TARTU ÜLIKOOL

LOODUS- JA TEHNOLOOGIATEADUSKOND

MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT

BIOINFORMAATIKA ÕPPETOOL

Märt Roosaare

Valgudomeenide analüüs koonusteos *Conus consors*

Magistritöö bioinformaatikas

Juhendaja: professor Maido Remm

Tartu 2014

Sisukord

Töös kasutatud lühendid:

aa (*amino acid*) – aminohappejääk

GPCR (*G-protein coupled receptor*) – G-valgu seoseline retseptor

HMM (*hidden Markov model*) – varjatud Markovi mudel

LDL-A (*low-density lipoprotein receptor domain type A*) – madala tihedusega lipoproteiini retseptori domeen, tüüp A

LDL-A-LGR (*low-density lipoprotein receptor domain type A-containing leucine-rich repeat-containing G-protein coupled receptor*) – madala tihedusega lipoproteiini retseptori tüüp A domeene ja leutsiinirikkaid kordusi sisaldav G-valgu seoseline retseptor

LGR (*Leucine-rich repeat-containing G-protein coupled receptor*) – leutsiinirikkaid kordusjärjestusi sisaldav G-valgu seoseline retseptor

LRR (*leucine-rich repeat*) – leutsiinirikas kordusjärjestus

MSA (*multiple sequence alignment*) – mitme järjestuse joondus

OR (*odds ratio*) – šansside suhe

ORF (*open reading frame*) – avatud lugemisraam

R7TM (*rhodopsin-like 7-transmembrane receptor*) – rodopsiini-sarnane 7-transmembraanne retseptor

TM (*transmembrane*) – transmembraanne

Sissejuhatus

Antud töös kasutati Euroopa Liidu FP6 teadusprojekti "CONCO, *the cone snail genome project for health*" (LSHB-CT-2007-037592) raames sekveneeritud koonusteo *Conus consors* genoomseid ja transkriptoomseid järjestusi. Projekti peamiseks eesmärgiks oli *C. consors'*i peptiidsete toksiinide baasil uute ravimikandidaatide väljatöötamine. Projekti raames sekveneeriti ka nimetatud liigi genoom ning transkriptoom kaheksast erinevast koest. Meie töörühm tegeles teo sekveneeritud genoomi assambleerimise ja selle edasise analüüsiga.

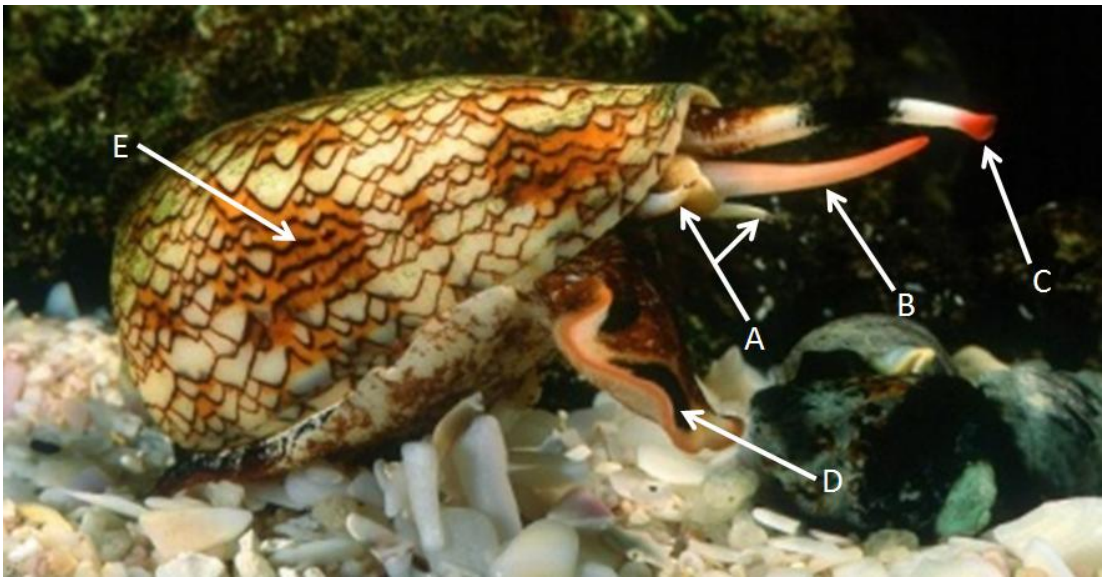
C. consors kuulub mürgiste, soojades meres elavate tigude perekonda *Conus*, mille esindajad on aktiivsed jahipidajad, kes saaki püüdes paralüseerivad selle miniatuursete mürgiste harpuunide abil. Nende peptiidsed neurotoksiinid – konopeptiidid – toimivad ioonkanalite blokaatoritena ning neid iseloomustab väga kiire toime, suur varieeruvus nii ühe liigi piires kui ka liikide vahel ning kõrge spetsiifika märklaua suhtes (Olivera jt., 1999). Tänapäeval kasutatakse konopeptiide peamiselt uute ravimite väljatöötamisel.

Molluskeid, mis on väga suur ja mitmekesine hõimkond, on lisaks *C. consors'*ile siiani sekveneeritud vaid neli liiki: pärlikarp *Pinctada fucata*, auster *Crassostrea gigas* ja teod *Lottia gigantea* ning *Aplysia californica* (merijänes). Molluskite genoomsete järjestuste hulk on võrreldes imetajatega oluliselt piiratum. Käesoleva magistritöö põhieesmärgiks oli välja selgitada, mis eristab *C. consors'*it teistest sekveneeritud molluskitest valgudomeenide tasemel, ning analüüsis keskenduda eelkõige neile domeenidele, mis on *C. consors'*is teiste molluskitega võrreldes suhteliselt rohkem esindatud. Esialgse analüüsi tulemuste põhjal seati peamiseks eesmärgiks rodopsiini-sarnaste 7-transmembraansete retseptorite täpsem analüüs *C. consors'*is – millised alamrühmad on koonusteos oluliselt rohkem esindatud kui teistes sekveneeritud molluskites keskmiselt ning milliseid bioloogilisi funktsioone need täita võiksid.

