661 aataagctag aattcaactt ctatgcaaaa gctcagtttc atagtaattc ataaagtttt
      721 gtaaaaaaca aaaagaaaaa aaaaagaaaa aaaaaaagaa aatgtgggac aaaaaaaaat
      781 caataagtga atgttaattc acatccattc attcattttc tccgtgtaga tgcgtttgga
      841 cttccaattc cctttcagac agttaaaacc aaatcttgtg cacaatactt tttctacatt
      901 tcacacccca taagaaaagc cagagactgg gcatatgtac tgacatagcc ctgggatttc
      961 atggtggctc acggatgttt actcgccggt ggtgtagggg gtgaacacct cgttcagctg
     1021 ctggctgacc cggatgaagg tggtgcggcg actgagctcc ttcagtttgg aggctcccac
     1081 ataggtgcag gtggagcgga cgcctcccag gatgtcctgt attgtcttct ccacgtcccc
     1141 ccggtattcg atctgaaccg tcttgccctc agacgccctg tcaaacacaa ccggtgatga
     1201 tgacactgca gctatcccct ctgcactcga agccctggca aacacaacta atgatactga
     1261 cgctgcagat atcccctctg tacttgatat caatgatttg gctgtccaaa accctgtcaa
     1321 atacaactag tgatgtcgac actgcagtta tcccctctgt actcagagcc ctggcaaaca
     1381 caactagtga tgttaactgt atagctatcc cctctgtcct cgaagccct
//
LOCUS       scaffold00348              7050 bp            DNA
FEATURES             Location/Qualifiers
     source          1..7050
```

Joonis 10. Andmed *genbank* formaadis. Genbank formaati kasutatakse GenBank andmebaasis andmete kuvamiseks. *Locus* sisaldab endas lookuse nime, järjestuse pikkust aluspaarides, molekuli tüüpi (DNA/cDNA/RNA/jne). *Features* sisaldab informatsiooni geenides ja nende produktidest, *source* sisaldab järjestuse pikkust, *CDS* koosneb kodeerivate alade koordinaatidest, koordinaatide alla on kombeks lisada ka geeni nimetus, kuhu see kodeeriv ala kuulub. *Base count* sisaldab mitu igat tüüpi alust sisaldab järjestus. *Origin* koosneb järjestusest endast, mis on jaotatud 60 nukleotiidi kaupa ridadesse. Formaadi lõpp sisaldab kahte kaldkriipsu.

3.2. Treeningfaili jaotus testandmestikuks ja treenimisandmestikuks

Treeningfaili jaotati juhuslikult kaheks osaks AUGUSTUSe paketiga kaasas oleva programmiga ***randomSplit.pl*** (http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#trainoptions). Esimeses failis oli 200 järjestust ja selle nimeks oli *conus_genes.gb.test*, teises kõik ülejäänud järjestused ja nimi *conus_genes.gb.train*. Esimest (test) faili oli vaja selleks, et hinnata programmi korrektsust, ning teist (treening) faili programmi "treenimiseks". Treenimine tähendab AUGUSTUSe kodeerivate regioonide parameetrite automaatset liigispetsiifilist määramist etteantud järejstuste põhjal.

3.3. AUGUSTUSe esmane treenimine

AUGUSTUSe esmane treening tehti temaga kaasas oleva programmi ***etraining.pl***, mis teeb algtreeningu (http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#etraining). ***Etraining.pl*** vajab, et ette oleks antud keskkonnamuutuja `$AUGUSTUS_CONFIG_PATH`, mis peab endas sisaldama rada AUGUSTUSe seadistuste (*config*) kausta juurde. Keskkonnamuutuja määramiseks tuli kasutada käsurida:

```
export $AUGUSTUS_CONFIG_PATH=home/user/AUGUSTUS/config
```

AUGUSTUSe seadistuste kaust sisaldab endas erinevate liikide kaustasid, mis omakorda sisaldavad parameetrite faile. Treeningprogramm käivitati järgneva käsureaga:

```
etraining --species=conus_consors conus_genes.gb.train
```

Siin ***etraining*** on kasutatava skripti nimetus, *--species = conus_consors* määrab, et kasutataks meid huvitava liigi kausta. Seal asub vastav metaparameetrite faili, kuhu programm hakkab parameetreid sisse kirjutama. Viimane parameeter on treeningfaili nimi.

Seejärel prooviti ennustada *ab initio* geene *conus_genes.gb.test* failist treeningu käigus loodud liigispetsiifiliste parameetritega. Augustus annab ise hinnangu geeniennustuse täpsusele. Antud töö jaoks AUGUSTUSe hinnang ei sobinud, sest see hindas õigeks ainult ennustatud täisgeene. *C.consors* genoomne järjestus oli aga väga fragmenteeritud ja seetõttu seal leiduvate täisgeenide hulk väike. Selle pärast huvitas meid rohkem eksonite ja intronite piiride ennustamise täpsus. Uutel liikidel ei pruugi geeniennustusprogrammid väga häid tulemusi anda, sest ennustuse kvaliteet sõltub treeningandmestiku kvaliteedist. Selleks, et ennustust parandada on võimalik AUGUSTUSe määratud parameetreid optimeerida.

3.4. AUGUSTUSe parameetrite optimeerimine

Optimeerimiseks on AUGUSTUSe paketiga kaasas programm ***optimize_augustus.pl*** (http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#etraining). Hindamisel jagab see programm *conus_genes.gb.train* treeningandmestiku juhuslikult kümneks võrdseks osaks ja kasutab 9 neist osadest treeningandmestikuna ja kümnendat testandmestikuna. Kõiki 10 osa kasutatakse juhuslikult nii treenimiseks, kui ka hindamiseks. Sealjuures tagatakse, et iga andmestikku kasutatakse vähemalt ühe korra tulemuse hindamiseks. Iga optimeerimistsükli käigus leitakse ühe parameetri parim väärtus. Optimeerimise aluseks on tundlikuse ja spetsiifilisuse kaalutud keskmised aluspaari, eksoni ja geeni tasemel. Metaparameetrite jaoks kordab programm ülalpool toodud evalueerimist ka iga erineva metaparameetri väärtuse jaoks. Kui ***optimize_augustus.pl*** leiab, et mõni väärtus annab parema tulemuse, siis ta korrigeerib metaparameetrite failis olevat väärtust. Ühe parameetri optimeerimine kestab nii kaua, kuni ennustuse täpsus enam ei parane, seejärel hakatakse järgmist parameetrit optimeerima. Kui programm on ühe korra kõik parameetrid ära optimeerinud, alustab ta optimeerimisringi algusest peale tehes maksimaalselt kokku 5 optimeerimistsüklit.
Optimize_augustus.pl nagu ka ***etraining.pl*** vajavab, et ette oleks antud keskkonnamuutuja `$AUGUSTUS_CONFIG_PATH.` Seega enne programmi käivitamist tuli anda käsk:

```
export $AUGUSTUS_CONFIG_PATH=home/user/AUGUSTUS/config
```

Lisaks nõuab ***optimize_augustus.pl***, et oleks määratud keskkonnamuutuja `$PATH` ja see sisaldaks rada AUGUSTUSe programmide (*bin*) kausta juurde. See kaust sisaldab skripte ***augustus.pl*** ja ***etraining.pl***, mida AUGUSTUSe optimeeriv skript vaheldumisi kasutab. Enne ***optimize_augustus.pl*** käivitamist tuli seega anda käsk:

```
PATH=$PATH:/home/user/AUGUSTUS/bin/
```

Optimeerimise protsess käivitati käsuga:

```
optimize_augustus.pl --species=conus_consors genes.gb.train
```

Kui skript on lõpetanud AUGUSTUSe optimeerimise või kui kasutaja selle katkestab, tuleb uuesti käivitada ***etraining.pl*** skript, et uute parameetritega AUGUSTUS üle treenida. Vajadusel võib uute parameetritega ümber treenitud AUGUSTUSe ennustuse kvaliteeti hinnata testandmestiku abil.

Ülalkirjeldatud metoodika abil saab AUGUSTUSe ennustuse muuta liigispetsiifiliseks. Samas ei piisa nendest etappidest väga täpse ennustuse saamiseks.

3.5. Lisaandmestiku kasutamine - „vihjete“ fail

Selleks, et programmi ennustust paremaks muuta, on vaja kasutada niinimetatud „vihjete“ faili *(hints file)* (http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/prediction.html#prephints). „Vihjete“ fail on geenide asukoha ja struktuuri väline tõestus GFF *(General Feature Format)* formaadis. Üldjuhul genereeritakse see joondades transkriptoomi genoomile. Vihjete faili näide on toodud Joonisel 11.

