TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Maarja Lepamets

Conus consors'i konopeptiidide geene ümbritsevate kordusjärjestuste statistiline analüüs

Bakalaureusetöö

Juhendaja: prof. Maido Remm
Juhendaja: Märt Möls, PhD

Autor: ..... ”.....”..... 2013

Juhendaja: ..... ”.....”..... 2013

Juhendaja: ..... ”.....”..... 2013

Lubatud kaitsmisele

Professor: ..... ”.....”..... 2013

Tartu 2013

Sisukord

Sissejuhatus **3**

1 Kirjanduse ülevaade **4**

1.1 Teema ja bioloogiliste terminite üldtutvustus 4

1.1.1 Lühidalt perekonnast *Conus* ja konotoksiinidest 4

1.1.2 Genoom, geenid ja kordusjärjestused 4

1.2 DUST 5

1.3 Kasutatud statistilised testid 6

1.3.1 Fisheri täpne test 6

1.3.2 Mann-Whitney test 8

1.4 Suhteline risk ja selle ligikaudsed usalduspiirid 10

2 Praktiline analüüs **12**

2.1 Töö eesmärgid 12

2.2 Algandmed 13

2.2.1 Andmestiku koostamine 13

2.2.2 Lõppandmestike kirjeldav analüüs 14

2.3 Metoodika ja tulemused 16

2.3.1 Lihtsate korduste sagedused 16

2.3.2 Lihtsate korduste pikkused 17

Kokkuvõte **18**

Summary **19**

Viited **20**

Lisad **21**

Sissejuhatus

Mereteo *Conus consors*'i nagu paljude teiste mürgiste loomade toksiinigeene iseloomustab suur koopiate arv ja varieeruvus. Selle mitmekesisuse bioloogiline tekkemehhanism ei ole täpselt teada. Üheks suure varieeruvuse põhjuseks on arvatud lihtsate korduste esinemist geenide ümbruses.

Käesoleva töö eesmärgiks on võrrelda lihtsate korduste esinemissagedust toksiini- ja teiste geenide (eksonite) ümbrustes. Lisaks testisime, kas leitud korduste pikkused on mõlemas valimis sarnased. Märgatavalt suurem korduste esinemistõenäosus või nende pikkuste oluline erinevus oleks potentsiaalseks kinnituseks hüpoteesile, et kordustel on oluline roll toksiinigeenide varieeruvuse tekitamisel.

Bakalaureusetöö koosneb kahest peatükist. Esimene peatükk sisaldab uurimuse bioloogilise tausta ja terminite tutvustust, samuti töös kasutatud statistiliste testide ja suuruste teoreetiliste tagamaade kirjeldusi. Teine peatükk koosneb analüüsitava andmestiku kokkupanemise metoodika kirjeldusest, andmestiku kirjeldavast analüüsist ja tulemuste kokkuvõttest.

Töö on kirjutatud Tartu Ülikooli Molekulaar- ja rakubioloogia instituudis Bioinformaatika õppetoolis. Töös kasutatud lähteandmed pärinevad Euroopa Liidu 6-nda raamprogrammi projektist CONCO – Cone Snail Genome Project for Health (LSHB-CT-2007/037592).

1 Kirjanduse ülevaade

1.1 Teema ja bioloogiliste terminite üldtutvustus

1.1.1 Lühidalt perekonnast *Conus* ja konotoksiinidest

Perekonda *Conus* ehk koonuskodalane kuulub ligikaudu 500 erinevat liiki. Tegemist on meres elavate mürgiste tigudega, kes toituvad kaladest, ussidest või teistest molluskitest. Kõik liigid sünteesivad tugevatoimelist mürki konotoksiini, mis sisaldab ligikaudu sadat erinevat valgulist konopeptiidi. Nende segu lastakse harpuunilaadse moodustise abil saaklooma kehasse. Seal mõjuvad toksiinid spetsiifiliselt närvirakkudes asuvatele ioonkanalitele, muutes ohvri liikumisvõimetuks. (Terlau *et al.*, 2004) Taolise mõju tõttu uuritakse konotoksiine eelkõige kui potentsiaalseid lõdvestava või tuimestava toimega aineid.

1.1.2 Genoom, geenid ja kordusjärjestused

Genoomiks nimetatakse kogu pärilikku materjali, mis sisaldub organismi ühes rakus. Keemiliseks pärilikkusekandjaks on desoksüribonukleiinhappe (DNA) molekulid, mis koosnevad lineaarselt ühendatud nukleotiididest (lühend: nt.). Nukleotiide on 4 tüüpi: adeniin (A), tsütosiin (C), guaniin (G) ja tümiin (T). Nende järgnevus määrab DNA vastava piirkonna funktsionaalsuse. Genoomi võib jagada valgugeenideks, ribonukleiinhappe (RNA) geenideks, regulaatorpiirkondadeks ning intergeenseteks aladeks. Valgugeenidelt sünteesitakse RNA, mille alusel pannakse kokku valgujärjestus. RNA geenidelt valke ei sünteesita. Nende geenide produktid jäävad rakku RNA kujul, viies seal läbi mitmeid eluliselt tähtsaid protsesse. Regulaatorpiirkondade ülesandeks on raku seest või teda ümbritsevast keskkonnast tulevatele signaalidele vastavalt geeniproduktide sünteesi kas aktiveerida või inhibeerida.

Organisme, mille rakkude tuumad on ülejäänud rakusisesest keskkonnast membraaniga eraldatud, nimetatakse eukarüootideks ehk päristuumseteks. Nende hulka kuuluvad kõik taimed ja loomad. Eukarüootide geenid koosnevad valke kodeerivatest eksonitest ja nende vahel paiknevatest intronitest, mis lõigatakse välja enne antud geeni lõpp-produkti sünteesi.

Üks oluline tunnus, mida genoomi iseloomustamiseks kasutatakse, on temas sisalduvate korduvate järjestuste osakaal. Kordused võivad paikneda nii intergeensetes alades kui ka geenide sees (eelkõige intronite piirkonnas). Üks eukarüootide genoomis leiduvaid korduste tüüpe on lihtsad kordused (vt. joonis 1), mida nimetatakse ka madala kompleksusega aladeks. Tegemist

on DNA regioonidega, mis enamasti koosnevad järjestikku korduvatest ühe kuni nelja nukleotiidi pikkustest motiividest. Lihtsa korduse täpne definitsioon varieerub sõltuvalt kasutatud korduste leidmise programmist. Need lihtsad kordused, mis asuvad geenide sees või ümbruses, võivad suurendada eksonite duplikatsioonide, deletsioonide või ümberpaiknemiste tõenäosust (Stankiewicz ja Lupski, 2006).

1.2 DUST

DUST on programm, mida kasutatakse DNA järjestustes paiknevate lihtsate kordustega alade tuvastamiseks. DUST kasutab heuristilist algoritmi, et arvutada alamsõnedele skoorid vastavalt sellele, kui suured on nendes paiknevate kolmetäheliste järjestuste (tripletite) sagedused. Mingi tripleti suur sagedus suurendab skoori ja viitab selle regiooni madalamale kompleksusele genoomis.