KIRJANDUSE ÜLEVAADE

1. Koonustigu *C. consors* ja teised sekveneeritud molluskid

Conus consors kuulub molluskite hõimkonna suurima klassi tigude (*Gastropoda*) hulka, suurde ja liigirikkasse perekonda *Conus*. Selle esindajad elavad soojades meredes ning jahivad väikseid kalu, molluskeid ning isegi teisi *Conus* perekonna liike. Saagi jahtimisel on *C. consors'*ile abiks teiste tigudega võrreldes tundlikum ja keerukam lõhnaelund (*osphradium*), mille abil ta keskkonnast tulenevaid keemilisi signaale detekteerib (Spengler ja Kohn, 1995). Kui suurel osal maismaatigudest on toidu purustamiseks hõõrel, kuhu kinnituvad kitiinist hambakesed, siis koonustigudel on hambad lahtised ja tigu kasutab neid saagi pihta „tulistamiseks". Jahti pidaval koonusteol on välja sirutatud lont (B joonisel 1), milles on laskevalmis mürgine harpuun. Sellega halvatakse lähidistantsilt saak, tõmmatakse see londi abil suhu ning neelatakse alla. Koonustigude mürk koosneb paljudest erinevatest toksiinidest, ühe liigi piires 50-200 erinevast toksiinist (Olivera jt., 1999). Sellise suure varieeruvuse üheks põhjuseks on ilmselt iga toksiini väga spetsiifiline mõju märklaudorganismile (enamik toksiine toimivad ioonkanalite blokaatoritena). Mürk on väga kiire toimega ja suuremad *Conus'*e liigid võivad olla eluohtlikud ka inimesele.



Joonis 1. Koonusteo üldanatoomia *Conus textile* näitel. Koonusteo üheks oluliseks tunnetuslikuks elundiks on sifoon (C), mille abil tigu ümbritsevast keskkonnast vett filtreerib ning seal leiduvaid keemilisi signaale püüab. Lont (B) on tähtis jahipidamisel – selle abil toimub „harpuunide tulistamine" saagi pihta ja saagi kinnihoidmine, kuni see alla neelatakse. Lihaseline jalg (D) on vajalik liikumiseks mööda merepõhja ning koda (E) on kaitseks pehmele kehale. Silmade (A) kohta ei ole teada, kuivõrd funktsionaalsed need on - saagi püüdmisel need olulist tähtsust ei oma, kuna tigu suudab saagi tabada ka liiva sisse kaevunult (foto: David Paul ja Bruce Livett, Melbourne ülikool).

Molluskeist on siiani sekveneeritud vaid mõned üksikud liigid – *Aplysia californica* (Remm jt., 2014), *Lottia gigantea* (Simakov jt., 2013), *Pinctada fucata* (Takeuchi jt., 2012) ja *Crassostrea gigas* (Zhang jt., 2012). Sekveneeritud on ka üks teine koonusteo liik - *Conus bullatus* (Hu jt., 2011), kuid saadud andmed olid väga madala kvaliteediga (3x katvus, Illumina metoodika) ja seepärast neid käesolevas töös ei kasutatud.

Tabel 1. Seni sekveneeritud molluskite genoomid (Remm jt., 2014 põhjal).

Liik	Genoom (Mbp)	Katvus	Assambleeritud järjestuse pikkus (Mbp)	Kasutatud tehnoloogia
Conus consors	3000	19x	2312	Roche 454, Illumina
Aplysia californica	1800	10x	716	Sanger
Lottia gigantea	500	9x	360	Sanger
Pinctada fucata	1150	40x	1413	Roche 454, Illumina
Crassostrea gigas	600	690x	559	Illumina

2. 7-transmembraansed G-valgu seoselised retseptorid

2.1 Ülevaade

Eukarüootses organismis peavad selle rakud toimima koos ühtse tervikuna ja selleks on aegade jooksul välja arenenud keerulised kontrollsüsteemid. Nende efektiivseks tööks on aga hädavajalik rakkude suhtlus nii omavahel kui ka väliskeskkonnaga, et tunnetada nii organismisiseseid kui ka keskkonnast tulenevaid muutusi. Erinevate signaalide (peamiselt keemiliste) vastuvõtmiseks on rakkudel spetsiifilised valgud, mida kutsutakse retseptoriteks. Signaalmolekule, mis retseptoritele seonduvad, on väga erinevaid. Nendeks võivad olla neurotransmitterid, peptiidid, hormoonid, glükoproteiinid ja isegi footonid. Retseptorid võivad paikneda nii raku pinnal, nagu suurem osa neist, kui ka raku sees. Raku pinna retseptorite suurimateks alamklassideks on ioonkanal-retseptorid, katalüütilised retseptorid ja 7-transmembraansed G-valgu seoselised retseptorid (GPCR), millele keskendub käesolev töö.

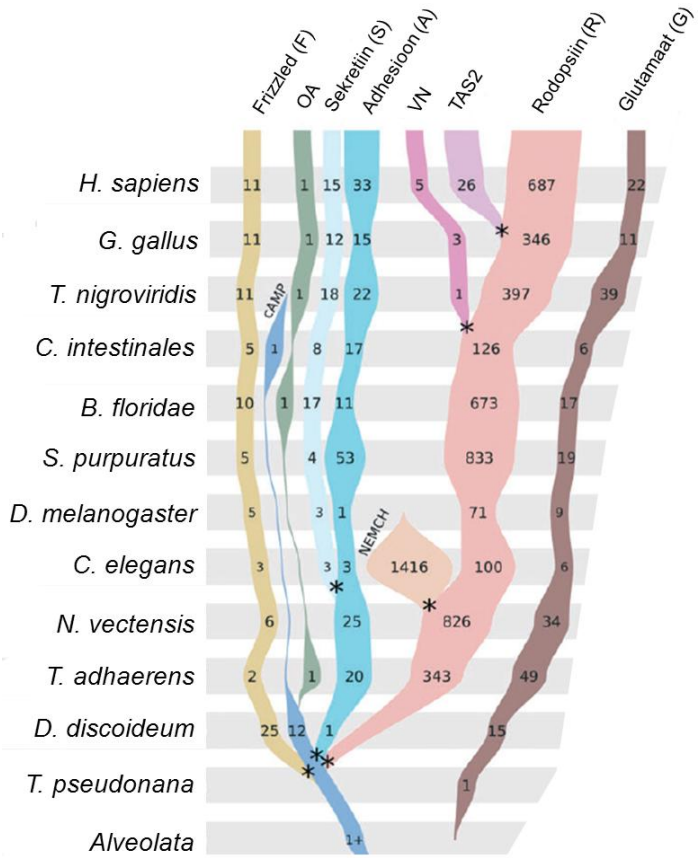
GPCR-id on inimese genoomis üks suuremaid ja mitmekesisemaid valguperekondi, mis mõjutavad väga erinevaid rakulisi protsesse nagu sekretsioon, metabolism, närvisignaalide ülekanne, rakkude diferentseerumine ja põletiku teke (Herbert ja Bouvier, 1998) ning neil on tähtis roll ka keskkonna tunnetamisel. GPCR-ide abil töötavad nägemis- ja haistmismeel ning osa maitsmismeelest.