```
scaffold32772   b2h     ep      6293    6504    0       .       .       grp=comp1000028_c0_seq1_Sample1;pri=4;src=E
scaffold00104   b2h     ep      37930   38322   0       .       .       grp=comp100002_c1_seq1_Sample1;pri=4;src=E
scaffold11466   b2h     ep      10805   10999   0       .       .       grp=comp1000039_c0_seq1_Sample1;pri=4;src=E
scaffold18320   b2h     ep      5684    6127    0       .       .       grp=comp100005_c0_seq1_Sample1;pri=4;src=E
scaffold18320   b2h     ep      6249    6258    0       .       .       grp=comp100005_c0_seq1_Sample1;pri=4;src=E
scaffold18320   b2h     intron  6128    6248    0       .       .       grp=comp100005_c0_seq1_Sample1;pri=4;src=E
scaffold15511   b2h     ep      3151    3482    0       .       .       grp=comp1000065_c0_seq1_Sample1;pri=4;src=E
scaffold63226   b2h     ep      4151    4360    0       .       .       grp=comp1000079_c0_seq1_Sample1;pri=4;src=E
contig764626    b2h     ep      11      469     0       .       .       grp=comp100009_c0_seq1_Sample1;pri=4;src=E
scaffold20296   b2h     ep      6582    6959    0       .       .       grp=comp100010_c0_seq1_Sample1;pri=4;src=E
contig1995131   b2h     ep      92      320     0       .       .       grp=comp1000140_c0_seq1_Sample1;pri=4;src=E
contig939903    b2h     ep      182     701     0       .       .       grp=comp1000292_c0_seq1_Sample1;pri=4;src=E
contig1160972   b2h     ep      101     552     0       .       .       grp=comp100033_c0_seq1_Sample1;pri=4;src=E
scaffold17717   b2h     ep      760     1404    0       .       .       grp=comp100034_c0_seq1_Sample1;pri=4;src=E
contig1094502   b2h     ep      96      577     0       .       .       grp=comp100036_c0_seq1_Sample1;pri=4;src=E
scaffold77559   b2h     ep      829     1224    0       .       .       grp=comp1000402_c0_seq1_Sample1;pri=4;src=E
contig1966220   b2h     ep      13      273     0       .       .       grp=comp1000406_c0_seq1_Sample1;pri=4;src=E
scaffold64318   b2h     ep      2222    2441    0       .       .       grp=comp1000437_c0_seq1_Sample1;pri=4;src=E
contig2089715   b2h     ep      52      305     0       .       .       grp=comp100043_c1_seq1_Sample1;pri=4;src=E
scaffold01394   b2h     ep      4013    5401    0       .       .       grp=comp100043_c2_seq1_Sample1;pri=4;src=E
scaffold06767   b2h     ep      11649   11821   0       .       .       grp=comp1000455_c0_seq1_Sample1;pri=4;src=E
scaffold06237   b2h     ep      16029   16429   0       .       .       grp=comp100045_c0_seq1_Sample1;pri=4;src=E
scaffold06237   b2h     ep      16480   17149   0       .       .       grp=comp100045_c0_seq1_Sample1;pri=4;src=E
scaffold06237   b2h     ep      14517   15974   0       .       .       grp=comp100045_c1_seq1_Sample1;pri=4;src=E
scaffold06237   b2h     intron  16430   16479   0       .       .       grp=comp100045_c0_seq1_Sample1;pri=4;src=E
scaffold03076   b2h     ep      7375    7691    0       .       .       grp=comp1000483_c0_seq1_Sample1;pri=4;src=E
contig757362    b2h     ep      35      254     0       .       .       grp=comp1000484_c0_seq1_Sample1;pri=4;src=E
scaffold05205   b2h     ep      19399   19691   0       .       .       grp=comp1000487_c0_seq1_Sample1;pri=4;src=E
scaffold00659   b2h     ep      30451   30791   0       .       .       grp=comp100048_c1_seq1_Sample1;pri=4;src=E
scaffold00659   b2h     ep      30851   31466   0       .       .       grp=comp100048_c1_seq1_Sample1;pri=4;src=E
scaffold00659   b2h     intron  30792   30850   0       .       .       grp=comp100048_c1_seq1_Sample1;pri=4;src=E
scaffold111980  b2h     ep      3496    3983    0       .       .       grp=comp100050_c0_seq1_Sample1;pri=4;src=E
contig2992950   b2h     ep      11      196     0       .       .       grp=comp1000611_c0_seq1_Sample1;pri=4;src=E
scaffold48421   b2h     ep      4144    4175    0       .       .       grp=comp1000621_c0_seq1_Sample1;pri=4;src=E
```

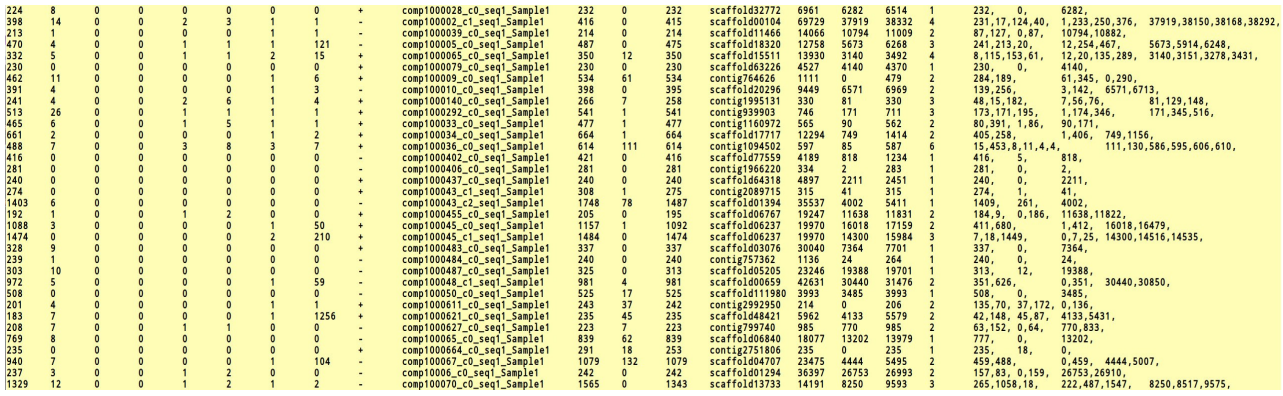
Joonis 11. GFF formaadis fail. I tulp – järjestuse nimetus, II – järjestuse allikas, III – detailid nt. ekson/intron, IV – algus, V – lõpp, VI – skoor, VII – ahel, VIII – raam, IX - omadused

Ennustuse käigus kasutab AUGUSTUS vihjeid, et korrigeerida geenistruktuuri kandidaatide tõenäosusi. Programm annab suurema skoori ja vastavalt väljastab eelistatult neid geenistruktuuri osi mis on „vihjetega“ kooskõlas.

Vihjete faili koostamiseks tuli alla laadida programm nimega BLAT versioon Src35 *(Blast-like Alignment Tool)(*hgwdev.cse.ucsc.edu/~kent/src/blatSrc35.zip*)*. BLAT joondab transkriptoomi genoomile, ning väljastab faili geeniosade koordinaatidega, mida AUGUSTUS saab kasutada vihjetena. Selleks, et genereerida koordinaatidega fail sisestati käsureale:

```
blat -noHead genome.fa transcriptome.fa hints.psl
```

Parameetrit `-noHead` on vaja selleks, et väljundfaili *hints.psl* ei kirjutataks päist. Kuna AUGUSTUS kasutab „vihjeid“ *gff* formaadis, ning vaikimisi BLATi väljund on *psl* formaadis (selle näide on toodud Joonisel 12), siis oli vaja see fail konverteerida õiges formaadis failiks.



```
224   8   0   0   0   0   0   0   +   comp1000028_c0_seq1_Sample1   232   0   232   scaffold32772   6961   6282   6514   1   232,   0,   6282,
398   14  0   0   2   3   1   1   -   comp100002_c1_seq1_Sample1    416   0   415   scaffold00104   69729  37919  38332  4   231,17,124,40,  1,233,250,376,  37919,38150,38168,38292,
213   1   0   0   0   0   1   1   -   comp1000039_c0_seq1_Sample1   214   0   214   scaffold11466   14066  10794  11009  2   87,127, 0,87,   10794,10882,
470   4   0   0   1   1   1   121 -   comp100005_c0_seq1_Sample1    487   0   475   scaffold18320   12758  5673   6268   3   241,213,20,     12,254,467,     5673,5914,6248,
332   5   0   0   1   1   2   15  +   comp1000065_c0_seq1_Sample1   350   12  350   scaffold15511   13930  3140   3492   4   8,115,153,61,   12,20,135,289,  3140,3151,3278,3431,
230   0   0   0   0   0   0   0   +   comp1000079_c0_seq1_Sample1   230   0   230   scaffold63226   4527   4140   4370   1   230,   0,   4140,
462   11  0   0   0   0   1   6   +   comp100009_c0_seq1_Sample1    534   61  534   contig764626    1111   0      479    2   284,189,        61,345, 0,290,
391   4   0   0   0   0   1   3   -   comp100010_c0_seq1_Sample1    398   0   395   scaffold20296   9449   6571   6969   2   139,256,        3,142,  6571,6713,
241   4   0   0   2   6   1   4   +   comp1000140_c0_seq1_Sample1   266   7   258   contig1995131   330    81     330    3   48,15,182,      7,56,76,        81,129,148,
513   26  0   0   1   1   1   1   +   comp1000292_c0_seq1_Sample1   541   1   541   contig939903    746    171    711    3   173,171,195,    1,174,346,      171,345,516,
465   6   0   0   1   5   1   1   +   comp100033_c0_seq1_Sample1    477   1   477   contig1160972   565    90     562    2   80,391, 1,86,   90,171,
661   2   0   0   0   0   1   2   +   comp100034_c0_seq1_Sample1    664   1   664   scaffold17717   12294  749    1414   2   405,258,        1,406,  749,1156,
488   7   0   0   3   8   3   7   +   comp100036_c0_seq1_Sample1    614   111 614   contig1094502   597    85     587    6   15,453,8,11,4,4,        111,130,586,595,606,610,
416   0   0   0   0   0   0   0   -   comp1000402_c0_seq1_Sample1   421   0   416   scaffold77559   4189   818    1234   1   416,   5,   818,
281   0   0   0   0   0   0   0   -   comp1000406_c0_seq1_Sample1   281   0   281   contig1966220   334    2      283    1   281,   0,   2,
240   0   0   0   0   0   0   0   +   comp1000437_c0_seq1_Sample1   240   0   240   scaffold64318   4897   2211   2451   1   240,   0,   2211,
274   0   0   0   0   0   0   0   +   comp100043_c1_seq1_Sample1    308   1   275   contig2089715   315    41     315    1   274,   1,   41,
1403  6   0   0   0   0   0   0   -   comp100043_c2_seq1_Sample1    1748  78  1487  scaffold01394   35537  4002   5411   1   1409,  261,  4002,
192   1   0   0   1   2   0   0   +   comp1000455_c0_seq1_Sample1   205   0   195   scaffold06767   19247  11638  11831  2   184,9,  0,186,  11638,11822,
1088  3   0   0   0   0   1   50  +   comp100045_c0_seq1_Sample1    1157  1   1092  scaffold06237   19970  16018  17159  2   411,680,        1,412,  16018,16479,
1474  0   0   0   0   0   2   210 +   comp100045_c1_seq1_Sample1    1484  0   1474  scaffold06237   19970  14300  15984  3   7,18,1449,      0,7,25, 14300,14516,14535,
328   9   0   0   0   0   0   0   +   comp1000483_c0_seq1_Sample1   337   0   337   scaffold03076   30040  7364   7701   1   337,   0,   7364,
239   1   0   0   0   0   0   0   -   comp1000484_c0_seq1_Sample1   240   0   240   contig757362    1136   24     264    1   240,   0,   24,
303   10  0   0   0   0   0   0   -   comp1000487_c0_seq1_Sample1   325   0   313   scaffold05205   23246  19388  19701  1   313,   12,  19388,
972   5   0   0   0   0   1   59  -   comp100048_c1_seq1_Sample1    981   4   981   scaffold00659   42631  30440  31476  2   351,626,        0,351,  30440,30850,
508   0   0   0   0   0   0   0   -   comp100050_c0_seq1_Sample1    525   17  525   scaffold111980  3993   3485   3993   1   508,   0,   3485,
201   4   0   0   0   0   1   1   +   comp1000611_c0_seq1_Sample1   243   37  242   contig2992950   214    0      206    2   135,70, 37,172, 0,136,
183   7   0   0   0   0   1   1256 +   comp1000621_c0_seq1_Sample1   235   45  235   scaffold48421   5962   4133   5579   2   42,148, 45,87,  4133,5431,
208   7   0   0   1   1   0   0   -   comp1000627_c0_seq1_Sample1   223   7   223   contig799740    985    770    985    2   63,152, 0,64,   770,833,
769   8   0   0   0   0   0   0   -   comp100065_c0_seq1_Sample1    839   62  839   scaffold06840   18077  13202  13979  1   777,   0,   13202,
235   0   0   0   0   0   0   0   +   comp1000664_c0_seq1_Sample1   291   18  253   contig2751806   235    0      235    1   235,   18,  0,
940   7   0   0   0   0   1   104 -   comp100067_c0_seq1_Sample1    1079  132 1079  scaffold04707   23475  4444   5495   2   459,488,        0,459,  4444,5007,
237   3   0   0   1   2   0   0   -   comp10006_c0_seq1_Sample1     242   0   242   scaffold01294   36397  26753  26993  2   157,83, 0,159,  26753,26910,
1329  12  0   0   1   2   1   2   -   comp100070_c0_seq1_Sample1    1565  0   1343  scaffold13733   14191  8250   9593   3   265,1058,18,    222,487,1547,   8250,8517,9575,
```

Joonis 12. Fail psl formaadis. Tavaliselt psl failid sisaldavad 21 tulpa. I – *match'*ide arv, mis ei ole kordused; II – *mismatch*'ide arv; III - *match'*ide arv, mis kuuluvad kordusjärjestusse; IV – 'N' aluspaaride arv; V – *insert*'ide arv päringus; VI – päringusse lisatud aluspaaride arv; VII – *insert*'ide arv sihtmärkjärjestuses; VIII - sihtmärkjärjestusse lisatud aluspaaride arv; IX- päringu ahel (+/- ehk *forward/reverse*); X – päringu nimetus; XI – päringu järjestuse pikkus; XII – päringu järjestuse algpunkt/algkoordinaat; XIII – päringu järjestuse lõppkoordinaat; XIV – sihtmärkjärjestuse nimetus; XV – sihtmärkjärjestuse pikkus; XVI – sihtmärkjärjestuse algkoordinaat; XVII – sihtmärkjärjestuse lõppkoordinaat; XVIII – blokkide arv järjestuses; XIX – komaga eraldatud blokkide suurused; XX – komaga eraldatud päringu blokkide alguspositsioonid; XXI – komaga eraldatud sihtmärkjärjestuse blokkide alguspositsioonid.

Enne failiformaadi muutmist tuli andmed sorteerida. Selleks kasutati AUGUSTUSega kaasas oleva skripti ***filterPSL.pl***, sisestades käsureale:

```
cat hints.psl | filterPSL.pl --best --minCover=80 > hints.f.psl
```

Käsk *cat* avab ette antud faili *hints.psl*, toru (*pipe* = |) käsk annab avatud faili edasi skriptile ***filterPSL.pl***, mis filtreerib välja need tulemused *(--best),* mis vastavad identsuse minimaalsele protsendile (vaikimisi 92) ja minimaalse lugemi katvuse protsendile (*--minCover=80*). Programmi väljund suunatakse etteantud faili *hints.f.psl*. Filtreeritud tulemuste konverteerimiseks õigesse formaati kasutati AUGUSTUSega kaasas olevat skripti ***blat2hints.pl*** , sisestades käsureale:

```
blat2hints.pl --nomult --in=hints.f.psl –out=hints.f.gff
```

Siin *--nomult* tähendab, et kui on mitu identset introni „vihjet“ siis neid ei summeerita üheks, *--in* parameeter määrab sisendfaili, ning *--out* parameeter määrab väljundfaili. Valmis „vihjete“ faili kasutati edaspidi lisaandmestikuna geenide ennustamiseks. Vihjete kasutamiseks käivitati AUGUSTUS käsuga:

```
augustus --species=conus_consors genome.fa --extrinsicCfgFile=extrinsic.conus_consors.cfg
                --hintsfile=hints.f.gff > augustus.hints.gff
```

Väljundiks oli fail *augustus.hints.gff*, milles igale ennustatud geenikomponendile oli lisatud juurde

kommentaar sellest, missugused „vihjed“ olid sobivad ja missugused mitte.

4. AutoAug.pl

Augustusega oli kaasas skript ***AutoAug.pl***, mis automaatselt treenib AUGUSTUSt ja ennustab geenistruktuure etteantud eukarüootsel genoomil, kasutades olemasolevaid cDNA tõendeid. See protsess koosneb mitmest etapist:

- Konstrueeritakse algne geenide treeningfail, kasutatdes programmi PASA väljundfaili.

- Treenitakse AUGUSTUSt ennustama kodeerivaid regioone (ilma *UTR'*ide ennustamiseta), kasutadest algset treeningandmestikku.

- Treenitakse UTR mudelit, kasutades *EST*'ide poolt toetatud geene ja *EST* joondusi.

- Ennustatakse geene koos UTR'idega uuritaval genoomil kasutades cDNA „vihjeid“.

UTR ennustuste kasutamiseks oli vaja lisaks programme PASA (versioon r2012-06-25) ja GMAP (versioon v3 20.07.2012).

4.1. PASA ja GMAP

Nagu ülalpool mainitud, on programm PASA vajalik treeningandmestiku koostamiseks, ning selleks, et joondada transkriptoom genoomile. Akronüüm „PASA“ tähendab splaisitud joonduste kokkupanemise programmi *(Program to Assemble Spliced Alignments)*(http://pasa.sourceforge.net/#A_rcdaap). Programm on mõeldud eükarüootse genoomi annoteerimiseks, ning kasutab ekspresseeritud transkriptide järjestuste splaisitud joondusi selleks, et modelleerida automaatselt geenistruktuuri ja salvestada annotatsioon, mis on kooskõlas eksperimentaalse järjestuse andmetega. Samuti identifitseerib ja klassifitseerib PASA kõiki splaisinguvariandid, mida kinnitavad transkriptide joondused.

Programm sisaldab järgmisi funktsioone:

- Täis- ja osalise geenistruktuuri modelleerimine, mis põhineb kokkupandud splaisitud joondustel
- Automaatset transkriptide joondustel põhineva geenistruktuuri juba olemasolevatesse geeniannotatsioonidesse lisamist. Annotatsioonide uuendamine sisaldab endas UTR'ide annotatsiooni, eksonite lisamist/eemaldamist/piiride korregeerimist, alternatiivsete splaisingvariantide mudelite juurde lisamist, geenide liitmist ja eraldamist ning uute geenide modelleerimist.

- Polüadenülatsiooni saitide kaardistamist genoomile
- Antisense transkriptide identifitseerimist
- Kõikide leitud splaisingvariatsioonide identifitseerimist ja klassifitseerimist
- *Ab initio* programmide treenimiseks mõeldud osaliste või täispikkade valku kodeerivate geenide leidmist transkriptide joonduste põhjal

Transkriptide kaardistamiseks ja joondamiseks genoomile kasutab PASA programmi GMAP (http://research-pub.gene.com/gmap/). GMAP on mõeldud mRNA ja EST järjestuste kaardistamiseks ja joondamiseks genoomile. Analüüsitakse ainult ühte parimat joondust iga transkripti kohta. Kaardistamise all on mõeldud, et kui on antud cDNA, siis programm leiab koha genoomil kuhu see kõige paremini joondub. Joondamine tagab antud cDNA eksonite vastavuse genoomi segmendiga nukleotiidse järjestuse tasemel.

Samuti GMAP'i kõrval PASA vajab kahte MySQL andmebaasi, kuhu salvestatakse liigispetsiifiline informatsioon. MySQL andmebaaside loomine nõuab, et programm käivitataks administraatori privileegidega. Selleks, et teha seda tavakasutajana on vaja teha skriptidesse muudatusi.

Treenimine käivitati käsuga:

```
./AUGUSTUS/scripts/autoAug.pl -g genome.fa --species=conus_consors -c transcriptome.fa -v -v
                                  --pasa --utr –useexisting
```

Kus argument -g määrab genoomifaili; --species liigi nimetuse (AUGUSTUSe parameetrite kausta); -c transkriptoomi faili; -v suurendab programmi väljastatava informatsiooni hulka (*verbosity*); --pasa määrab, et kasutataks programmi PASA; --utr määrab, et treenitaks ning ennustaks ka *UTR*'e; --useexisting määrab, et kasutataks juba olemasolevat kausta nimetatud liigi parameetritega/muude failidega. Hiljem läheb PASA poolt loodud treeningfaili vaja selleks, et tekitada ja treenida AUGUSTUSe UTR mudelit.

4.2. UTR mudeli tekitamine

„Vihjed" pole ainuke viis, kuidas AUGUSTUSt täpsemaks teha. Selleks on võimalik kasutada veel treenitud UTR mudelit (http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.UTRTraining). UTR'ide kasutamise jaoks oli vaja luua eraldi skriptiga ***gff2gbSmallDNA.pl*** treeningandmestik (*utr.gb.train*) kasutades sisendfailina PASA poolt tekitatud gff formaadis faili. Saadud fail *genbank* formaadis

jagatakse skripti ***randomSplit.pl*** abil kaheks andmestikuks.

Treeningandmestikku kasutati algseks treenimiseks ***etrainig.pl*** skriptiga:

```
etraining --species=conus_consors utr.gb.train
```

Parameetrite optimeerimiseks kasutati ***optimize_augustus.pl*** skripti, sisestades käsurealt:

```
optimize_augustus.pl --species=conus_consors --rounds=3 utr.gb.train --UTR=on
–metapars=/AUGUSTUS/config/species/conus_consors/conus_consors_metapars.utr.cfg
–trainOnlyUtr=1
```

Kus argument --rounds määrab mitu optimeerimisringi skript teostab, --UTR=on määrab, et UTR mudel on sisse lülitatud, --trainOnlyUtr=1 tähendab, et treenitakse ainult UTR mudeleid (nt. kui teised mudelid on juba treenitud). Pärast optimeerimise lõpetamist käivitati jälle ***etraining.pl*** ning kui kõik mudelid olid valmis, siis käivitati uus ennustus sisestades käsureale:

```
augustus --species=conus_consors genome.fa --UTR=on
--extrinsicCfgFile=extrinsic.conus_consors.cfg --hintsfile=hints.f.gff >
conus_consors_with_hints_and_utr.gff
```

Tulemuseks on gff formaadis fail (Joonis 13.), kus olid kirjas transkriptsiooni start ja lõpp saidid,

```
# start gene g4
scaffold00001   AUGUSTUS        gene    99686   105138  0.03    +       .       g4
scaffold00001   AUGUSTUS        transcript      99686   105138  0.03    +       .       g4.t1
scaffold00001   AUGUSTUS        tss     99686   99686   .       +       .       transcript_id "g4.t1"; gene_id "g4";
scaffold00001   AUGUSTUS        exon    99686   100530  .       +       .       transcript_id "g4.t1"; gene_id "g4";
scaffold00001   AUGUSTUS        exon    102821  105138  .       +       .       transcript_id "g4.t1"; gene_id "g4";
scaffold00001   AUGUSTUS        start_codon     104253  104255  .       +       0       transcript_id "g4.t1"; gene_
scaffold00001   AUGUSTUS        CDS     104253  104585  0.69    +       0       transcript_id "g4.t1"; gene_id "g4";
scaffold00001   AUGUSTUS        stop_codon      104586  104588  .       +       0       transcript_id "g4.t1"; gene_
scaffold00001   AUGUSTUS        tts     105138  105138  .       +       .       transcript_id "g4.t1"; gene_id "g4";
# protein sequence = [MLCSCLSTLFLVLSSSSRCPQLTCPGPPSPDKQFSILCSSLGLIPHFSNSSHKRVCGMGQMVLECLPMEMLVTSSNHR
# WGGIFSAGKPLLSSGEWVTDSDQERLKVAEGEG]
# Evidence for and against this transcript:
# % of transcript supported by hints (any source): 80
# CDS exons: 1/1
#       E:   1
# CDS introns: 0/0
# 5'UTR exons and introns: 2/3
#       E:   2
# 3'UTR exons and introns: 1/1
#       E:   1
# hint groups fully obeyed: 10
#       E:  10 (comp1650958_c0_seq1_Sample2,comp1041858_c0_seq1_Sample2,comp776420_c0_seq1_Sample8,...)
# incompatible hint groups: 2
#       E:   2 (comp951742_c0_seq1_Sample8,comp843905_c0_seq1_Sample11)
# end gene g4
###
```

Joonis 13. Treenitud AUGUSTUSe eennustus koos *UTR*'idega, kus kasutati „vihjete“ faili. Esimene tulp sisaldab *scaffold*'i nimetust, II tulp – programmi nimetust, III – detail (geen/exon/intron/CDS-*coding DNA sequence*/tss-*transcription start site*/tts – *transcription terminal site*), IV – alguspunkt/algkoordinaat, V-lõppunkt/lõppkoordinaat, VI – skoor, VII – ahel(+/-, '.'- ahel ei ole oluline), VIII – raam(0 või '.' tähendab, et regioon on raami sees), IX – omadused. Kommentaarides on valgu järjestus, ning natuke informatsiooni kui suur osa transkriptist on „vihjete“ poolt tõestatud(kaetud),mitu CDS'i kokku leitud, mitu UTR'i on leitud, mitu „vihjegruppi“ olid sobilikud ja mittesobilikud.

eksonid, intronid ja kodeerivad/mittekodeerivad regioonid.
Selleks, et hinnata liigispetsiifiliseks treenitud AUGUSTUSE ennustuse täpsust võrreldes teise liigi kvaliteetse mudeli abil tehtud ennustusega kasutati AUGUSTUSega kaasas olevat inimese geenide mudelit. Inimparameetrite treeningandmestik (http://augustus.gobics.de/datasets/) sisaldas 1284 geeni GenBank (http://www.ncbi.nlm.nih.gov/genbank/) andmebaasist ja 11739 splaisingsaiti (Guigó, R., jt 2000). Meie ennustasime inimparameetritega AUGUSTUS'e abil geene nii *Conus consors*'il kui ka võrdlusorgamismidel *Aplysia californica* (http://www.broadinstitute.org/ftp/pub/assemblies/invertebrates/aplysia/), *Pinctada fucata* (http://marinegenomics.oist.jp/pinctada_fucata), *Crassostera gigas* (http://www.ncbi.nlm.nih.gov/sra/?term=Crassostrea+gigas), *Lottia gigantea* (http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html) ja inimese (ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests) genoomil.

5. AUGUSTUSe ennustuse kvaliteedi ja eksonite pikkuste hindamine.

Selleks, et hinnata AUGUSTUSe ennustuse kvaliteeti loodi eraldi *fasta* formaadis fail, mis koosnes kõikidest valgulistest järjestustest AUGUSTUSe väljundist. Seejärel leiti kõigi selliste järjestuste homoloogiad UniProtKB/Swiss-Prot andmebaasiga kasutades programmi BLASTP:

```
  blastp -db /uniprot/uniprot_sprot.fasta -query conus_consors_prot.fasta -evalue 0.000001
-num_threads 2 -outfmt "6 qseqid qlen sseqid slen pident length mismatch gapopen qstart qend
                sstart send evalue bitscore" > conus_consors_vs_sprot.txt
```

Siin *blastp* – on BLAST, mis joondab ühe valgujärjestuse teise vastu; *db* määrab kasutatava andmebaasi; *query* määrab otsitavate valgujärjestuste faili; *evalue* määrab joonduse statistilise olulisuse nivoo; *num_threads* on kasutatavate protsessorite arv; *outfmt* määrab väljundfaili formaadi. Selle järel on jutumärkides antud, mis väljad ja mis järjekorras kirjutatakse faili.
Lisaks uuriti kui palju valguhomoloogia meetodiga genereeritud oletatavatest eksonitest langevad kokku AUGUSTUSe ennustatutega. Selleks kirjutati programmi, mis võrdleb omavahel kahe ennustuse kattuvaid regioone. Kattuvuse täpsuseks nõuti ±20 nukleotiidi.
Selleks, et hinnata eksonite pikkusi võrreldi omavahel AUGUSTUSe ennustatud eksonite pikkusi ja valguhomoloogia meetodiga leitud eksonite pikkusi. Selleks leiti kattuvad regioonid lubades ±20 aluspaari suurust nihet ja arvutati, kui palju keskmiselt olid ühed eksonid teistest pikemad või lühemad.

III. Tulemused

1. Valguhomoloogia järgi leitud eksonid

Valguhomoloogia metoodika järgi leiti *Conus consors* 3024,7 Mb genoomist vasted 9383 unikaalsele valgufunktsioonile. Keskmine valgu katvus (leitud vastete pikkuste summa suhe terve valgu pikkusesse) oli 46%, mediaankatvus 44%. Kokku leiti 30453 oletatavat eksonit. Eksoni minimaalne pikkus oli 45 nukleotiidi, maksimaalne 11166 nukleotiidi, keskmine 215 ja mediaan 165 nukleotiidi. Kokku tuvastati 4221 oletatavat intronit. Minimaalne introni pikkus oli 27, maksimaalne 39298, keskmine 1676 ja mediaan 1266 nukleotiidi.

Sama meetodiga uuriti peale *Conus consors'i* ka teiste molluskite - *Aplysia californica, Pinctada fucata, Crassostera gigas* ja *Lottia gigantea* genoome. Tabelis 1 on toodud valguhomoloogia põhjal ennustatud eksonite ja intronite koguhulgad erinevate molluskite genoomides, eksonite ja intronite pikkuste mediaanid, leitud unikaalsete valgufunktsioonide hulgad ning ennustatud eksonite keskmine ja mediaanarv kogu valgu kohta. Eksonite keskmine ja mediaanarv valgu kohta arvutati välja kasutades valgu katvust. Kuna enamikul juhtudel katvus ei ole 100%, siis on ilmselt leitud ainult osa eksonitest. Eksonite tegelik arv peab vastama 100% katvusele. Selleks, et leida tegelik eksonite hulk funktsiooni kohta, kasutati valemit: $\frac{X \times 100}{Y}$, kus X on leitud eksonite hulk valgu kohta, Y – leitud valgu katvus, 100 – 100% valgu katvus.

Tabel 1. Valguhomoloogia meetodil leitud eksonite ja intronite koguhulgad erinevates molluskites, välja arvutatud eksonite ja intronite pikkuste mediaanid, unikaalsete funktsioonide koguhulgad genoomi kohta ning keskmised ja mediaanid eksonite hulgast funktsiooni kohta.

Liigid	**Kokku eksoneid**	**Kokku intoneid**	**Mediaan eksoni pikkus (bp)**	**Mediaan introni pikkus (bp)**	**Kokku funktsioone**	**Keskmiselt eksoneid funktsiooni kohta**	**Mediaan eksoneid funktsiooni kohta**	**Genoomi suurus (Mb)**
P.fucata	25288	4000	161	489	7450	9.7	7	1150
C.gigas	14482	5624	182	331	4445	8	6	559
L.gigantea	5031	1654	197	391	1882	8.45	5	359.5
A.californica	16703	5555	164	1006	5525	10.1	8	1800
C.consors	30453	4221	164	1265	9383	8.47	6	3024.7

Nagu on näha tabelist 1, on kõige rohkem eksoneid leitud teol *Conus consors*. Teisel kohal on karp

Pinctada fucata. Leitud eksonite ja intronite hulgad molluskitel on väga varieeruvad. Kõige rohkem oletatavaid introneid leiti karbil *Crassostera gigas.* Eksonite mediaanpikkused on kõigil molluskitel enam-vähem sarnased, maksimaalne erinevus nende vahel on 33 bp. Samas on intronite mediaanpikkused tigudel märgatavalt suuremad kui karpidel, erandiks *Lottia gigantea.* Võimalik, et intronite väiksem mediaanpikkus on *Lottia* puhul seotud sellega, et tema genoom on teistest molluskitest väiksem. Leitud valgufunktsioonide hulk sõltub genoomi suurusest ja erineb liigiti. Eksonite keskmine hulk valgu kohta varieerub 8 – 10 ja mediaan 5 – 8 vahel. Leitud eksonite keskmist hulka valgu kohta kasutati hiljem geenide arvu hindamisel.

2. AUGUSTUSe ennustused

Kasutades *Conus consors*'i liigispetsiifilisi treenitud parameetreid (koos „vihjete“ ja UTR mudeliga), ennustas AUGUSTUS *C.consors*'i genoomist 120076 geeni ja 214847 valku kodeerivat järjestust (CDS). Viimast võib lugeda kodeerivate eksonite hulgaks. Kuna genoom oli väga fragmenteeritud, siis ei saa AUGUSTUSe ennustatud geenide hulka tõeseks lugeda. AUGUSTUSel ei ole informatsiooni, mille põhjal saaks liita kokku eri kontiigides paiknevad sama geeni fragmendid. Samas võib pidada tõeseks kodeerivate regioonide koguhulka, sest see ei sõltu olulisel määral genoomi fragmenteerituse astmest eeldusel, et fragmentide katkevuskohad ei asu CDS'ide sees.

Selleks, et hinnata uuritava liigi geenide koguhulka, kasutati AUGUSTUSe ennustatud kodeerivate regioonide hulka ja valguhomoloogia meetodi abil leitud keskmist eksonite hulka valgu kohta. Jagades need omavahel läbi, saadi ligikaudne hinnang unikaalsete valkude, seega geenide, arvule. Selle metoodika kohaselt on liigil *Conus Consors* umbes 25000 geeni.

Nagu üldse AUGUSTUSe ennustused, sõltus ennustatud geenide koguhulk tugevasti sellest, kas kasutati “vihjete” faili ja liigispetsiifiliselt treenitud UTR mudelit.

Tabelis 2 on näidatud, kuidas sõltub AUGUSTUSe ennustatud kodeerivate regioonide koguhulk ja sellest tuletatud hinnang geenide arvule sellest, milliseid geeniennustust täpsustavaid lisaandmeid kasutati.

Tabel 2. *Conus consors* geenide hulk liigispetsiifiliste parameetritega treenitud AUGUSTUSe väljundist, kus lisati juurde erinevaid täpsust soosivaid andmeid („vihjed“, *UTR*'id).

	Ilma „vihjedeta“ ja *UTR*'ideta	„vihjetega“ ilma *UTR*'ideta	„vihjete“ ja *UTR*'idega
Kogu eksonite hulk	390987	326612	214847
Geenide arv(keskmine = 8.47 eksonit funktsiooni kohta)	46161	38561	25365
Geenide arv (mediaan = 6 eksonit funktsiooni kohta)	65164.5	54435	35807.8

AUGUSTUSel on tendents ennustada valepositiivseid eksoneid. Kui lisada täpsust tõstvaid andmeid siis valepositiivsete eksonite hulk langeb. Koos eksonite arvu langusega langeb ka oletatavate geenide arv.

Lisaks ennustati võrdluseks AUGUSTUSe abil geenid teiste molluskite genoomidest , kasutades eraldi *Conus consors*'i mudelit ja inimese geenimudelit. Inimese geenimudel valiti seetõttu, et kuigi ta on fülogeneetiliselt kauge, on inimese geenid kõige põhjalikumalt uuritud ja seega mudel kõige täpsem. Huvi pakkus eelkõige küsimus, kas inimese geenimudel ennustab molluski genoomist geene edukamalt või vähemedukalt, kui sama või mõne teise molluski teadaolevatel geenidel treenitud mudel. Ehk siis kas mudeli kvaliteet sõltub rohkem lähteandmete heast kvaliteedist (inimmudel) või fülogeneetiliselt võimalikult sarnase treengandmestiku kasutamisest (*C. Consors*'i mudel). Tabelis 3 on näidatud kahe mudeliga ennustatud eksonite hulk kõigis molluskites.

Tabel 3. Kahe mudeliga ennustatud eksonite hulk erinevatel molluskitel. Paremaks võrdluseks on juurde lisatud uuritav liik *C.consors* . Mudelid ei kasutanud täpsust tõstvaid andmeid.

Liigid	***Conus consors* mudel**	**Inimese mudel**
P.fucata	27394	64057
C.gigas	69295	108692
L.gigantea	9267	34762
A.californica	163959	183433
C.consors	390987	352532

Inimudeliga ennustatud molluskite eksonite kaudu arvutati välja ka molluskite geenide oletatavad hulgad. Tulemused on toodud tabelis 4.

Tabel 4. Liikide inimese ja *Conus consors* parameetritega ennustatud eksonite hulgad jagatud kas keskmise või mediaan eksonite hulgaga funktsiooni kohta. [* - (http://genome.ucsc.edu/cgi-bin/hgGateway?org=Sea+hare&db=aplCal1&hgsid=154586973)]

Liigid	**Inimmudeli eksonite hulk jagatud keskmisega**	**Inimmudeli eksonite hulk jagatud mediaaniga**	***C.consors* mudeli eksonite hulk jagatud keskmisega**	***C.consors* mudeli eksonite hulk jagatud mediaaniga**	**Artiklitest võetud oletatavate geenide hulgad**
P.fucata	6604	9151	2824	3913	23257
C.gigas	13587	18115	8662	11549	28027
L.gigantea	4114	6952	1097	1853	23851
A.californica	18162	22929	16234	20495	CDS = 186384[*] ca 18453
C.consors	41621	58755	46161	65165	-

3. AUGUSTUSe ennustuse kvaliteedi hinnang

BLASTX abil leiti joondused AUGUSTUSe ennustatud geenide valguliste järjestuste ja UniProtKB/Swiss-Prot andmebaasi valkude vahel. Kokku leiti 28757 vastavust „vihjete“ ja UTR mudeliga ennustatud järjestuste ning 27116 „vihjetega“ aga ilma UTR mudelita ennustatud järjestuste jaoks. Neile järjestustele vastas Swissproti unikaalseid funktsioone 7832 („vihjete“ ja UTR mudeliga) ning 7066 („vihjetega“ aga ilma UTR mudelita). See tõestab veelkord, et täpsust tõstvad andmed parandavad programmi ennustustulemust. Võrreldes omavahel AUGUSTUSe eksoneid ja valguhomoloogia meetodil leitud oletatavaid eksoneid, leiti kokku 12210 kokkulangevat eksonit. Valguhomoloogia meetodil oli üldse kokku leitud 30453 eksonit, seega ei ole see sugugi halb tulemus, arvestades genoomi fragmenteeritust.

Nende kokkulangevate eksonite pikkuste erinevusi hinnates leiti, et keskmine erinevus AUGUSTUSe ennustatud ja valguhomoloogia põhjal leitud eksonite pikkuste vahel oli *ca* 10%.

IV. Arutelu

1. II põlvkonna sekveneerimise probleemid