DUSTi käivitades on võimalik määrata piirskoor, millest alates soovitakse järjestust nimetada lihtsaks korduseks. Tavaliselt võetakse see vahemikust 20 kuni 30. Mida suurem on etteantud skoor, seda vähem genoomiregioone leitakse ehk seda ”rangemad” on lihtsa kordusjärjestuse tuvastamise kriteeriumid. (Morgulis *et al.*, 2006)

a. GGTAGGAAaTGAAGGGGAAAAAAAAAAGTACCGTTTCTGATGATGACACT

b. TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTGTGTCGGACAG

c. GATGATGATGATGATGATGATGATGATGAGATTGGAAGTGGTTACAGTAG

d. CGGGCAGACAGACAGACAGACAGACAGGCAGACAGACAGGCAGAGAGACA

Joonis 1: *Lihtsad kordused Conus consors'i genoomist eksonite 50 nt. ümbrustest. Punasega on märgitud DNA regioon, mida DUST (piirskoori väärtusega 30) on tuvastanud kui madala kompleksusega ala. Joonisel on toodud näited (a) A-nukleotiidi kordusest, (b) dinukleotiidsest kordusest, korduv motiiv: TC, (c) trinukleotiidsest kordusest, korduv motiiv: GAT ja (d) tetranukleotiidsest kordusest, korduv motiiv: CAGA.*

1.3 Kasutatud statistilised testid

1.3.1 Fisheri täpne test

Fisheri täpne test on *sir* Ronald A. Fisheri demonstreeritud kahe- ja enamamõõtmeliste sagedustabelite analüüsiks mõeldud statistiline test (Fisher, 1925). Võrreldakse binaarse tunnuse suhtelisi sagedusi mitmes erinevas grupis. Praktikas kasutatakse seda eelkõige väikeste valimimahtude ja 2×2 sagedustabelite korral, kuid test töötab analoogiliselt ka teistel juhtudel.

Leidugu testgrupp mahuga n_1 ja kontrollgrupp mahuga n_2, kus mõlema puhul on igal uuritaval objektil mingi binaarse tunnuse väärtus kas 1 või 0, kusjuures testgrupis on a ja kontrollgrupis b objekti väärtusega 1 ning vastavalt $(n_1 - a)$ ja $(n_2 - b)$ objekti väärtusega 0. Tähistagu X_0 saadud tabelit (vt. tabel 1).

Tabel 1: *Näide genereeritud kahemõõtmelisest sagedustabelist.*

	testgrupp	kontrollgrupp	kokku
1	a	b	$a + b$
0	$n_1 - a$	$n_2 - b$	$n_1 + n_2 - a - b$
kokku	n_1	n_2	$n_1 + n_2$

Soovime teada saada, kas erinevates gruppides on väärtuse 1 esinemistõenäosused erinevad ehk kas vaadeldav binaarne tunnus sõltub grupist. Nullhüpoteesiks on seega

$$H_0 : \pi_t = \pi_k,$$

kus π_t on väärtuse 1 esinemistõenäosus testgrupis ja π_k vastav esinemistõenäosus kontrollgrupis.

Tähistame saadud kahemõõtmelise tabeli üldjuhu tähega X (vt. tabel 2).

Tabel 2: *Üldkuju* 2×2 *tabelist valimimahtudega* n_1 *ja* n_2*.*

	testgrupp	kontrollgrupp	kokku
1	x_t	x_k	$x_t + x_k$
0	$n_1 - x_t$	$n_2 - x_k$	$n_1 + n_2 - x_t - x_k$
kokku	n_1	n_2	$n_1 + n_2$

Suvalise sellise tabeli X saamise tõenäosus on

$$P(X|\pi) = \binom{n_1}{x_k}\binom{n_2}{x_t}\pi^{x_k+x_t}(1-\pi)^{n_1+n_2-x_k-x_t},$$

kus $\pi_t = \pi_k = \pi$. Täpne π väärtus on teadmata. Seega ülaltoodud valemist lähtudes täpset p-väärtust leida ei õnnestu. Olgu

$$\Gamma = \{X : X\text{on tabel 2-ga analoogiline } 2 \times 2 \text{ tabel}\}.$$

Fisheri test võtab vaatluse alla kõik sellised tabelid $X \in \Gamma$, mille korral $x_t + x_k = a + b$. Tähistame neid järgmiselt:

$$\Gamma(a+b) = \{X : X\text{on } 2 \times 2 \text{ sagedustabel, kus } x_t + x_k = a + b\}.$$

Test põhineb tähelepanekul, et nullhüpoteesi kehtides on iga sellise tabeli saamise tõenäosus hüpergeomeetriline, s.t.

$$P(X|X \in \Gamma(a+b)) = \frac{\binom{n_1}{x_t}\binom{n_2}{x_k}}{\binom{n_1+n_2}{a+b}}.$$

Täpne kahepoolne p-väärtus on kõigi selliste tabelite X, mis on sama ekstreemsete või ekstreemsemate osakaaludega kui tabel X_0 saamise tõenäosuste summa. Tarkvarapaketi R (R Core Team, 2013) standardfunktsioon Fisheri täpse testi sooritamiseks defineerib meid huvitava tabeliga sama ekstreemse või veel ekstreemsema tabeli kui tabeli, mille saamise tõenäosus on meid huvitava tabeli saamise tõenäosusega võrdne või sellest väiksem. Seega avaldub Fisheri täpse testi p-väärtus kujul

$$P_{\mathcal{F}} = \sum_{P(X|\pi)\leq P(X_0|\pi)} P(X|X \in \Gamma(a+b)) = \sum_{P(X|\pi)\leq P(X_0|\pi)} \frac{\binom{n_1}{x_t}\binom{n_2}{x_k}}{\binom{n_1+n_2}{a+b}}.$$

Fisheri testi p-väärtust saab leida ka $2 \times k$ sagedustabelite korral. Sel juhul avaldub sama valem kujul

$$P_{\mathcal{F}} = \sum_{P(X|\pi)\leq P(X_0|\pi)} \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}...\binom{n_k}{x_k}}{\binom{n_1+n_2+...+n_k}{a_1+a_2+...+a_k}},$$

kus $n_1...n_k$ on vaadeldavate valimite mahud, $x_1...x_k$ nende objektide arvud, mille korral vaadeldava tunnuse väärtus võrdub 1-ga ja $a_1 + a_2 + ... + a_k$ väärtusega 1 objektide koguarv, mis Fisheri täpse testi puhul on fikseeritud.

Selle illustreerimiseks leiame näitena ühe konkreetse 2×3 sagedustabeli (vt. tabel 3) saamise tõenäosuse.