GPCR-ide ligandid varieeruvad suurel määral – nendeks võivad olla nii footonid, ioonid, neurotransmitterid, lipiidid, aminohapped kui ka peptiidid (Bockaert ja Pin, 1999) ning seetõttu erinevad GPCR-id üksteisest küllaltki palju ka aminohappelise (aa) järjestuse tasemel – isegi alamperekondade siseselt võivad erinevate retseptorite järjestuste identsused olla alla 15%. (Chabbert jt., 2012). GPCR-ide väga suur mitmekesisus ja laialdane mõju rakus toimuvatele protsessidele teeb neist väga hea märklaua ravimitele – umbes pooled uuel aastatuhandel kasutusse lubatud ravimitest mõjutavad just GPCR-e (Klabunde ja Hessler, 2002). GPCR-ide uurijaid on tunnustatud ka kõrge teaduspreemiaga: 2012. aastal olid keemia valdkonnas Nobeli preemia laureaatideks Robert Lefkowitz ja Brian Kobilka, kelle töörühmad on pikka aega tegelenud just GPCR-ide tööpõhimõtete ja funktsioonide uurimisega.

GPCR-idel, vaatamata nende suurele heterogeensusele, on ka mitmeid ühiseid struktuurseid tunnuseid, neist põhiliseks on polüpeptiidahelas seitsme 25-35 aa pikkuse transmembraanse α-heeliksi olemasolu. Ahela N-terminaalne osa paikneb rakust väljas, C-terminaalne raku sees ning vahepeal läbib ahel membraani seitse korda, moodustades kolm ekstratsellulaarset ja kolm tsütoplasmaatilist lingu (Schiöth ja Fredriksson, 2005a). Üldjoontes on sarnane ka GPCR-ide tööpõhimõte: ligandi seostumisel GPCR-ga muutub retseptori konformatsioon - see aktiveeritakse ning sellega seostub mitteaktiivne G-valk, millega on seotud guaniindifosfaat. Aktiivne GPCR toimib G-nukleotiidi vahetusfaktorina ning G-valgus vahetub guaniindifosfaat guaniintrifosfaadi vastu ja seeläbi aktiveeritud G-valk mõjutab rakus mitmeid erinevaid bioloogilisi protsesse (Davies jt., 2007). Sellest tuleneb ka GPCR-ide nimi - nende toime avaldub läbi G-valkude.

2.2 Klassifikatsioon

GPCR-ide suur hulk ja heterogeensus teeb nende süstematiseerimise keeruliseks. Inimese genoomi sekveneerimine 2001. aastal (Lander jt., 2001) võimaldas leida genoomist kõikide GPCR-ide DNA-järjestused ja kasutada fülogeneetilisel analüüsil põhinevat klassifitseerimist. Sellel põhineb 2003. aastal inimese GPCR-ide põhjal koostatud GRAFS-süsteem (joonis 2), mis jagab GPCR-id viieks peamiseks põhiklassiks: glutamaat (G), rodopsiin (R), adhesioon (A), Frizzled (F) ja sekretiin (S) (Fredriksson jt., 2003). Mitmed hilisemad tööd on näidanud, et GRAFS sobib hästi ka teistest organismidest pärit GPCR-ide klassifitseerimisks (Schiöth ja Fredriksson, 2005b, Gloriam jt., 2007, Kamesh jt., 2008, Chabbert jt., 2012).



Joonis 2. GPCR-ide fülogeneesipuu ja GRAFS põhiklassid. Rodopsiin (R), Frizzled (F), sekretiin (S) ja adhesioon (A) klassi GPCR-id on lahknenud cAMP retseptorite perekonnast, nagu ka *Ocular albinism* tüüp 1 retseptor (OA). Glutamaadi (G) perekonna retseptorid on evolutsioneerunud eraldiseisva rühmana. Rodopsiini perekonnast on hiljem lahknenud nematoodide kemoretseptorite perekond (NEMCH), vomeronasaalsed retseptorid (VN) ja Taste2 mõru maitse retseptorid (TAS2). Fülogeneesipuu harudel on näidatud ka vastavat tüüpi retseptorite arv igas organismis (Nordström jt., 2011 põhjal).

3. Rodopsiini-sarnased 7-transmembraansed retseptorid

3.1 Ülevaade

Suurima grupi selgroogsetes esinevate GPCR-ide superperekonnast moodustavad rodopsiini-sarnased 7TM retseptorid (R7TM). Neid iseloomustab mitmekesine ligandide hulk, väga suur varieeruvus valgujärjestuses (omavahel võrreldes võib identsus olla alla 15%) ja mõningad väga konserveerunud motiivid, nagu näiteks N-P-x-x-Y-motiiv seitsmendas TM-heeliksis ja D(E)-R-Y(F)-motiiv kolmanda TM-heeliksi ja teise rakusisese lingu piiril (Schiöth ja Fredriksson, 2005b, Fredriksson jt., 2003). Teistest GPCR perekondadest eristab neid ka N-terminaalse osa struktuur – kui ülejäänud perekondadel on see pikk ning võib sisaldada mitmeid erinevaid domeene, siis R7TM on see üldjuhul lühike ja ligandi seondumissaidiks on transmembraansete heeliksite vahele jääv ruum (Nygaard jt., 2009,

Granier ja Kobilka, 2012). Eranditeks on glükoproteiinseid hormoone siduvad retseptorid, nagu näiteks inimese relaksiini, lutropiini, follitropiini ning türeotropiini retseptorid, mille ligandi siduvaks osaks on just pikk, mitmeid domeene sisaldav valgu N-terminaalne osa (Schiöth ja Fredriksson, 2005b).