Teise põlvkonna sekveneerimine on endaga kaasa toonud palju uusi võimalusi, aga ka uusi probleeme. Võimalikuks on osutanud liikide hulgaline sekveneerimine odava hinnaga. See avas paljude laborite jaoks võimaluse sooritada uuringuid ja katseid palju efektiivsemalt. Nagu paljudel teistel puhkudel on paraku ka siin hind ja kvaliteet antikorrelatsioonis. Teise põlvkonna sekveneerimistehnoloogiate lugemid on võrreldes esimese põlvkonna lugemitega lühikesed ning madala kvaliteediga. Kasutatavad veakorrektsiooni algoritmid võimaldavad sekveneerimisvigade hulka vähendada, kuid nende efektiivsusel on piirid. Madala kvaliteediga lugemite tõttu on genoomi lähteandmetest kokkupanek keerukas ja ka saadav genoomne järjestus tavaliselt fragmenteeritud. Lisaks on uudsete liikide värskelt sekveneeritud genoomid sageli madala katvusega. Lähteandmete madalast kvaliteedist tekkivaid probleeme võimendavad veel kordusjärjestused ja polümorfismid, mida eriti rohkelt leidub eukarüootide genoomides. Algandmete kvaliteedi parandamiseks on mitu võimalust, nagu näiteks sekveneerimisandmete kombineerimine, katvuse tõstmine, referentsgenoomi kasutamine jne, kuid probleemi universaalset lahendust ei ole olemas. Esimest korda sekveneeritud uudsete liikide genoomi kokku panemisel on sekveneerimisandmete kvaliteet eriti määrava tähtsusega, sest nendel puudub referentsgenoom ja tavaliselt ka täpsem informatsioon genoomi ülesehituse kohta. Sageli ei ole võimalik tervet genoomi järjestust täielikult kokku panna, ning see jääb fragmenteeritud kujule. Juhul kui fragmendid on lühikesed pole osasid analüüse, näiteks geenistruktuuri määramist, võimalik teha.
Näiteks oli vaöguhomoloogia meetodil leitud intronite hulk eksonite hulgast umbes 7 korda väiksem. Sellest võib võib järeldada, et enamik assembleeritud genoomi fragmente sisaldavad ainult mõnda eksonit tervest valgust. Seega ei saa homoloogial põhinevatmeetodit kasutada usaldusväärsete intronregioonide leidmiseks.

2. AUGUSTUSe mudelite treeningu analüüs – kas ja kuidas töötab

Uute sekveneeritud liikide jaoks puudub enamasti informatsioon nende geenide arvu ja struktuuri kohta. See tuleb igal üksikul juhul leida genoomsest järjestusest, kasutades lisaks kõikvõimalikku täiendavat informatsiooni, eelkõige transkriptoomi ja varem uuritud valkude järjestusi. Paljude uudsete liikide genoomide annoteerimise keerulisus seisneb veel selles, et tihtipeale puudub täpsem informatsioon ka sarnaste liikide kohta. Enamik geenidest on tundmatud ja seni uurimata ning geenide ülesleidmine on raskendatud genoomi fragmenteerituse tõttu. Selleks, et geenide annoteerimist

lihtsustada on välja töötatud automaatsed geeniennustusprogrammid. Kuna enamiku liikide jaoks ilmselt kehtivad eukarüootide geenistruktuuri üldised reeglid, siis on võimalik nende reeglite alusel ennustada geene ka uudsete liikide genoomides, kasutades erinevaid algoritme ja mudeleid. Samas on igal liigil mingid geneetilised eripärad ja sellepärast tuleks kasutatavaid mudeleid mingil määral modifitseerida või võtta kasutusele uued mudelid, et tõsta ennustuse täpsust. Eriti oluline on see liigispetsiifiliste geenide ennustamisel, mille jaoks pole võimalik kasutada homoloogial põhinevat otsingut teada olevate valkude andmebaasist. Selleks, et ennustusalgoritm saaks võtta arvesse liikide spetsiifilisi eripärasid tuleb selle parameetreid muuta, treenides programmi mingil sõltumatul viisil leitud sama liigi geenide mudelitega. Tänapäeva geeniennustusprogrammid enamjaolt omavad sellist liigispetsiifilist treenimisvõimalust, see kehtib ka meie poolt kasutatud programmi AUGUSTUS puhul.

Liigispetsiifilise treenimise käigus võtab AUGUSTUS arvesse järjestuste GC sisaldust, koodoneid ja 1-5 aluspaariste sõnade keskmisi sagedusi geeni eri osades. Treenimiseks on kõige olulisemad kodeerivad geeniosad ning lisaks lühike järjestus geeni algusest ülesvoolu. Oluline on ka see, et geeni kodeeriva osa algus ja lõpp oleksid võimalikult täpselt annoteeritud. Ette antud geenijärjestuste kogum ei tohiks sisaldada redundantseid geene ja kahe geeni sarnasus ei tohi ületada 70%. Seda on vaja selleks, et vältida ületreenimist spetsiifilistele järjestustele mis alandab uudsete järjestuste tuvastamise võimet. AUGUSTUS langetab otsuse eksoni, introni või geeni olemasolust/puudumisest, võttes arvesse erinevate pikkustega aluspaaride mustrite esinemissagedusi. Nende sageduste oodatavad väärtused on programmi poolt salvestatud treeningandmestiku põhjal. Seega mida täpsem ja õigem on koostatud treeningandmestik, seda täpsemad on mustrite tõenäosused ning seda õigemaid tulemusi programm väljastab.

AUGUSTUS võimaldab mudelit treenida kahes etapis. Esimene treening on mudeli parameetrite seadistamine olemasolevate lähteandmete pealt. Sellele järgneb parameetrite optimeerimine. Optimeerimise käigus tehakse korduvalt geeniennustusi lähteandmetelt, muutes iga kord automaatselt ühte programmi parameetrit. Igas sammus salvestatakse parameetri optimaalne väärtus ning hakatakse seejärel optimeerima järgmist parameetrit.

Liigispetsiifiliselt treenitud AUGUSTUS peaks olema võimeline leidma üles nii teada olevad varem annoteeritud geenid kui ka ennustama uudseid liigispetsiifilisi geene. Tegelikkuses jääb palju geene ennustamata/leidmata (valenegatiivsed), või leitakse/ennustatakse geene seal kus neid ei ole (valepositiivsed). Uudsete liikide treeningandmestiku koostamine võib osutuda keeruliseks

geeniinformatsiooni puudulikkuse tõttu. Siin võiks abiks olla informatsioon lähedaste liikide kohta. Mida lähedasem on treeningandmestiku koostamiseks kasutatud liik, seda suurem on tõenäosus, et selle liigi andmed kõlbavad uuritava liigi geeniennustamismudeli treenimiseks.

Samas on näha, et *Conus consors* mudel ennustab teistel liikidel oluliselt väiksema arvu kodeerivaid regioone kui inimese mudel. Samuti on hinnatud geenide arv teistel molluskitel mõlema mudeli puhul väiksem kirjanduses toodud oletatavast geenide arvust. Et inimmudeliga ennustatud geenide arv on lähedasem kirjanduses toodule võime väita, et hea kvaliteediga inimmudel töötab uute organismide puhul geenide ennustamiseks paremini, kui madala kvaliteediga fülogeneetiliselt lähedasema orgamismi järjestustel treenitud mudel. Ilmselt on oluliseks kriteeriumiks ka organismide genoomide sarnasus. Molluskid on väga heterogeenne ja fülogeneetiliselt vana hõimkond, seega erinevad nende genoomid olulisel määral. On tõenäoline, et sama sugukonna või perekonna piires töötab teise liigi geenimudel uue liigi geenide ennustamiseks paremini, kui näiteks inimese mudel.

AUGUSTUSe programmiga on mõnede liikide geenimudelid vaikimisi kaasas. Juhul kui uuritav liik on mõnele neist lähedane, tuleks kasutada seda mudelit. Programmis olevad mudelid on koostatud tuntud mudelorganismide (inimene, äädikakärbes) kvaliteetsetest andmetest ja on kontrollitud, seega annavad ka paremaid tulemusi. Kui aga uuritav liik erineb märgatavalt mudelorganismist, tuleks kindlasti treenida uuritava liigi jaoks spetsiifiline mudel, mitte kasutada lähedase liigi mudelit. Valmis mudelit tuleks siis kasutada ainult sellisel juhul, kui liigispetsiifilist mudelit ei õnnestu koostada, näiteks puuduliku genoomi tõttu.

Antud töös selgus, et kuigi kvaliteetsema geenimudeli (inimese) kasutamine molluski geenide ennustamiseks annab parema tulemuse, kui teise molluski geenimudeli kasutamine, ei suuda kumbki mudel leida üles suuremat osa geenidest (lähtudes kirjanduses toodud geenide arvust). Seevastu samal liigil treenitud mudeli kasutamine tema geenide ennustamiseks annab ilmselt parema tulemuse, kui kvaliteetse aga fülogeneetiliselt kauge inimmudeli kasutamine.

3. Vihjete (*hints*) kasutamine ennustuse täpsuse tõstmiseks

Kui kasutada geenide ennustamiseks ainult liigispetsiifiliselt treenitud mudelit on ennustus küllaltki vigane sisaldades palju valepositiivseid eksoneid ja geene. Ennustuse parandamiseks võimaldab AUGUSTUS kasutada välist informatsiooni, näiteks transkriptoomi järjestusi. See võimaldab oluliselt tõsta programmi täpsust ja spetsiifilisust. Sellist informatsiooni nimetatakse „vihjeteks“ (*hints*).

„Vihjeid“ saab tekitada näiteks kaardistades sama või lähedase liigi transkriptoomi genoomile, soovitatavalt kasutades programmi BLAT. „Vihjete“ olemasolul võtab AUGUSTUS need arvesse hinnangu andmisel enustatud eksoni usaldusväärsuse kohta ning elimineerib suure osa valepositiivseid ja valenegatiivseid geene. Samuti paneb ta kokku rohkem õigeid geenimudeleid ja eksonite/intronite piirid hinnatakse täpsemalt.

Lisaks on AUGUSTUSe ennustust võimalik muuta veelgi täpsemaks, pannes teda geenimudeli koostamisel arvestama mittetransleeritavaid mRNA otsi (UTR'e). Selleks, et programm ennustaks UTR'e tuleb eraldi treenida vastav mudel. Mudel treenitakse PASA programmi poolt väljastatud andmestiku alusel, mis sisaldab mRNA piire. Võttes arvesse CDS piire ja mRNA piire tekitab AUGUSTUS UTR'ide mudeli, mille treenimine ja optimeerimine käib nagu ka teiste mudelite treenimine.

Kui panna kokku kõik liigispetsiifilised mudelid, sealhulgas ka UTR mudel ja lisada juurde „vihjete“ fail, muutub AUGUSTUSe ennustus küllaltki täpseks ning seda võib pidada usaldusväärseks. Ideaalis tuleks kõik programmi poolt ennustatud geenid ükshaaval läbi vaadata, aga see on väga mahukas ja aeganõudev töö. Kiiremalt saab ennustuse täpsust ja korrektsust hinnata võrreldes ennustatud geenimudelitele vastavaid valke teada olevate valkude andmebaasidega. See meetod võimaldab ka ühendada osalisi geenimudeleid, mida AUGUSTUS ei ole suutnud kokku panna sest geeni eri osad asuvad eri kontiigidel. Kuna aga võrdlus olemasolevate andmebaasidega ei anna vastust liigispetsiifiliste geenide kohta, tuleb nende kontrollimiseks ikkagi geenid ükshaaval läbi vaadata. Kumba kahest teest valida, sõltub ainult püstitatud eesmärkidest. Samas võib loota, et eri organismide geenide arv andmebaasides kasvab tulevikus kiirest ja nii on aina suuremat osa geenidest võimalik võrrelda juba olemasolevatega.

4. *C. consors*'i geenistruktuur

Tigudel on eukarüootidele tüüpiline geenistruktuur, sisaldades eksoneid, introneid, start ja stop koodoneid ja 5' ning 3' UTR'i. Tigude täielikke genoome on vähe sekveneeritud ja seega vähe uuritud.

Kuigi meie töörühmas kokku pandud *C.consors* genoom oli küllaltki fragmenteeritud, saab selle põhjal siiski *C.consors*'i geenistruktuuri kohta mõndagi teada. Uuritava teoliigi genoomi suurus on 3 Gb. Valguhomoloogia metoodiga tuvastati, et geenid sisaldavad keskmiselt 8.47 eksonit. *C.consors* eksonite arvu mediaan on 6, eksoni pikkus 164 aluspaari ja introni pikkus on 1265 aluspaari. Need tulemused võivad olla kallutatud seetõttu, et UniProt/SwissProt valkude andmebaasis on eelistatult