Tabel 3: 2×3 *sagedustabeli näide.*

	esimene grupp	teine grupp	kolmas grupp	kokku
1	2	1	6	9
0	9	3	1	13
kokku	11	4	7	22

Sellise tabeli saamise tõenäosus avaldub kujul

$$p = \frac{\binom{2+9}{2}\binom{1+3}{1}\binom{6+1}{6}}{\binom{22}{2+1+6}} = 0.0031.$$

Kõikvõimalikke antud valimimahtudega tabeleid on kokku 480. Neid, mille saamise tõenäosus on tabel 3 saamise tõenäosusest väiksem või sellega võrdne ning mille korral väärtusega 1 saadud tulemuste summa on võrdne 9-ga, on kokku 14. Liites viimaste saamise tõenäosused, saame tulemuseks ligikaudu 0.014, mis ongi otsitavaks p-väärtuseks. Olles eelnevalt valinud usaldusnivooks 0.05, võime väita, et erinevates gruppides on väärtusega 1 objektide osakaalud erinevad (vt. lisad).

1.3.2 Mann-Whitney test

Mann-Whitney test (Mann ja Whitney, 1947) on mitteparameetriline statistiline test kahe sõltumatu populatsiooni võrdlemiseks. Eelkõige kasutatakse seda tuvastamiseks, kas ühe populatsiooni mingi konkreetse tunnuse väärtused on suuremad teise populatsiooni sama tunnuse väärtustest.

Leidugu testgrupp mahuga n_1 ja kontrollgrupp mahuga n_2 ning tähistagu

$$n = n_1 + n_2$$

kõigi vaadeldavate objektide koguarvu. Mann-Whitney testi nullhüpotees on

$$H_0 : P(X > Y) + 0.5P(X = Y) = 0.5,$$

kus X ja Y on vastavalt testgrupist ja kontrollgrupist vaadeldava tunnuse juhuslikult valitud väärtused. Mann-Whitney testi sooritamiseks moodustatakse uuritava tunnuse kõigist väärtustest

variantsioonrida ning leitakse igale neist vastav astak ehk positsioon selles reas. Võrdsete väärtuste korral seatakse nende kõigiga vastavusse nende astakute keskmine, mis avaldub valemiga

$$r = \frac{a_1 + a_2 + ... + a_k}{k},$$

kus k on ühe konkreetse väärtuse esinemiste arv ning $a_1...a_k$ nende väärtuste positsioonid variantsioonreas.

Olgu

$$R = r_1 + r_2 + ... + r_{n_1} \quad \text{ja} \quad S = s_1 + s_2 + ... + s_{n_2}$$

vastavalt testgrupis ja kontrollgrupis esinevate vaadeldava tunnuse väärtuste astakute summad. Omavahel on nad seotud valemiga

$$R = 0.5n(n+1) - S.$$

S minimaalne võimalik väärtus on $0.5n_1(n_1 + 1)$ ning R minimaalne võimalik väärtus on $0.5n_2(n_2 + 1)$. Mann-Whitney teststatistik testgrupile on defineeritud järgmiselt:

$$U = R - 0.5n_1(n_1 + 1).$$

Vastav statistik kontrollgrupile on

$$U_k = S - 0.5n_2(n_2 + 1),$$

kusjuures eelnevatest avaldistest lähtub, et

$$U + U_k = R - 0.5n_1(n_1 + 1) + S - 0.5n_2(n_2 + 1) = n_1n_2.$$

Sisuliselt loetleb Mann-Whitney statistik U juhte, kus, vaadates kõikvõimalikke väärtuste paare, on testgrupist võetud väärtus kontrollgrupist võetud väärtusest suurem ehk

$$U = \#_{i,j}\{X_i > Y_j\},$$

kus $i = 1...n_1$ ja $j = 1...n_2$.

P-väärtuse leidmiseks tuleb leida teststatistiku jaotus nullhüpoteesi kehtides. Selle jaotuse tuvastamiseks tuleb vaadelda test- ja kontrollgrupis vaadeldava tunnuse väärtuste kõiki võimalikke astakute kombinatsioone antud valimimahtude korral. Kui välistada võrdsete astakute esinemisvõimalus, on taoliseid kombinatsioone kokku $\binom{n}{n_1}$. Nullhüpoteesi kehtides on

kõigi selliste kombinatsioonide esinemistõenäosused võrdsed. Edasi tuleb kõigi kombinatsioonide jaoks leida statistikute R ja U väärtused. Leides kõikide võimalike U väärtuste jaoks nende esinemiste osakaalud, mille tähistame $p(u_i)$, kus u_i tähistab statistiku U mingit konkreetset väärtust, olemegi leidnud tema jaotuse nullhüpoteesi kehtides. Tähistame U_t-ga meid huvitava valimi põhjal leitud U väärtuse. Ühepoolse p-väärtuse leiame valemist

$$p = \sum_{u_i \geq U_t} p(u_i).$$

Täpsele Mann-Whitney testi p-väärtusele lähedast tulemust on võimalik saada, genereerides Monte-Carlo meetodil palju juhuslikke testgrupi mahuga astakute komplekte, arvutades nende kõikide puhul teststatistiku U väärtuse ja hinnates saadud tulemuste korral U jaotust nullhüpoteesi kehtides. Sellisel juhul ei tea me täpset teststatistiku jaotust, kuid võime seda ligilähedaselt hinnata. Taoline lähenemine kasutab vähem ressursse kui täpse jaotuse leidmine.

Rakendustarkvara R võimaldab Mann-Whitney testi sooritada Wilcoxoni astaksummatesti nime all. Selle testi standardfunktsioon R-s ei oska aga arvestada võimalike korduvate astakute väärtustega. Seetõttu kasutatakse eelistatult Wilcoxoni astaksummatesti, mis sisaldub R lisapaketis *coin* (Hothorn *et al.*, 2008).

1.4 Suhteline risk ja selle ligikaudsed usalduspiirid

Suhteline risk (RR) on eelkõige epidemioloogias laialt kasutatav suurus, mis hindab, kuivõrd mingi riskifaktori mõju muudab tagajärje tekkimise tõenäosust. Üldisemalt saab seda kasutada kahes grupis olevate objektide mingi tunnuse esinemistõenäosuste võrdlemiseks. Olgu meil antud 2×2 sagedustabel (vt. tabel 4), kus on nii uuritava kui ka kontrollgrupi jaoks välja toodud nende objektide sagedused, millel vaadeldav tunnus esines ja millel seda tunnust ei esinenud.

Tabel 4: *Kahemõõtmeline sagedustabel.*

	uuritav grupp	kontrollgrupp	kokku
tunnus esines	A	C	$A + C$
tunnust ei esinenud	B	D	$B + D$
kokku	$A + B$	$C + D$	$A + B + C + D$

Kasutades tabelis 4 olevaid tähistusi, avaldub suhtelise riski hinnang valemiga

$$\hat{RR} = \frac{\frac{A}{A+B}}{\frac{C}{C+D}}.$$

Kui suhteline risk on võrdne ühega, siis gruppide riskid ei erine. Ühest suurema suhtelise riski korral on uuritavas grupis vaadeldava tunnuse esinemise tõenäosus suurem kui kontrollgrupis ning ühest väiksema suhtelise riski puhul on sama tunnuse esinemistõenäosus kontrollgrupis suurem kui uuritavas grupis.