R7TM-ide tööpõhimõtete täpsemaks uurimiseks oli väga oluline samm nende kolmedimensionaalse struktuuri kindlaksmääramine, mis võimaldas analüüsida ligandide seondumissaite ja interaktsioone. Väga oluline läbimurre R7TM uurimises toimus 2000. aastal, kui K. Palczewski töögrupp suutis lõpuks kristalliseerida veise rodopsiini ning saadi esimene kõrge resolutsiooniga R7TM kolmedimensionaalne struktuur (Palczewski jt., 2000). R7TM ja G-valgu omavaheliste interaktsioonide analüüsi aitas oluliselt edasi viia β-2 adrenergilise retseptori ja sellega seondunud G-valgu kompleksi kristallstruktuur, mille põhjal saadi esmakordselt eksperimentaalselt kindlaks määrata G-valgu retseptoriga seostumise iseärasused (Rasmussen jt., 2011).

3.2 Klassifikatsioon

Võttes aluseks inimeses kirjeldatud R7TM-id, mida on ligikaudu 700 (Gloriam jt., 2007, Chabbert jt., 2012) ning toetudes nende fülogeneetilisele analüüsile, jaotati R7TM-id nelja põhiklassi ja kolmeteistkümne alamklassi (Fredriksson jt., 2003; Schiöth ja Fredriksson, 2005a). Inimese puhul moodustab väga suure osa R7TM-de hulgast lõhnaretseptorite alamklass (tabel 2). Loomariigi piires on R7TM-ide alamklassid esindatud väga erinevalt – mõned, evolutsiooniliselt ilmselt vanemad klassid nagu PEP, AMIN, LGR ja SOG esinevad ka üksteisest väga kaugetes liikides nagu näiteks *C. elegans* ja inimene. Teisi, evolutsiooniliselt ilmselt hiljem lahknenud retseptorite klasse on leitud vaid vähestes organismirühmades, nagu näiteks klass MRG, mida on siiani kirjeldatud vaid imetajates ja lindudes (Chabbert jt., 2012).

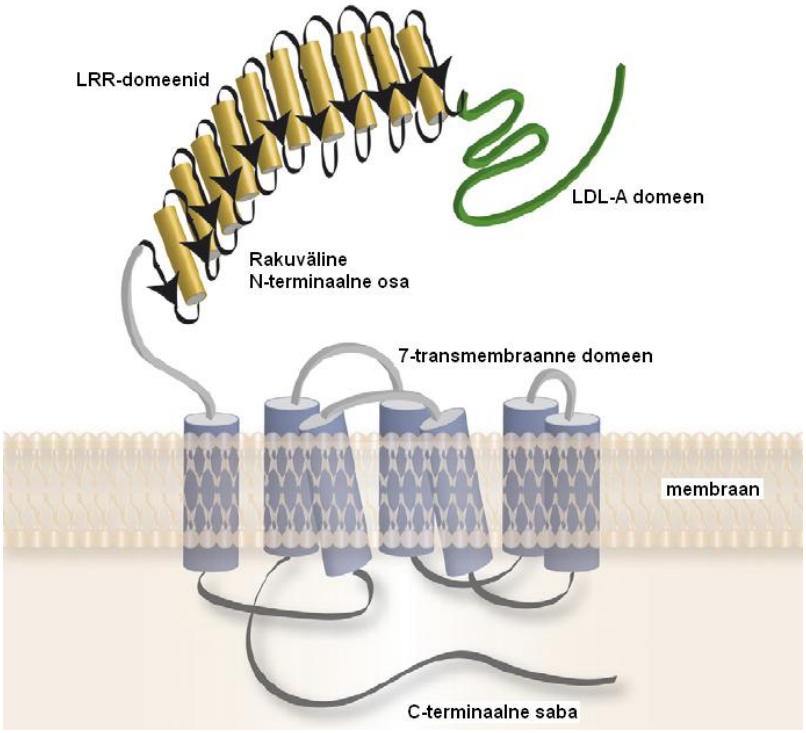
Tabel 2. Rodopsiini-sarnaste GPCR-ide alamklassid inimeses (Chabbert jt., 2012 põhjal).

Klass	Alamklass	Arv inimeses	Peamised retseptorid
α	AMIN	44	biogeeniliste amiinide retseptorid (serotoniin, dopamiin, histamiin...)
	MEC	18	melanokortiini ja kannabinoidide retseptorid ning adenosiine siduvad retseptorid
	MTN	3	melatoniini retseptorid
	OPN	11	opsiinid (kepikeste ja kolvikeste pigmendid)
	PTGR	13	prostaglandiini retseptorid
β	PEP	43	peptiide siduvad retseptorid (oksütotsiini, neuromediini, neuropeptiidide...)
γ	CHEM	43	kemokiinide, bradükiniini ja angiotensiini retseptorid
	MCHR	1	melaniini kontsentreeriva hormooni retseptor
	SOG	10	somatostatiini ja galaniini retseptorid
δ	LGR	7	glükoproteiinsete hormoonide (relaksiin, lutropiin, follitropin, türeotropiin) retseptorid
	MRG	7	MRG ja MAS-retseptorid
	OLF	494	lõhnaretseptorid
	PUR	35	puriini retseptorid
		20	klassifitseerimata

3.3 Rodopsiini-sarnaste GPCR-ide ülesehitus valgudomeenide tasemel

Valdava enamiku R7TM struktuuris on vaid üks kompaktne, seitsmest TM-ahelast koosnev domeen, millega seonduvad ka ligandid, sest valguahela rakuväline N-terminus on lühike ja ei sisalda senikirjeldatud funktsionaalseid domeene (Lagerström ja Schiöth, 2008). Inimeses ja paljudes teistes organismides on eranditeks LGR-alamklassi retseptorid, mille koosseisus esinevad ka leutsiinirikkad kordusjärjestused (*leucine-rich repeats*) (LRR) ning relaksiini retseptorite (joonis 3) puhul lisaks ka madala tihedusega lipoproteiini retseptori klass A domeen (LDL-A) (van der Westhuizen jt., 2008). Teiste R7TM alamklasside struktuuris muid domeene peale 7TM leitud ei ole, vähemalt mitte inimese puhul. Uute organismide sekveneerimine on andnud ka teistsuguseid kombinatsioone, näiteks meritupes *Ciona intestinalis* (Kamesh jt., 2008) on leitud LRR ja LDL-A domeene koos 7TM-osaga imetajatest erinevates kombinatsioonides – näiteks potentsiaalsed R7TM-id, milles lisaks 7TM osale ainult LDL-A domeenid ilma LRR-ideta. R7TM koosseisus on LRR ja LDL-A domeene leitud ka teos – mudakuke *Lymnea stagnalis*e (Tensen jt., 1994) puhul on arvatud, et sellised retseptorid võivad osaleda lipoproteiinsete signaalide ülekandel kesknärvisüsteemi. Merisiilikus *Strongylocentrotus purpuratus* on leitud ka seni lähemalt kirjeldamata kombinatsioon, kus R7TM-s esinevad veel lisaks LRR ja LDL-A-le ka CUB-domeenid