pikemate valkude järjestused ja vastavalt leiti ka homoloogilised alad eelistatult pikematele valkudele. Kombineeritud meetoditega hinnati *C.consors*'i geenide koguhulk, milleks on 25000 geeni. Võrdluseks on karbi *P.fucata* genoomi suurus 1150 Mb ja valguhomoloogia meetodiga leiti keskmiselt 9.7 eksonit geeni kohta. Samas oli kirjanduse andmetel hinnatud *P.fucata* eksonite arvuks geeni kohta 3.2. Meie hinnangul on kirjanduses toodud eksonite arv tugevalt alahinnatud. Eksonite arvu mediaan valgufunktsiooni kohta on 7, eksoni pikkuse mediaan 161 ja introni pikkus 489 aluspaari. Kirjanduse andmetel on pärlikarbil *ca* 23000 geeni. Nii inimese kui *C.consors*'i geenimudeli kasutamine võimaldas tuvastada ainult väikese osa neist geenidest. Karbi *C.gigas* genoomi suurus on 559 Mb. Valguhomoloogia meetodiga leiti keskmiselt 8 eksonit geeni kohta. Eksonite arvu mediaan valgufunktsiooni kohta on 6, eksoni pikkuse mediaan 182 ja introni pikkus 331 aluspaari. Kirjanduse alusel on sellel karbil *ca* 28000 geeni. Teol *L.gigantea* on 359.5 Mb suurune genoom. Valguhomoloogia meetodiga leiti keskmiselt 8.45 eksonit geeni kohta. Kirjanduse andmetel on sellel teol keskmiselt 6 eksonit geeni kohta, keskmine eksoni pikkus 213 ja keskmine introni pikkus 787 aluspaari ning geenide hulk *ca* 24000. Valguhomoloogia meetodi alusel leiti eksonite arvu mediaaniks funktsiooni kohta 5, eksoni pikkuse mediaaniks 197 ja introni pikkuse mediaaniks 391 aluspaari. Teol *A.californica* on 1800 Mb suurune genoom. Valguhomoloogia meetodiga leiti keskmiselt 10.1 eksonit geeni kohta. Eksonite arvu mediaan funktsiooni kohta oli 8, eksoni pikkuse mediaan 164 ja introni pikkuse mediaan 1006 aluspaari. Kirjandusel põhinevat geenide hulka ei ole teada. Arvutuste järgi tuleb selleks ca 18000 – 19000 geeni (arv on saadud kirjandusel põhineva CDSide hulga jagamisega valguhomoloogia meetodiga leitud keskmise eksonite arvuga funktsiooni kohta).

Inimese eksoni pikkuse mediaan on 122 aluspaari (keskmine on 145), eksonite arvu mediaan geeni kohta 7 (keskmine 8.8). Inimesel on keskmiselt on 7.8 intronit geeni kohta ning intronite pikkus varieerub 20 – 11000 aluspaarini (Sakharkar, M. K.,2004). Inimese genoomi suurus on 3300 Mb (http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html) ning see sisaldab umbes 20000 – 28000 geeni. Äädikakärbse (*Drosophila melanogaster*) genoomi suurus on 139.5 Mb ja see sisaldab *ca* 13600 geeni. Äädikakärbsel on keskmiselt 4.67 intronit (Malko, D. B., 2006) ja umbes 4 eksonit geeni kohta (Adams, M. D., jt 2000).

Nendest andmetest võib järeldada, et karpide ja tigude genoomide suurused on liigiti küllaltki varieeruvad. Samas karbid ei erine tigudest eksoni mediaanpikkuste poolest, mis on kõigil vaadeldud liikidel 150 - 200 aluspaari ringis. On võimalik, et tigudele on iseloomulikud pikemad intronid kui karpidel, kuigi genoomide vähesus ei võimalda kindlat järeldust teha. Erandiks on tigu *L.gigantea* kelle

intronite pikkus on karpide intronite pikkusega samas suurusjärgus. See võib olla seletatav väikese genoomi suurusega. Kõigil teadaolevatel molluskitel näib olevat korrelatsioon genoomi suuruse ja intronite keskmise pikkuse vahel. Siin vaadeldud molluskite liikidest on *C.consors*'i genoom kõige suurem. See võib olla seotud tema organismi keerukusega. Röövtoidulise teona peavad *C.consors*'il olema geenid mis kontrollivad tema jahimehhanisme ja aparaate. Lisaks vajavad aktiivselt saaki püüdvad röövteod palju täiuslikumaid meeleorganeid kui planktonit filtreerivad karbid.

5. AUGUSTUS *vs* valguhomoloogia meetod (plussid ja miinused)

Käesolevas töös otsiti *C.consors*'i genoomist geene kahe meetodi abil. Geeniennustusprogramm AUGUSTUS kasutab treenitud liigispetsiifilisi mudeleid ja välist informatsiooni ja ennustab seda kasutades täielikke või osalisi geenimudeleid. Valguhomoloogia meetodi puhul leitakse BLASTX programmi abil UniProtKB/Swiss-Prot andmebaasi valkude homoloogid ning koostatakse neist tõenäoline valkude nimekiri koos oletatavate eksonite asukohtadega genoomil.

Antud töös leiti, et valkude homoloogia alusel leitud ja AUGUSTUSega ennustatud eksonite asukohad erinevad umbes 10%. Seega võib väita, et mõlemad meetodi võimaldavad saada ettekujutust geenide struktuurist.

See, kumba meetodit kasutada uue genoomi geenide kaardistamisel sõltub püstitatud eesmärkidest. AUGUSTUS sobib sellistes olukordades, kui on vaja leida geene, mis veel ei ole UniProtKB/Swiss-Prot andmebaasis olemas. Kasutades liigispetsiifiliselt treenitud mudeleid, suudab programm leida ka liigispetsiifiliste geenide asukohti. Samuti võimaldab AUGUSTUS, eriti transkriptoomi „vihjete“ olemasolul ennustada täpsemalt splaisingsaite. Valguhomoloogia metood sobib annoteeritud geenide ülesleidmiseks ja nende asukoha määramiseks genoomis. AUGUSTUS võimaldab hinnata ka eksonite koguarvu genoomis, mida saab kasutada geenide koguarvu leidmiseks. Juhul, kui uuritav genoom ei ole fragmenteeritud, annab AUGUSTUS ka hinnangu geenide koguarvule. Valguhomoloogia metood leiab üles küll ainult väikese hulga kõigist geenidest, kuid võimaldab hinnata nende struktuuri ja anda hinnang eksonite arvule geeni kohta tänu sellele, et leitavad geenid on homoloogsed juba andmebaasis olevatele. Samuti sobib valguhomoloogia meetod selleks, et genereerida treeningandmestik AUGUSTUSe jaoks. AUGUSTUSe miinuseks on see, et programm kipub ennustama valepositiivseid geene, samuti ei suuda ta alati geene õigesti kokku panna, eriti fragmenteeritud genoomi puhul. AUGUSTUSe puhul tuleks võimalusel igat ennustatud geeni käsitsi kontrollida. Valguhomoloogia metoodi miinuseks võib pidada seda, et leitakse ainult andmebaasis

olevate valkude geene. Andmebaasis ei ole kunagi homolooge kõigile uue organismi geenidele. Lisaks ei leia valguhomoloogia meetod vasteid valkude vähekonserveerunud osadele. Samuti on nii geenide kui eksonite alguse- ja lõpukoordinaadid sageli ebatäpsed.

Kui tegemist on uudse liigi värskelt sekveneeritud genoomiga, siis parim viis saada täiuslikumat pilti genoomist, oleks mõlema meetodi kombineerimine, kuna kasutades neid mõlemaid koos, saab elimineerida nende nõrku külgi.

Mereteo *Conus Consors* geenide arvu ja struktuuri ennustamine bioinformaatiliste meetoditaega.

Viktorija Kukuškina

Kokkuvõte

Teise põlvkonna sekveneerimine on avanud võimalused kiirelt ja odavalt sekveneerida uute seni vähe uuritud kuid huvipakkuvate liikide genoome. Üheks selliseks liigiks on röövtoiduline meretigu *Conus Consors,* kelle neuroaktiivsed konopeptiidid pakuvad suurt farmakoloogilist huvi. Töörühma eesmärgiks oli uudsete konopeptiidide leidmine ja uurimine, samuti oli huvi kätte saada maksimaalselt palju informatsiooni genoomi, geenide koguhulga, nende struktuuri, kordusjärjestuste ja valgustruktuuri kohta. Sellest lähtuvalt sai minu ülesandeks analüüsida, kui edukalt on võimalik kasutada automaatseid geeniennustusalgoritme selleks, et analüüsida uudsete organismide genoome, sealhulgas kirjeldada nende geenide struktuure ja hinnata geenide koguarvu.

Töö tulemusi saab jagada kaheks rühmaks

1) AUGUSTUSe automaatsete geenialgoritmide kasutamise edukus:
    a) Kvaliteetne ja usaldusväärne liigispetsiifiline treeningandmestik, tagab täpsema ja usaldusväärsema tulemuse.
    b) Täiendava informatsiooni lisamine ja lisamudelite (nt. UTR mudel) treenimine aitab tulemuse kvaliteeti tõtsta.
    c) Iga liigi jaoks tuleks võimalusel treenida eraldi liigispetsiifiline geenimudel.

    Kui ei ole võimalik treenida liigispetsiifilist mudelit, tuleks eelistada võimalikult kvaliteetset mudelit. Fülogeneetiliselt lähedase liigi madala kavaliteediga andmestiku pealt treenitud mudeli kasutamine ei ole õigustatud.
2) *C.consors* geenide koguarvu ja struktuuri hinnang:
    a) Valguhomoloogia meetodiga saadud andmetel ja AUGUSTUSe ennustusel põhjal sai oletada, et liigil *C.consors* on *ca* 25000 geeni.
    a) Valguhomoloogia metoodiga tuvastati et antud liigil on keskmiselt 8.47 eksonit geeni kohta.
    b) *C.consors*'i eksonite arvu mediaan valgu kohta on 6, eksoni pikkus 164 aluspaari ja introni

pikkus on 1265 aluspaari.

Käesolevas töös uurisin ning hindasin erinevaid meetodeid, mis aitavad kaasa informatsiooni kättesaamisel teise põlvkonna tehnoloogiatega sekveneeritud genoomidest. Samuti määrasin geenide koguhulga röövtoidulise mereteo liigil *Conus consors.* Et projekti põhieesmärk uudsete konopeptiidide leidmisel on täidetud siis projekt lähiajal lõpetatakse. Edaspidine töö ning selle eesmärgid pole veel paika seatud.

Predicting gene number and structure of sea snail *Conus Consors* with bioinformatic methods.

Viktorija Kukuškina

Summary

Second-generation sequencing technology has opened opportunities for cheap and fast sequencing of genomes of little researched yet interesting species. One of such species is the carnivorous sea snail *Conus Consors*, which provides great interest to pharmacology for its neuroactive conopeptides. The goal of the work-group was finding and studying novel conopeptides, as well as to acquire as much information as possible about genome, total gene count, gene structure, repeats and protein structure. Based on this, my assignment was to analyze the possibility of using automated gene prediction algorithms to analyze novel organisms and describe their gene structure and estimate their gene count. The results of this work can be separated into two groups:

1) Successful usage of AUGUSTUS' automatic gene prediction algorithms:
    a) Reliable and high-quality species-specific training dataset guarantees better precision and reliability of results
    b) Addition of extra information and training of additional models (e.g. UTR model) helps further to raise the quality of the results
    c) It is advisable to train a separate species-specific gene model for each species, when possible. If it is not possible to train a species-specific model, the highest-quality model should be used. Using low-quality model of phylogenetically close organism is not justified.