Suhtelise riski ligikaudse $(1-\alpha)$-usaldusintervalli piirid avalduvad valemiga

$$UI = e^{\ln(\hat{RR}) \pm z_{\alpha/2} * \hat{SE}},$$

kus $\ln(RR)$ on naturaallogaritm suhtelisest riskist, $z_{\alpha/2}$ on normaaljaotuse $\alpha/2$-kvantiil ning

$$\hat{SE} = \sqrt{\frac{B}{A*(A+B)} + \frac{D}{C*(C+D)}}$$

on riskide suhte standardvea hinnang.

2 Praktiline analüüs

2.1 Töö eesmärgid

Meie uurimisgrupi töö on keskendunud kaladest toituva *Conus consors*'i genoomi kokkupanemisele ja analüüsile. Olulise osa kogu uurimistööst moodustab nimetatud liigi konotoksiinigeenide leidmine ja iseloomustamine. Varemalt on täheldatud, et nende mürkide koostis varieerub suures ulatuses isegi samast liigist isendite piires (Dutertre *et al.*, 2010) ning meid huvitab, millise bioloogilise mehhanismiga taoline varieeruvus on tekkinud.

Üheks võimalikuks varieeruvuse tekkepõhjuseks on geenijärjestuste või selle kodeerivate osade duplitseerumine, deleteerumine või ümberpaiknemine (Chang ja Duda Jr, 2012). Seda laadi struktuursete ümberkorralduste keskmisest sagedasemat esinemist mingis geenide grupis või perekonnas võivad põhjustada mitmesugused genoomis leiduvad järjestused, näiteks lühikesed korduvad motiivid. Käesolevas töös on vaatluse all konotoksiinigeenide eksonitega külgnevad alad genoomis ja neis leiduvad lihtsad kordusjärjestused.

Töö eesmärgid on püstitatud järgnevalt:

- Võrrelda lihtsate korduste esindatust konotoksiini ja muude geenide eksonite ümbrustes.
- Võrrelda konotoksiini ja muude geenide eksonite ümbrustes paiknevate lihtsate korduste pikkuseid.

2.2 Algandmed

2.2.1 Andmestiku koostamine

Käesoleva töö põhiliseks algandmestikuks on *Conus consors*'i genoomne järjestus. Siinkohal tasub märkida, et nimetatud genoomi terviklik järjestus ei ole täielikult teada. Selle kindlakstehtud osa koosneb erinevate pikkustega DNA lõikudest, mida nimetatakse kontiigideks ning mille paiknemine üksteise suhtes on teadmata. Lisaks on töös kasutatud konopeptiidide ja teiste geenide eksonite asukohti sisaldavat tabelit. Nimetatud tabelist saab vaadata, millisel kontiigil konkreetne ekson asub ning milliselt positsioonilt ta algab ja lõppeb.

Esmalt kirjutati Pythonis programm (vt. lisad), mis väljastas etteantud eksonite asukohtade järgi kasutaja poolt määratud pikkusega eksonite ümbrused (juhul, kui antud pikkusega ümbrust oli täispikkuses võimalik väljastada). Sobiva pikkuse valimiseks lasti programmil väljastada kõik kuni 1000 nt. pikkused lõigud konopeptiidide eksonite ümbrustest ning, jälgides, et andmemahud ei jääks liiga väikseks, valiti analüüsiks 50 nt. ja 100 nt. pikkused ümbrused.

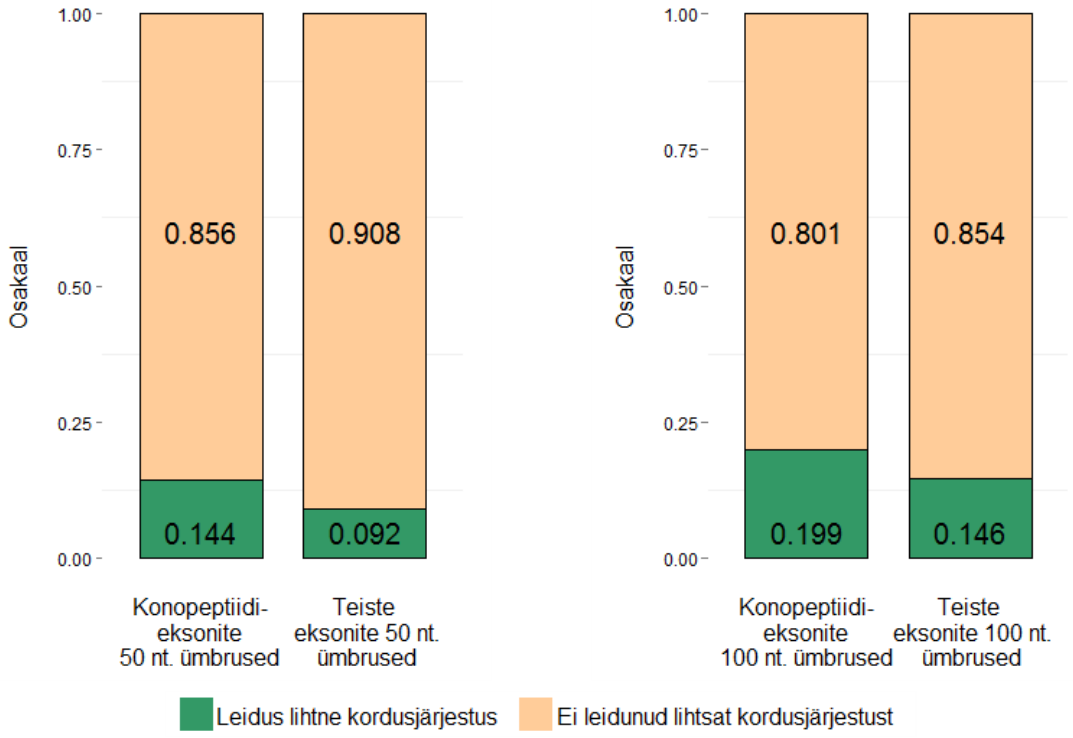
Saadud DNA järjestused anti sisendiks programmile DUST, valides DUSTi piirskoori väärtuseks 30. Saadud väljundi põhjal saadi iga eksoni nii 50 nt. kui ka 100 nt. ümbrusega vastavusse seada kas 1 või 0, olenevalt, kas DUST tuvastas vaadeldavast ümbrusest lihtsa kordusjärjestuse või mitte.

Lisaks kirjutati Pythonis teine programm (vt. lisad), mis tuvastas DUSTi poolt leitud kordusjärjestuse ning väljastas selle pikkuse. Ülalnimetatud programmi kasutati ainult 100 nt. ümbruste analüüsiks, et vältida pikemate korduste alaesindatusest tulenevat süstemaatilist viga.

Kirjeldatud metoodikat kasutades loodi kaks andmetabelit, millest üks koosnes konopeptiidide ja teine ülejäänud eksonite ümbruste kohta saadud andmetest. Näite saadud tabelist võib leida töö lisadest (vt. lisad).