(UniProtKB/TrEMBL andmebaas, valgud H3IE72_STRPU, H3JAK8_STRPU, H3I0L3_STRPU), kuid selliste valkude reaalne eksisteerimine on veel isegi transkripti tasemel tõestamata ning nende funktsioonid ei ole teada.



Joonis 3. Inimese klass C LGR-retseptori skeem. Inimeses kuuluvad nende hulka relaksiini retseptorid 1 ja 2. Nagu kõigil GPCR-idel, on ka nende struktuuri põhiosaks 7-transmembraanne domeen (sinine). Erinevalt teistest rodopsiini-sarnastest 7TM-retseptoritest, mille N-terminaalne osa on lühike ja ei sisalda ühtki domeeni, on relaksiini retseptoritel pikk ekstratsellulaarne osa, mis koosneb 9-10-st LRR-domeenist (kollased) ning N-terminuses olevast LDL-A domeenist (roheline) (van der Westhuizen jt., 2008).

4. Genoomipõhise fülogeneetilise analüüsi metoodika

4.1 Valgudomeenid

Antud töös kasutatud *C. consors*'i assambleeritud genoom oli väga fragmenteerunud (assambleeritud genoomi suurus 2312 Mbp, 4,5 miljonit kontiigi, N50 = 819 bp), mistõttu oli analüüsis võimalik kasutada vaid eksoneid, mitte täispikki geene. Sellest tulenevalt otsustati fülogeneetiline analüüs teha valgudomeenide põhjal. Valgudomeenid on üheks valkude tertsiaarstruktuuri põhiosaks – nad on kindlalt piiritletud, kompaktsed struktuurid, mis on võimelised iseseisvalt funktsioneerima, evolutsioneeruma ja voltuma. Domeenide puhul on tähtis just sõltumatus ühest kindlast valgust: üks ja sama domeen võib esineda nii paljudes erinevates valkudes kui ka mitme koopiana ühes ja samas valgus ning lisaks ka erinevates

kombinatsioonides koos teiste domeenidega. Sellise plastilisuse tõttu on neid nimetatud ka valkude looduslikeks „ehituskivideks“ (Ponting ja Russell, 2002).

4.2 Valgudomeenide modelleerimine ja Pfam andmebaas

Valgudomeenide modelleerimiseks koondatakse ühtse evolutsioonilise päritoluga domeenid neid kirjeldavasse profiili, mis luuakse mitme järjestuse joonduse (MSA) põhjal ja kujutab endast positsiooni-spetsiifilist hindamismaatriksit. Väga oluline on siin just hindamise sõltuvus positsioonist – paarikaupa joondamises kasutatavad hindamismaatriksid nagu PAM ja BLOSUM ei arvesta järjestuse spetsiifilisi osi (Henikoff 1996), profiilid aga küll. Sellest tulenevalt on profiilid evolutsiooniliselt kaugemate järjestuste detekteerimisel tundlikumad kui paarikaupa joondamine (Sonnhammer jt., 1997). Profiilidel põhineva metoodika üks oluline puudus on nende koostamise keerukus ja raskesti kirjeldatavad hindamissüsteemid. Kui positsiooniliselt sõltumatuid hindamissüsteeme nagu PAM on hästi kirjeldatud ning need on üsnagi universaalsed (Altschul ja Gish, 1996), siis profiilide puhul on hindamissüsteemi loomine keerulisem. Üldjuhul tehti see iga profiili iseärasustele vastavalt, kuid sellistel süsteemidel puudus tihtipeale toetav teooria ja kirjeldus, kuidas ja miks nad töötavad ning mille alusel nad loodi. Lahendusena võeti profiilide puhul aluseks varjatud Markovi mudelite (HMM) teooria, mis on matemaatiliselt hästi kirjeldatud (Eddy, 1998).

Valgujärjestuste analüüsi neis esinevate domeenide põhjal lihtsustas oluliselt 1997. a. loodud Pfam andmebaas (Sonnhammer jt., 1997), mis erines teistest kaasaegsetest eelkõige selle poolest, et perekondadesse jaotamise ja andmebaasist tehtavate otsingute aluseks olid võetud terved domeenijärjestusted, mitte ainult neis esinevad lühikesed konserveerunud motiivid. Domeeniperekondade kirjeldamiseks kasutati HMM-e, mis lubasid arvestada ka võimalike insertsioonide ja deletsioonidega ning ühtlasi kasutada järjestuste analüüsil HMMER-tarkvara (Eddy, 2011). Pfam andmebaas jaguneb üldjoontes kaheks – kõrgekvaliteediline, spetsialistide poolt üle vaadatud Pfam-A ja automaatselt, algoritmi abil klasterdatud perekondadega Pfam-B. Uue Pfam-A perekonna loomine koosneb järgmistest põhietappidest:

- UniProtKB andmebaasist valitakse hoolikalt loodavat domeeniperekonda kõige paremini iseloomustavad valgujärjestused ning tehakse neist MSA (*seed alignment*).
- HMMER-paketiga ehitatakse saadud joondust kasutades uut perekonda kirjeldav HMM.

- Saadud HMM-i abil otsitakse UniProtKB andmebaasist kõik valgujärjestuste osad, mis ületavad teatavat bit-skoori väärtust ning need loetakse antud domeeniperekonna alla kuuluvaiks.

Aja jooksul on Pfam'i andmebaasi pidevalt edasi arendatud ja sinna uusi domeene lisatud, antud töös kasutati Pfam-A versiooni 26.0 (Punta jt., 2012).