2) Estimate of *C.Consors* gene amount and structure:
    a) Using protein homology method on acquired data and AUGUSTUS' prediction allows us to estimate *C.consors* gene count as *ca* 25000 genes.
    b) Protein homology method estimated approx. 8.47 exons per gene.
    c) The median number of exons per protein is approximately 6, exon length is 164 base pairs, and intron length is 1265 base pairs.

In this work I researched and evaluated different methods that help retrieve information from genomes produced by second-generation sequencing technologies. Also, I estimated the total amount of genes in carnivorous sea snail species *Conus consors*. Due to the project's primary goal of finding novel conopeptides is completed, this project is going to be completed in the near future. Further work and its goals are not yet determined.

Kasutatud kirjandus

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., ... & Brokstein, P., 2000, The genome sequence of *Drosophila melanogaster*: Science, v. 287, no. 5461, p. 2185-2195.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J., 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs: Nucleic acids research, v. 25, no. 17, p. 3389-3402.

Brauer, A., Kurz, A., Stockwell, T., Baden-Tillson, H., Heidler, J., Wittig, I., Kauferstein, S., Mebs, D., Stöcklin, R., Remm, M., 2012, The mitochondrial genome of the venomous cone snail Conus consors: PLoS One, v. 7, no. 12, e51528.

Dewilde, S., Winnepenninckx, B., Arndt, M. H., Nascimento, D. G., Santoro, M. M., Knight, M., ... & Moens, L., 1998, Characterization of the myoglobin and its coding gene of the mollusc *Biomphalaria glabrata*: Journal of Biological Chemistry, v. 273, no. 22, p. 13583-13592.

Ferris, S. D., Sage, R. D., Prager, E. M., Ritte, U., & Wilson, A. C., 1983, Mitochondrial DNA evolution in mice: Genetics, v. 105, no.3, p. 681-721.

Gallardo, M.H., Bickham, J.W., Kausel, G., Köhler, N., Honeycutt, R.L., 2002, Gradual and quantum genome size shifts in the hystricognath rodents: Journal of Evolutionary Biology, v 16, p. 163-169.

Gregory, T. R., & Mable, B. K., 2005, Polyploidy in animals: The evolution of the genome, v. 171, p. 427-517.

Guigó, R., Agarwal, P., Abril, J. F., Burset, M., & Fickett, J. W., 2000, An assessment of gene prediction accuracy in large DNA sequences: Genome Research, v. 10, no. 10, p. 1631-1642.

Hu, H., Bandyopadhyay, P. K., Olivera, B. M., & Yandell, M., 2011, Characterization of the *Conus bullatus* genome and its venom-duct transcriptome: BMC genomics, v. 12, no. 1, p. 60.

International Human Genome Sequencing Consortium, 2001, Initial sequencing and analysis of the human genome: Nature, v. 409, no. 6822, p. 860–921.

Kanda, A., Takuwa-Kuroda, K., Iwakoshi-Ukena,E., Minakata, H., 2003, Single exon structures of the oxytocin/vasopressin superfamily peptides of octopus, Biochemical and Biophysical Research Communications, v. 309, no. 4, p. 743-748.

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics, v. 27, no. 6, p. 757-763.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., ... & Wang, J., 2010, De novo assembly of human genomes with massively parallel short read sequencing. Genome research: v. 20, no. 2, p. 265-272.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., ... & Law, M., 2012, Comparison of next-generation sequencing systems: BioMed Research International , 2012.

Mahon, A. C., Nambu, J. R., Taussig, R. O. N. A. L. D., Shyamala, M. A. L. L. A. D. I., Roach, A. R. T. H. U. R., & Scheller, R. H., 1985, Structure and expression of the egg-laying hormone gene family in *Aplysia*, The Journal of neuroscience, v. 5, no. 7, p. 1872-1880.

Makałowski, W., 2001, The human genome structure and organization: Acta Biochim. Pol, v. 48, p. 587-598.

Malko, D. B., Makeev, V. J., Mironov, A. A., & Gelfand, M. S., 2006, Evolution of exon–intron structure and alternative splicing in fruit flies and malarial mosquito genomes: Genome research, v. 16, no. 4, p. 505-509.

Montooth, K. L., Abt, D. N., Hofmann, J. W., & Rand, D. M., 2009, Comparative genomics of *Drosophila* mtDNA: novel features of conservation and change across functional domains and lineages: Journal of molecular evolution, v. 69, no. 1 , p. 94-114.

Nata, K., Sugimoto, T., Tohgo, A., Takamura, T., Noguchi, N., Matsuoka, A., ... & Okamoto, H., 1995, The structure of the *Aplysia kurodai* gene encoding ADP-ribosyl cyclase, a second-messenger enzyme, Gene, v. 158, no. 2, p. 213-218.

Parra, G., Bradnam, K., & Korf, I., 2007, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes: Bioinformatics, v. 23, no. 9, p. 1061-1067.

Parra, G., Bradnam, K., Ning, Z., Keane, T., & Korf, I., 2009, Assessing the gene space in draft genomes: Nucleic acids research, v. 37, no. 1, p. 289-297.

Pennisi, E., 2012, ENCODE Project Writes Eulogy For Junk DNA: Science, v. 337**,** no. 6099, p. 1159–1160.

Pop, M., & Salzberg, S. L., 2008, Bioinformatics challenges of new sequencing technology: Trends in Genetics, v. 24, no. 3 , p. 142-149.

Proux-Wéra, E., Armisén, D., Byrne, K. P., Wolfe, K. H., 2012, A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach: BMC Bioinformatics, v. 13, no. 1, p. 237.

Rice, P., Longden, I., Bleasby, A., 2000, EMBOSS: The European Molecular Biology Open Software Suite: Trends Genet, v. 16, p. 276–277.

Sakharkar, M. K., Chow, V. T., & Kangueane, P., 2004, Distributions of exons and introns in the human genome: *In silico biology*, v. 4, no. 4, p. 387-393.

Schmieder, R., & Edwards, R., 2011, Fast identification and removal of sequence contamination from genomic and metagenomic datasets: PLoS one, v. 6, no. 3, e17288.

Simakov, O., Marletaz, F., Cho, S. J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., ... & Rokhsar, D. S., 2012, Insights into bilaterian evolution from three spiralian genomes: Nature.

Simison, W. Brian., Boore, Jeffery L., 2010, Molluscan Evolutionary Genomics.: Lawrence Berkeley National Laboratory, LBNL Paper, LBNL-59179.

Stanke, M., & Waack, S., 2003, Gene prediction with a hidden Markov model and a new intron submodel: Bioinformatics, v. 19, suppl. 2, ii215-ii225.

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S., 2006, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources: BMC bioinformatics, v. 7, no. 1, p. 62.

Takeuchi, T., Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Fujiwara, M., Shinzato, C., Hisata, K., Satoh, N., 2012, Draft genome of the pearl oyster *Pinctada fucata:* a platform for understanding bivalve biology: DNA research, v. 19, no.2, p. 117-130.

Vinogradov, A.E., 2000, Larger genomes for molluskan land pioneers: Genome, v. 43, p. 211-212.

Zenkevitš, L., 1969, Loomade elu: 2. kd., lk 7 -115

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., ... & Wang, J. ,2012, The oyster genome reveals stress adaptation and complexity of shell formation: Nature, v. 490, no. 7418,p. 49-54.

Kasutatud veebiaadressid

Algorithms of Molecular Biology lecture (1999) part „Gene structure in eukaryotes“: http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec07/node8.html (16.05.2013)

Animal Genome Size Database (Mammals): http://www.genomesize.com/statistics.php?stats=mammals (16.05.2013)

Animal Genome Size Database (Molluscs): http://www.genomesize.com/statistics.php?stats=molluscs (16.05.2013)

Ap*lysia californica* genome assembly: http://www.broadinstitute.org/ftp/pub/assemblies/invertebrates/aplysia/ (16.05.2013)

Ap*lysia californica* genome browser: http://genome.ucsc.edu/cgi-bin/hgGateway?org=Sea+hare&db=aplCal1&hgsid=154586973 (16.05.2013)

AUGUSTUS program manual/readme: http://augustus.gobics.de/binaries/README.TXT (16.05.2013)

AUGUSTUS training/test sets for human/fruitfly/arabidopsis models: http://augustus.gobics.de/datasets/ (16.05.2013)

AUGUSTUS UTR model training tutorial: http://bioinf.uni-greifswald.de/bioinf/wiki/pmwiki.php?n=Augustus.UTRTraining (16.05.2013)

BLAT (Blast Like Alignment Tool): hgwdev.cse.ucsc.edu/~kent/src/blatSrc35.zip (16.05.2013)

CONCO: the cone snail genome project for health, Cone snails: http://www.conco.eu/cone_snails.html (16.05.2013)

Co*nus* Biodiversity Website, catalogue of recent and fossil *conus,* „*Conus consors* Sowerby ii, 1833“: http://biology.burke.washington.edu/conus/recordview/record.php?ID=717l66l10129713421llll&tabs=21101011&frms=0&pglimit=&offset=&res=gengrp&srt=&sql2= (16.05.2013)

Cr*assostera gigas* genome assembly: http://www.ncbi.nlm.nih.gov/sra/?term=Crassostrea+gigas (16.05.2013)

Encyclopedia of Life(EOL), Conus (about Cone snails): http://eol.org/pages/50322/details (16.05.2013)

GenBank database: http://www.ncbi.nlm.nih.gov/genbank/ (16.05.2013)

GMAP (Genomic Mapping and Alignment Program for mRNA and EST Sequences): http://research-pub.gene.com/gmap/ (16.05.2013)

Human genome GRCh37: ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests (16.05.2013)

Human Genome Project Information (Genome size): http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml#genomesize (16.05.2013)

Kimball's Biology Pages (Genome Sizes): http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html (16.05.2013)

Lo*ttia gigantea* genome assembly: http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html (16.05.2013)

PASA (Program to Assemble Spliced Alignments): http://pasa.sourceforge.net (16.05.2013)

Pi*nctada fucata* genome assembly: http://marinegenomics.oist.jp/pinctada_fucata (16.05.2013)

Predicting Genes with Augustus (prepare hints): http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/prediction.html#prephints (16.05.2013)

Training AUGUSTUS (compile training and test sets): http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#trainoptions (16.05.2013)

Training AUGUSTUS (create metaparameters file/optimize parameters): http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#meta (16.05.2013)

Training AUGUSTUS (etraining/optimize parameters): http://bioinf.uni-greifswald.de/augustus/binaries/tutorial/training.html#etraining (16.05.2013)