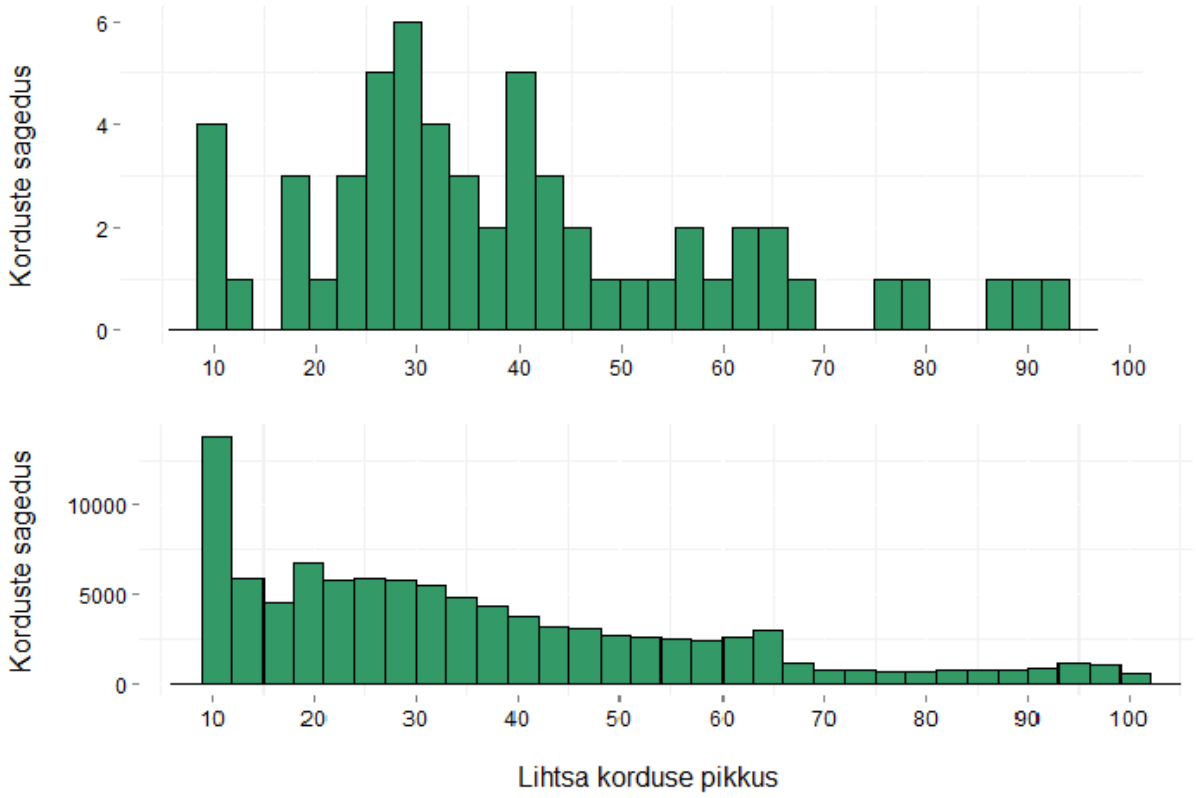
2.2.2 Lõppandmestike kirjeldav analüüs

Koostatud andmestikus on 50 nt. pikkuseid konopeptiidieksonite ümbruseid 355 ja 100 nt. pikkuseid ümbruseid 281. Lihtsaid korduseid tuvastati esimesel juhul 51 ja teisel juhul 56 ümbrusest, mis moodustavad kõikidest sama pikkusega ümbrustest vastavalt $14,4\%$ ja $19,9\%$. *Conus consors*'i ülejäänud eksonite ümbert saadi 50 nt. pikkusega DNA lõike 789402 ja 100 nt. pikkusega DNA lõike 653621, millest korduseid tuvastasime vastavalt 73010 $(9,2\%)$ ja 95649 $(14,6\%)$ eksoni ümbrusest (vt. joonis 2).



Joonis 2: *Lihtsate kordusjärjestuste sagedusjaotus konopeptiidide ja teiste eksonite 50 nt. (vasakul) ja 100 nt. (paremal) ümbrustes.*

Konopeptiidide eksonite ümbrustest leitud lihtsate korduste keskmine pikkus oli $39,6$ nt. ja standardhälbeks $20,4$ nt. Teiste eksonite puhul oli keskmiseks pikkuseks $35,3$ nt. ja standardhälbeks $22,3$ nt. (vt. joonis 3). Ligikaudsed 95% usaldusvahemikud keskmisele olid vastavalt $(34,4...44,9)$ ja $(35,1...35,4)$. Leiti ka kordustes asuvate nukleotiidide hulgad iga ümbruse kohta (korduse puudumise korral loeti tema pikkuseks 0). Konopeptiidide puhul oli tulemuseks $8,1779$ $(8,0488...8,3071)$ nt. ja teiste eksonite puhul $5,31011$ $(5,31007...5,31016)$ nt.



Joonis 3: *Lihtsate kordusjärjestuste pikkuste jaotus konopeptiidide eksonite (üleval) ja teiste eksonite (all) ümbrustes. Korduste pikkuseid vaadeldi vaid 100 nt. ümbrustes.*

2.3 Metoodika ja tulemused

2.3.1 Lihtsate korduste sagedused

Töö peamiseks eesmärgiks oli uurida, kas lihtsate korduste esinemistõenäosused konopeptiidide ja teiste geenide eksonite ümbrustes on erinevad. Selleks moodustati eelmises peatükis väljatoodud sagedustest 2×2 sagedustabelid (vt. tabeleid 5 ja 6).

Tabel 5: *Lihtsate korduste esindatused konopeptiidide ja teiste eksonite 50 nt. ümbrustes.*

	konopeptiidid	%	teised eksonid	%	kokku
lihtne kordusjärjestus	51	$14,4\%$	73010	$9,2\%$	73061
kordusjärjestust ei tuvastatud	304	$85,6\%$	716392	$90,8\%$	716696
kokku	355	100%	789402	100%	789757

Tabel 6: *Lihtsate korduste esindatused konopeptiidide ja teiste eksonite 100 nt. ümbrustes.*

	konopeptiidid	%	teised eksonid	%	kokku
lihtne kordusjärjestus	56	$19,9\%$	95649	$14,6\%$	95705
kordusjärjestust ei tuvastatud	225	$80,1\%$	557972	$85,4\%$	558197
kokku	281	100%	653621	100%	653902

Vaadeldud kordusjärjestuste esinemissageduste põhjal hinnati suhet

$$RR = \frac{P(\text{leidub kordusjärjestus} \mid \text{konopeptiidi ekson})}{P(\text{leidub kordusjärjestus} \mid \text{muu ekson})}$$

ehk kordusjärjestuse esinemise suhtelist riski koos 95% usalduspiiridega. Usaldusnivooks määrati $0,05$ ning kontrolliti saadud suhete statistilist olulisust kahepoolse Fisheri täpse testi abil. Eksonite 50 nt. ümbruste analüüsil saadi riskide suhteks $1,55$ $(1,20...2,00)$. Kuna terve usaldusvahemik jääb reaalarvude sirgel 1-st paremale, viitab saadud tulemus lihtsate korduste suuremale esinemistõenäosusele konopeptiidide hulgas. Fisheri täpne test andis p-väärtuseks $0,0017$. Seega on leitud erinevus valitud usaldusnivool statistiliselt oluline.