4.3 Valgudomeenide fülogeneetiline analüüs

Suurema grupi homoloogsete valgujärjestuste täpsemaks fülogeneetiliseks analüüsiks kasutatakse fülogeneesipuid, mille ehitamiseks on omakorda vajalik MSA olemasolu. MSA loomise käigus võrreldakse korraga kõiki sama grupi valke ning see aitab tuvastada ka erinevaid mutatsioone – asendusi, insertsioone, deletsioone ja ka järjestuste konserveerunud blokke ning motiive, mis viitavad üldjuhul funktsionaalsetele osadele. MSA koostamiseks saab teoreetiliselt kasutada ka täpseid dünaamilise programmeerimise meetodeid, kuid järjestuste arvu suurenedes kasvab ajakulu eksponentsiaalselt ja suurte andmemahtude puhul ei ole selline lähenemine praktiliselt võimalik. Seetõttu on MSA-de tegemisel rakendatud eelkõige heuristilisi meetodeid nagu progressiivne joonduse koostamine, mille puhul määratakse algul järjestuste omavaheline sarnasus ja seejärel alustatakse joondust kõige sarnasemast järjestuste paarist ning lõpetatakse kõige kaugemalt seotutega. Sellise lähenemise plussiks on suur kiirus, kuid see ei anna tihti väga head tulemust – iga järjestus joondatakse algoritmi töö käigus lõplikult, selle sobivust MSA edasisel ehitamisel enam teistkordselt ei analüüsita. Seda suunda esindavad näiteks varasemad Clustal-perekonna programmid (Higgins ja Sharp, 1988). Progressiivse joonduse edasiarenduseks on iteratiivsed meetodid nagu MUSCLE (Edgar, 2004), mis analüüsivad MSA loomise käigus ka juba olemasolevat joondust ning vajadusel optimiseerivad seda, andes täpsema lõpptulemuse. Väga suurte MSA-de loomiseks sobivad HMM-idel põhinevad meetodid, mille eeliseks on suur kiirus – järjestusi võrreldakse ainult ühe kindla HMM-iga, mitte omavahel paarikaupa (Price jt., 2009).

Suuremahulistest MSA-dest fülogeneesipuude ehitamiseks kasutatakse peamiselt metoodikat, mis põhineb järjestuste omavahelise evolutsioonilise kauguse hindamisel, sest fülogeneesipuu ehitamine suurima tõepära meetodil on palju ajamahukam ja samas mitte oluliselt täpsem (Price jt., 2009). Ka tavapärased evolutsioonilisel kaugusel põhinevad meetodid on suurte andmehulkade puhul aeglased, kuid mitmete heuristikute kasutamisega (võimalike kandidaatide hulga vähendamine puu harude ühendamisel; lahknemiste tõepära kontrolliks ei

arvutata välja täies mahus alternatiivseid fülogeneesipuid) on neid oluliselt kiiremaks muudetud ning loodud programme, näiteks FastTree (Price jt., 2009), mis võimaldavad ka väga suurtest joondustest minutite jooksul fülogeneesipuu ehitada.

4.4 Rodopsiini-sarnaste GPCR-ide otsimine genoomidest bioinformaatiliste meetoditega

Tänu aasta-aastalt üha suurenevale, vabalt kättesaadavale infohulgale, mis on avalikes bioloogiliste järjestuste andmebaasides nagu GenBank, Ensembl ja UniProtKB ning teaduspublikatsioonides, muutub uute teadusavastuste tegemisel üha tähtsamaks olemasoleva info leidlik kasutamine. 2003. aastal inimese genoomis esinevate R7TM-ide täielikku repertuaari koostades (Fredriksson jt., 2003) alustati just kirjanduse läbitöötamisega, et koondada kogu teadaolev info R7TM-idest ning selle põhjal otsiti GenBanki andmebaasist BLAST-i abil välja kõik sarnased järjestused, mis võiksid potentsiaalselt olla R7TM-id. Saadud andmed klasterdati fülogeneetilise analüüsi abil ning iga klassi kohta käiv info koondati neid kirjeldavasse HMM-mudelisse ning korrati otsingut ka saadud mudelitega, mis andis veel uusi potentsiaalseid R7TM-sid.

Hilisemates töödes on kasutatud üldjoontes sama lähenemist (Gloriam jt., 2007, Kamesh jt., 2008), kuid oluliselt on arenenud bioinformaatiline pool – R7TM erinevate alamklasside kohta käiv info on suures osas juba HMM-idesse koondatud ja Pfam andmebaasis arhiveeritud ning HMM-ide kasutamiseks on olemas kiire tarkvara eesotsas HMMER-paketiga. Fülogeneetiliste analüüside tegemine on tänu protsessorite kiiruse kasvule ja heuristiliselt täiustatud algoritmidel põhinevate programmide kasutamisele kiirenenud kümneid kordi, võimaldades analüüsida korraga järjest suuremaid andmehulki ja saada ülevaatlikumaid tulemusi.

PRAKTILINE OSA

1. Töö eesmärgid

- Leida rodopsiini-sarnaste retseptorite alamrühmad, mis on *C. consors*'is, võrreldes teiste analüüsitud nelja molluski keskmisega, oluliselt rohkem esindatud ning selgitada välja nende võimalik bioloogiline funktsioon koonusteos.

2. Metoodika

2.1 Genoomsete järjestuste kogumine ja transleerimine

HMMER3 paketi (http://hmmer.janelia.org/) (Eddy, 2011) alamprogramm hmmsearch, mida antud töös domeenide otsimiseks kasutati, võimaldas sisendina kasutada vaid valgujärjestusi, seega oli esimeseks sammuks domeenide otsimisel genoomide transleerimine. Kuna *C. consors'i* assambleeritud genoomijärjestus (Remm jt., 2014) oli väga fragmenteerunud ja kodeerivate eksonite täpne määramine komplitseeritud, siis otsustati transleerida kogu genoomijärjestus kuues raamis, leida potentsiaalsed stoppkoodonid ja nende põhjal avatud lugemisraamid (ORF), miinimumpikkusega 150 aa. Optimaalse miinimumpikkuse otsimisel katsetati ka madalamaid väärtusi, mis ei tõstnud oluliselt leitud domeenide hulka, küll aga suurendasid kordades andmemahtu. ORF-ideks loeti kõik stoppkoodonite vahelised lõigud, mis olid vähemalt miinimumpikkusega. Sama metoodikat kasutati ka teiste nelja molluski (*Aplysia californica*, *Pinctada fucata*, *Crassostrea gigas* ja *Lottia gigantea*) ning kolme referentsorganismi (*C. elegans*, *D. melanogaster* ja *H. sapiens*) genoomide puhul, et tulemused oleksid võrreldavad. Genoomid saadi järgmistest allikatest: Ensembl'i andmebaasist *Caenorhabditis elegans* (WBcel215.69), *Drosophila melanogaster* (BDGP5.69) ja *Homo sapiens* (GRCh37.69). Teistest võrgus olevatest allikatest *Aplysia californica* (ver. 2.0: http://www.broadinstitute.org/ftp/pub/assemblies/invertebrates/aplysia/), *Lottia gigantea* (ver. 1.0: http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html), *Crassostrea gigas* (http://public-contigbrowser.sigenae.org:9090/Crassostrea_gigas/index.html) ja *Pinctada fucata* (ver. 1.0: http://marinegenomics.oist.jp/genomes/downloads?project_id=20).