Eksonite 100 nt. ümbruste korral saadi riskide suhteks $1,36$ $(1,08...1,72)$. Ka siin viitab saadud tulemus lihtsate korduste üleesindatusele konopeptiidide ümbrustes. Fisheri täpse testiga saadud p-väärtus $(0,014)$ kinnitab valitud usaldusnivool leitud erinevuse olulisust.

Kuigi lihtsate korduste suur sagedus mingis genoomi piirkonnas ei viita otseselt selle regiooni kiiremale muutusele evolutsiooni käigus, siis on teada, et kordusjärjestused võivad suurendada DNA järjestuse struktuursete ümberkorralduste tõenäosust, näiteks teatud piirkonna deleteerumist, duplitseerumist või ümberpaiknemist. Seega võib lihtsate korduste üleesindatus konopeptiidieksonite ümbrustes olla üheks põhjuseks konotoksiinide suurele varieeruvusele.

2.3.2 Lihtsate korduste pikkused

Teiseks sooviti teada saada, kas konopeptiidide ja teiste eksonite ümbrustes asuvad lihtsad kordusjärjestused erinevad mingi neid iseloomustava tunnuse alusel. Nimetatud tunnuseks valiti selle DNA regiooni pikkus, mille DUST märkis kui leitud lihtsa korduse. Kirjeldavas analüüsis väljatoodud ligikaudsete usaldusvahemike põhjal tuleks jääda nullhüpoteesi juurde. Kuna aga puudub informatsioon korduste pikkuste jaotuse kohta, siis viidi läbi ka mitteparameetriline test. Usaldusnivooks määrati taaskord $0,05$. Korduseid võrreldi Mann-Whitney testiga rakendustarkvara R paketi *coin* abil.

Võrreldi nende lihtsate korduste pikkuseid, mille DUST konopeptiidide või teiste eksonite ümbrustest tuvastas. Sooritati Mann-Whitney täpne test, mis andis p-väärtuseks $0,11$. Määratud usaldusnivood arvesse võttes ei osutunud saadud tulemus statistiliselt oluliseks.

Lisaks võrreldi korduvates elementides leiduvat nukleotiidide hulka iga eksoni ümbruse kohta, märkides nendes ümbrustes, kust lihtsat kordust ei leitud, korduse pikkuseks 0. Andmete kirjeldava analüüsi peatükis on näha, et sellisel juhul korduste keskmiste pikkuste usaldusintervallid ei kattu. Monte-Carlo meetodiga leiti ligikaudne Mann-Whitney testi p-väärtus. Tulemuseks saadi $0,0048$. Seega on pikkuste erinevus valitud usaldusnivool statistiliselt oluline. Järelikult paikneb konopeptiidide eksonite ümbrustes suurem arv nukleotiide lihtsate korduvate motiividena. Sellised korduvad motiivid võivad olla seotud genoomi struktuursete ümberkorraldustega, mis seletaks konopeptiidide suurt varieeruvust.

Kokkuvõte

Käesolevas töös uuriti, kas mereteo *Conus consors*'i toksiinigeenide ümbrustes leidub rohkem lihtsaid kordusjärjestusi võrreldes teiste geenide ümbrustega.

Esmalt loeti kokku programmi DUSTi poolt tuvastatud korduste regioonid. Konopeptiidigeenide eksonite 50 nt. ümbrustes leidus lihtsaid korduseid $1,55$ ja 100 nt. ümbrustes $1,36$ korda rohkem kui teiste geenide eksonite ümbrustes. Fisheri täpne test tunnistas leitud erinevused mõlemal juhul statistiliselt olulisteks ($p = 0,0017$ ja $p = 0,014$).

Lisaks kontrolliti konopeptiidigeenide ja teiste geenide eksonite ümbrustest leitud lihtsate korduste pikkuste erinevust. Konopeptiidide puhul oli korduse keskmiseks pikkuseks $39,6$ nt. ja teiste eksonite puhul $35,3$ nt., kuid Mann-Whitney testi sooritamisel leiti, et see erinevus ei ole statistiliselt oluline ($p = 0,11$).

Viimaks loeti nukleotiidide hulka kordustes. Konopeptiidigeenide eksonite ümbrustes oli keskmiselt $8,18$ korduvas motiivis asuvat nukleotiidi iga ümbruse kohta. Teiste eksonite puhul saadi tulemuseks $5,31$ nt. iga ümbruse kohta. Mann-Whitney testi sooritamisel ja ligikaudse p-väärtuse arvutamisel leiti, et saadud erinevus on statistiliselt oluline ($p = 0,0048$).

Seega võime väita, et konopeptiidide eksonite ümbrustes on mõnevõrra rohkem lihtsaid kordusjärjestusi. Kuivõrd varem on näidatud, et lihtsad kordused võivad suurendada genoomi mingi piirkonna varieeruvust, võivad nad antud juhul osaleda konotoksiinide geenide suure arvu ja mitmekesisuse evolutsioonis.

Toksiinigeenide evolutsiooni mõistmine aitab paremini leida bioloogiliselt aktiivseid aineid, mis on potentsiaalsed ravimikandidaadid. Käesolevas töös tuvastatud lihtsate korduste suurem osakaal annab põhjenduse uurimistöö jätkamiseks, leidmaks nende korduste bioloogilist rolli geneetika ja molekulaarbioloogia meetoditega.

Statistical analysis of *Conus consors*' conotoxin genes flanking regions

Bachelor thesis

Maarja Lepamets

Summary

The purpose of this bachelor thesis is to analyse the flanking regions of the genes of the marine cone snail *Conus consors*. A common feature of many venomous animals including *Conus consors* is a wide variety among their toxin genes. The biological mechanism which generates that kind of variety is currently unknown. One of the possibilities is the existence of simple repeated sequences. It has been shown that these repeats increase the frequency of structural variance in gene sequences.

First, the repeated regions found by DUST were counted separately for conopeptide and other genes (exons). There were 55% more repeats in the conopeptide 50 nt. flanking regions and 36% more repeats in the conopeptide 100 nt. flanking regions compared to other exons. According to Fisher's exact test the difference was statistically significant ($p < 0,05$).

Secondly, the length of the repeated regions were determined and compared. Although on average the repeated regions in the conopeptide flanking regions were about 5 nt. longer, acoording to Mann-Whitney test the difference was not statistically significant ($p = 0,11$). Furthermore, the number of nucleotides in repeats per one flanking region were calculated and compared. There were approximately $8,18$ and $5,31$ nucleotides in repeated motifs per flanking region in conopeptide and other genes, respectively. According to Mann-Whitney test the difference was statistically significant ($p < 0,05$).

Knowing the evolutionary mechanism behind the variety of toxin genes helps to discover new biologically active substances which serve as potential drugs. The results of this paper indicate the fact that there are more simple repeats near conopeptide genes. In order to prove the biological significance of current result the matter should be taken to the wet lab.

Viited

Chang, D., Duda Jr, T. F. (2012). Extensive and Continuous Duplication Facilitates Rapid Evolution and Diversification of Gene Families. Mol Biol Evol. 29: 2019-29.

Dutertre, S., Biass, D., Stöcklin, R., Favreau, P. (2010). Dramatic Intraspecimen Variations Within the Injected Venom of *Conus consors*: an Unsuspected Contribution to Venom Diversity. Toxicon 55: 1453-62.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Hothorn, T., Hornik, K., van de Wiel, M. A., Zeileis, A. (2008). Implementing a Class of Permutation Tests: The coin Package. Journal of Statistical Software 28: 1-23. URL: http://www.jstatsoft.org/v28/i08/.

Mann, H. B., Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Stat. 18: 50-60.

Morgulis, A., Gertz, E. M., Schäffer, A. A., Agarwala, R. (2006). WindowMasker: Window-Based Masker for Sequenced Genomes. Bioinformatics 22: 134-41.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Stankiewicz, P., Lupski, J. R. (2006). The Genomic Basis of Disease, Mechanisms and Assays for Genomic Disorders. Genome Dyn. 1: 1-16.

Terlau, H., Olivera, B. M. (2004). *Conus* Venoms: A Rich Source of Novel Ion Channel-Targeted Peptides. Physiol Rev 84: 41-68.

Lisad

Andmestike näited

Lõppandmestik

Tabel: *Näidis autori koostatud andmestikust. Sisaldab informatsiooni, kas kordus leidus eksoni* $50ja100$ *nt. ümbruses ning mis oli korduse pikkus juhul, kui ta leidus 100 nt. ümbruses. Igal eksonil leidub kaks ümbrust. Kriipsuga on tähistatud ümbrused, mida polnud võimalik väljastada (soovitud ulatuses polnud teo genoom teada).*

ekson	kordus 50 nt. ümbruses	kordus 100 nt. ümbruses	korduse pikkus (nt.)
exon00001	1	1	11
exon00001	0	1	23
exon00002	0	0	0
exon00002	1	-	-
exon00003	0	0	0
exon00003	1	1	45
exon00004	1	0	0
exon00004	0	0	0
exon00005	1	1	10
exon00005	1	1	42
exon00006	0	1	15
exon00006	0	1	21
exon00007	1	1	17
exon00007	0	0	0
exon00008	0	0	0
exon00008	0	1	23
exon00009	1	0	0
exon00009	-	-	-
exon00010	1	1	10
exon00010	1	-	-

Genoomi andmed

Teo genoom on antud FastA formaadis tekstifailina. Failis on kirjas kontiigide nimed ja neile vastavad nukleotiidsed järjestused. Väikesed tähed tähistavad madalama kindlusega määratud nukleotiide. Näide vastavast andmestikust on toodud allpool.

```
>contig1324548 length=501   numreads=72
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGCAGTAGATCCAT
AATTATATACAGAGGTAGACAAATATGTCACTGCAAAAGATCCAAAGCCGATAtATATGA
TAAGATATACAGGAATTTCCTAGcGTGGTGCTTTCAAATTTaaagatagatagatagata
GATAGATAGATAGATAGATAGATGGGTTTGTCTGGGTTGGTTTtCATCGATCTTAACTGA
GGCAAAACTGGTCGTACAACAGCGACTGgATTACACCATCAcGCCTACCAgTCTATACAC
ACTCAAATCAGAATCGTGCgGCAGCTCCATGACAAAAGGGAACGCAACATCATGGTCATA
TCACCCTAGGAGTTAttaGCACAGTTGATTCGTATCATGAGCTCTCCACTCTTGATAGAG
GATGTATGTAGGCATTTGATGGTAGTTAAAGAGATGAAGTGGGCTTACGTAGCACAGTTG
ATACAAATTATGAGCTCTTTT

>contig1333836 length=509   numreads=77
aaaaaaaaaaaaaaaaaaaaaaaaaaaccAACCAACAGTCTGATGAACAAGACATAGCTA
AGAGAGACGAATAAAAAAGAGTCGAAGAAAACATACAAGATTAGCAACAAGAAAAaaaaa
GATTAACCAAAAAAAAAGAAAAAATCTGAATACATTTGAAAGAAAAAGAACAACAAAACA
ATTCAAGCCTGTAAATGTTTTTTAAAAAAAGGGCAAGCTACAAAATAACAGAAGTAAACT
TAAGTCAAATGATGAATTTGCAAAGTGAATAGCAAGACAGGAAAAAAAAACATGAAATAA
ATAAAATAAACAAAAGAAATAAACAACCAACCAACCCTGCCATGCATTTTGGTGAAGGAt
GTACTTCGGTGTCAAAACAGTAAAAATGATCTGTATTTCTGGTTCTTTCTGGGTTTTCTT
TTCAAAGAgAaTAAAATGATTCAGTATGTACTTAGGCTTCAAAGTAGTACTTAAATTATT
gCTGTTTGACAGCTCAAGAACGGTGGGTT
```

Eksonite koordinaatide andmed

Teo eksonite koordinaadid on antud tabuleeritud tekstifailina. Tabeli tulbad on järgmised:

1. Genoomse järjestuse fragmendi ID (nimi)
2. Genoomse järjestuse fragmendi pikkus
3. Transkriptoomi järjestuse nimi (mRNA)
4. Transkriptoomi järjestuse pikkus
5. Järjestuste joonduse (ühtiva ala) pikkus
6. DNA ahel (suund)
7. Joonduse alguskoordinaat genoomi järjestusel
8. Joonduse lõppkoordinaat genoomi järjestusel
9. Identsete nukleotiidide suhtarv joonduses
10. Joonduse skoor (sõltub joonduse pikkusest ja identsuse suhtarvust)
11. Joonduse E-väärtus

Lõppandmestiku kokkupanemiseks olid olulised tulbad numbritega 1, 7 ja 8. Lisaks kontrolliti joonduse sisulist korrektsust tulpade 9 ja 10 põhjal. Ülejäänud tulpade väärtused polnud antud töö seisukohalt olulised. Näide koordinaatide failist on toodud järgmisel leheküljel.