2.2 Transleeritud genoomiandmete analüüs HMMER3 paketi abil

ORF-idest valgudomeenide leidmiseks kasutati HMMER3 paketi alamprogrammi hmmsearch (parameetrid Z=1 000 000, domZ=5 000, E=0,00001, domE=0,00001), et leida valgujärjestuste osad, mis on vastavat domeeni kirjeldava HMM-iga sarnased. Otsingu käigus

võrreldi leitud ORF-e kõikide Pfam-A (versioon 26.0, november 2011, 13 672 erinevat domeeniperekonda kirjeldavat HMM-i) (Punta jt., 2012) andmebaasis olevate mudelitega. Kui ORF-ides tuvastati mitu üksteisega rohkem kui 20 aa pikkuses kattuvat domeeni, arvestati vaid seda, millel oli väikseim e-väärtus. Hmmsearch'i vastetele seati ka maksimum- ja miinimumpikkused, mis olid vastavalt 50% ja 150% vastava HMM-i pikkusest. Miinimumpikkuse määramine aitas vältida olukorda, kus suuri domeene, mis võivad esineda mitmes eksonis, loetakse mitmekordselt.

2.3 Transkriptoomsete andmete analüüs

Kasutati *C. consors*'i transkriptoome (Remm jt., 2014), mis sisaldasid järjestusi kaheksast erinevast koest: mürgijuha, närviganglion, süljenääre, lõhnaelund, mantel, jalg, lont ja mürgipõis. Järjestused transleeriti ning analüüsiti sama metoodikaga nagu genoomseid järjestusi.

2.4 Domeenide üle-esindatuse määramine *C. consors*'i puhul

Üle-esindatud domeenide kindlaksmääramiseks kasutati šansside suhet (OR) (*C. consors* võrreldes ülejäänud nelja molluski keskmisega) kujul:

$$OR = (A/B)/(C/D).$$

Valemis tähistab A uuritava domeeni esinemiste arvu ja B ülejäänud domeenide koguarvu *C. consors*'is. C on uuritava domeeni esinemiste arv neljas ülejäänud molluskis ning D ülejäänud domeenide koguarv neljas ülejäänud molluskis.

Domeeni piires arvulise erinevuse (*C. consors* võrreldes ülejäänud nelja molluski keskmisega) olulisuse leidmiseks kasutati Fisheri täpset testi. Domeeni üle-esindatust defineeriti järgmiselt: OR > 1,3, domeeni on *C. consors*'is leitud vähemalt 10 korda ja Fisheri testi p-väärtus on väiksem kui 0,01. Mitmese võrdlemise tõttu korrigeeriti piiriks seatud p-väärtust Bonferroni meetodiga (antud domeeni p-väärtus/domeenide koguarv, mida võrdluses kasutati, antud töös 705) ja korrigeeritult saadi uueks p-väärtuse piirmääraks $1{,}41*10^{-5}$.

2.5 7tm_1 domeeni fülogeneesipuu konstrueerimine

Fülogeneesipuu tegemisel kasutati molluskite puhul ORF-e, milles oli eelnevalt hmmsearch'i abil tuvastatud R7TM retseptorit kirjeldava Pfam-A domeeni 7tm_1 (Pfam-A ID - PF00001) esinemine. 7tm_1 domeen koondab endas seitset TM-domeeni ja nendevahelisi ekstra- ning

intratsellulaarseid linge. Antud töös loeti potentsiaalseteks R7TM-ideks kõik valgud, kust leiti 7tm_1 domeen. Kolme referentsorganismi puhul kasutati ORF-ide asemel proteoome (UniProtKB/TrEBML, *C. elegans* – 24 982, *D. melanogaster* – 17 536, *H. sapiens* – 70 101 erinevat potentsiaalset valku), et kasutada leitud 7tm_1 sisaldavaid valke fülogeneesipuu analüüsimisel.

Kasutatud metoodikaga saab tuvastada 7tm_1 domeene, millest vähemalt 50% paikneb ühes eksonis. 7tm_1 domeenide täpset koguarvu molluskites on raske hinnata, sest puuduvad eelnevad teadmised, kui suur hulk 7tm_1 domeenidest paikneb mitmes eksonis. Inimese puhul on neist suurem osa 1-2 eksonis, aga esineb ka erandeid (relaksiini retseptorites RXFP1 ja RXFP2 on 7tm_1 domeen neljas eksonis).

Saadud ORF-id ja valgufragmendid joondati HMMER3 paketi alamprogrammi hmmalign abil (parameetrid: -trim) 7tm_1 HMM-i järgi. Võimalikult korrektse fülogeneesipuu saamise huvides puhastati joondus madala katvusega piirkondadest JalView programmi (Waterhouse jt., 2009) abil (kõik joonduse tulbad, kus JalView määratud „*quality score*“ oli alla 5% parima skooriga tulba omast, eemaldati). Lisaks eemaldati kõik järjestused, milles oli tühikute arv suurem kui 20% joonduse pikkusest, sest nende asukohta ei suudetud fülogeneesipuul üheselt määrata. Eemaldati ka üksteisega rohkem kui 99% identsed järjestused (redundantsuse vähendamine). Saadud lõplik järjestus oli joondatud FASTA formaadis ja sisaldas ainult R7TM transmembraanseid domeene ning inter- ja ekstratsellulaarsete lingude piirkondi. N- ja C-terminaalsed osad eemaldati.