Näide koordinaatide failist:

contig642325	733	comp0_c0_seq1_Sample1	373	259	1	294	552	1.00	473	2.0e-131
contig642325	733	comp0_c0_seq1_Sample1	373	154	1	385	540	0.95	243	5.0e-62
scaffold32772	6961	comp1000028_c0_seq1_Sample1	232	232	1	6283	6514	0.97	385	5.0e-105
scaffold00104	69729	comp100002_c1_seq1_Sample1	416	415	-1	37920	38332	0.96	665	0.0e+00
scaffold42787	7011	comp100002_c1_seq1_Sample1	416	291	1	2117	2401	0.87	322	7.0e-86
scaffold15669	12972	comp100002_c1_seq1_Sample1	416	118	1	8188	8305	0.93	174	2.0e-41
scaffold05183	22059	comp100002_c1_seq1_Sample1	416	116	1	13032	13147	0.93	171	3.0e-40
scaffold43694	7285	comp100002_c1_seq1_Sample1	416	116	1	536	651	0.92	165	1.0e-38
scaffold66416	4394	comp100002_c1_seq1_Sample1	416	116	-1	3316	3431	0.91	159	6.0e-37
scaffold114249	2421	comp100002_c1_seq1_Sample1	416	118	1	963	1079	0.91	156	8.0e-36
scaffold19895	10540	comp100002_c1_seq1_Sample1	416	116	1	2980	3095	0.91	154	3.0e-35
contig1083730	620	comp100002_c1_seq1_Sample1	416	304	-1	65	388	0.76	143	6.0e-32
scaffold04803	25215	comp100002_c1_seq1_Sample1	416	115	1	3945	4058	0.89	139	8.0e-31
scaffold11466	14066	comp1000039_c0_seq1_Sample1	214	214	-1	10795	11009	0.99	385	4.0e-105
scaffold18320	12758	comp100005_c0_seq1_Sample1	487	465	-1	5674	6137	0.98	819	0.0e+00
scaffold15511	13930	comp1000065_c0_seq1_Sample1	350	338	1	3141	3492	0.94	521	6.0e-146
scaffold63226	4527	comp1000079_c0_seq1_Sample1	230	230	1	4141	4370	1.00	425	3.0e-117
contig764626	1111	comp100009_c0_seq1_Sample1	534	473	1	1	479	0.96	785	0.0e+00
contig1684230	393	comp100009_c0_seq1_Sample1	534	388	1	1	393	0.97	649	0.0e+00
contig2953033	217	comp100009_c0_seq1_Sample1	534	205	1	1	209	0.97	344	2.0e-92

Pythoni programmide koodid

Programm eksonite ümbruste väljastamiseks

```
   # Kasureale: koordinaatide fail, genoomi fail,
    # soovitud ymbruse pikkus, valjundfail
import sys
import re

coordinates = open(sys.argv[1])
Dict = {}
Dict2 = {}
# loeb koik read sonastikku, mille voti on
    # kontiigi nimi ja vaartus massiiv puntidest
entry = 0
for line in coordinates:
  line_split = line.strip().split()
  if line_split != []:
    # name - kontiigi nimi, beg - eksoni alguskoordinaat,
        # end - eksoni loppkoordinaat
    (name, beg, end) = (line_split[0], int(line_split[6]),
            int(line_split[7]))
    if name in Dict:
        Dict[name].append((beg, end))
    else:
        Dict[name] = [(beg, end)]
  entry += 1
coordinates.close()

# yhendab ylekattes olevate jarjestuste pundid yheks
def order(lst, algne):
    if len(lst) == 1:
        algne.append(lst[0])
        return algne
    maximum = 0
```

```
    for i in lst:
        if i[0] >= maximum:
            maximum = i[0]
            suurim = i
    lst.pop(lst.index(suurim))
    order(lst, algne)
    algne.append(suurim)
    return algne

Lst = list(Dict.keys())
for i in Lst:
    Lst2 = []
    temp = order(Dict[i], [])
    for j in temp:
        (beg_a, end_a) = (j[0], j[1])
        kattuv = 0
        for l in Lst2:
            if not (l[0] > end_a or l[1] < beg_a):
                kattuv = 1
        if not kattuv:
            for k in temp:
                    (beg_b, end_b) = (k[0], k[1])
                    if not(beg_b > end_a or end_b < beg_a):
                        beg_a = min(beg_a, beg_b)
                        end_a = max(end_a, end_b)
            Lst2.append((beg_a, end_a))
    Dict2[i] = Lst2


# Votab valja eksonite ymbrused
names = list(Dict2.keys())
genome = open(sys.argv[2])
n = int(sys.argv[3])
```

```
output = open(sys.argv[4], 'w')
k = 0
found = 0
seq = ""
for line in genome:
  if re.search(r'>', line, re.I):
    if seq != "":
        Tuples = Dict2[name]
        for i in range(len(Tuples)):
            Tuple = Tuples[i]
            beg = Tuple[0] - 1
            end = Tuple[1] - 1
            a = beg - n
            b = end + n
            Max = max(a, 0)
            if i > 0:
                end_prev = Tuples[i - 1][1]
                Max = max(Max, end_prev + 1)
            Min = min(len(seq) - 1, b)
            if i < len(Tuples) - 1:
                beg_next = Tuples[i + 1][0]
                Min = min(Min, beg_next - 1)
            out = ""
            out2 = ""
            if beg - Max == n:  # eksonist eespool
                out = seq[Max: beg ]
                if 'N' not in out:
                    output.write('>' + name)
                    output.write(out + "\n")
            if Min - end == n:  # eksonist tagapool
                out2 = seq[(end + 1): Min + 1]
                if 'N' not in out2:
                    output.write('>' + name)
```

```
                    output.write(out2 + "\n")
            seq = ""
            found = 0
        line = line.strip()
        name = line[1:].split()[0]
        if name in names:
            found = 1
      elif found:
        line = line.strip()
        seq += line

output.close()
genome.close()
```

Programm korduste pikkuste leidmiseks

```
import sys
import re

# kasureale: kordustega umbruste fail
FastA = sys.argv[1]
fasta = open(FastA)

asd = ""
line = fasta.readline()
while line != "":
  line = line.strip()
  i = 0
  while i < len(line):
    if line[i] == "N":
      j = 0
      while i < len(line) and line[i] == "N":
        j += 1
        i += 1
      print(str(j))
    else:
      i += 1
  line = fasta.readline()

fasta.close()
```

R kood Fisheri täpse testi näite jaoks

```
tabel = rbind(c(2, 1, 6), c(9, 3, 1))
tulpSum = colSums(tabel)
ridaSum = rowSums(tabel)
kokku = sum(tulpSum)

toen1 = choose(tulpSum[1], tabel[1,1])*choose(tulpSum[2], tabel[1,2])*
            choose(tulpSum[3], tabel[1,3])/ choose(kokku, ridaSum[1])

p = 0 #p-vaartus
loendur = 0 #sobivate tabelite arv
loendur2 = 0  #koikide tabelite arv
for (i in 0:tulpSum[1]){
  for (j in 0:tulpSum[2]) {
    for (h in 0:tulpSum[3]) {
      loendur2 = (loendur2 + 1)
      toen = choose(tulpSum[1], i)*choose(tulpSum[2], j)*
                choose(tulpSum[3], h)/choose(kokku, i+j+h)
      if (toen <= toen1 && i+j+h==ridaSum[1]) {
        loendur = (loendur + 1)
        p = (p + toen)
      }
    }
  }
}
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina _____Maarja Lepamets_____
(*autori nimi*)
(sünnikuupäev: _____23. detsember 1990_____)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Conus consors'i konopeptiidide geene ümbritsevate kordusjärjestuste statistiline analüüs_____,
(*lõputöö pealkiri*)

mille juhendajad on _____Maido Remm ja Märt Möls_____,
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus **06.05.2013**