Fülogeneesipuu tegemiseks kasutati standardparameetritega FastTree programmi (Price jt., 2009), mis võimaldab suurt kiirust ka mahukate andmehulkade puhul. Puu saadi Newick-formaadis ja edasine analüüs ning töötlus toimus MEGA 5.10 paketi abil (Tamura jt., 2011).

2.6 7tm_1 sisaldavate transkriptide vastandamine genoomsetele ORF-idele

Kasutati vaid hmmsearch’i poolt kindlaks määratud ala ORF-ides ja transkriptides, mis andis HMM-iga vaste (väljundfailis tulbad „*ali from*“ ja „*ali to*“). Genoomist ja transkriptoomist leiti lõikudele vastavad DNA järjestused ning neid võrreldi omavahel mõlemat pidi (kord genoomsed järjestused andmebaasiks ja transkriptoomsed päringjärjestuseks, kord vastupidi), kasutades programmi BLASTN (parameetrid: -dust no -evalue 0,00001). Vaste (transkriptoomne ORF vastab genoomsele) defineeriti järgmiselt: mõlemat pidi võrdlusel olid

järjestused üksteise parimad vasted, lisatingimustega: e-väärtus < 0,00001, identsus > 95% ja vaste pikkus > 50 nukleotiidi.

3. Tulemused

3.1 Genoomsete järjestuste transleerimine ja analüüs paketiga HMMER3

Domeenide leidmiseks transleeriti *C. consors'*i assambleeritud genoomne järjestus, mis sisaldas ligikaudu 4,5 miljonit kontiigi, kuues lugemisraamis. Kokku leiti *C. consors'*i genoomist 298 653 ORF-i miinimumpikkusega 150 aa, millest suurem osa on tõenäoliselt valgu eksonid. Sama metoodikat kasutati ka teiste molluskite ning kolme referentsorganismi (*C. elegans*, *D. melanogaster* ja *H. sapiens*) genoomide puhul. Hmmsearch'i kasutades leiti *C. consors'*ist kokku 22 796 domeeni, mis jagunesid 1 383 erineva domeenitüübi vahel.

Domeenide otsimiseks kasutatud metoodika täpsust kontrolliti 7tm_1 domeeni sageduse põhjal, sest R7TM retseptorite arv inimeses on küllaltki täpselt teada (Gloriam jt., 2007). Selgus, et antud metoodikaga on võimalik R7TM arvu küllaltki täpselt määrata: *H. sapiens'*is leiti 7tm_1 domeene 305 ning lisaks veel 619 7-TM olfaktoorset retseptorit. Kirjanduse andmetel on inimeses vastavalt R7TM geene 311 (284 täispikka ja 27 pseudogeeni) ning 7TM olfaktoorsete retseptorite geene 867 (388 täispikka ja 479 pseudogeeni) (Gloriam jt., 2007). Seega töötas metoodika hästi R7TM leidmisel, kuid ei saanud kätte kõiki 7TM olfaktoorseid pseudogeene, mille tõenäoliseks põhjuseks võivad olla raaminihked. Antud metoodikaga pole võimalik tuvastada ka R7TM-e, mis jagunevad paljude eksonite vahel (tuvastamiseks peab vähemalt 50% domeeni HMM-ist asuma ühes eksonis).

3.2 Üle-esindatud domeenid *C. consors'*is

Töö põhieesmärgi saavutamiseks oli vaja leida, millised funktsioonid ja domeenid on *C. consors'*is teiste molluskitega võrreldes rohkem esindatud, et edasises analüüsis keskenduda just nendele. Üle-esindatus defineeriti OR kaudu, võrreldes omavahel *C. consors'*i ja ülejäänud nelja molluski keskmist. Kui OR oli rohkem kui 1,3, loeti domeen üle-esindatute hulka. Esikohal oli LDL-A (Pfam-A ID: PF00057) ja teiste hulgas ka kaht liiki LRR-domeenid (PF12799 ja PF13855), R7TM retseptorit kirjeldav domeen 7tm_1 (PF00001) ning 7-TM kemoretseptor Srw (PF10324). Lisaks olid üle esindatud ka mitmed kordusjärjestuste ning transponeeruvate elementidega seotud domeenid (PF00078, PF01498, PF03184, PF10551), kuid neid käesolevas töös põhjalikumalt ei analüüsitud. Kõik üle-esindatud domeenid on täpsemalt välja toodud tabelis 3.

Tabel 3. Üle-esindatud domeenid *C. consors*'is (Cco), võrrelduna teiste molluskitega: *A. californica* (Aca), *L. gigantea* (Lgi), *P. fucata* (Pfu) ja *C. gigas* (Cgi). Võrdluseks on toodud ka *C. elegans* (Cel), *D. melanogaster* (Dme) ja *H. sapiens* (Hsa).

Domeeni kirjeldus	Cco	Aca	Lgi	Pfu	Cgi	Cel	Dme	Hsa	šansside suhe Cco vs mol.	Pfam-A ID
Madala tihedusega lipoproteiini retseptori domeen, klass A	1724	76	15	17	18	67	89	48	34.3	PF00057
EB moodul	21	0	0	2	0	21	0	0	24.4	PF01683
SCAN domeen	70	0	2	14	13	0	0	52	5.6	PF02023
FLYWCH tsink-sõrme motiivi domeen	50	6	4	7	10	2	11	8	4.3	PF04500
Leutsiinirikas kordusjärjestus, tüüp 4	600	36	83	132	109	27	70	57	4.0	PF12799
7-transmembraanne GPCR kemoretseptor Srw	340	152	25	52	38	52	1	0	3.0	PF10324
Leutsiinirikas kordusjärjestus, tüüp 8	2007	226	273	682	531	40	240	355	2.9	PF13855
Rodopsiini-sarnane 7-transmembraanne retseptor	2252	483	268	787	646	14	35	305	2.5	PF00001
MULE transposaasi domeen	58	11	1	21	25	0	0	0	2.3	PF10551
Transposaas	375	207	127	20	118	36	63	0	1.9	PF01498
Pöördtranskriptaas (RNA-sõltuv DNA polümeraas)	5602	4272	676	1380	1574	134	2409	14985	1.9	PF00078
DDE superperekonna endonukleaas	130	64	8	48	56	4	8	139	1.7	PF03184

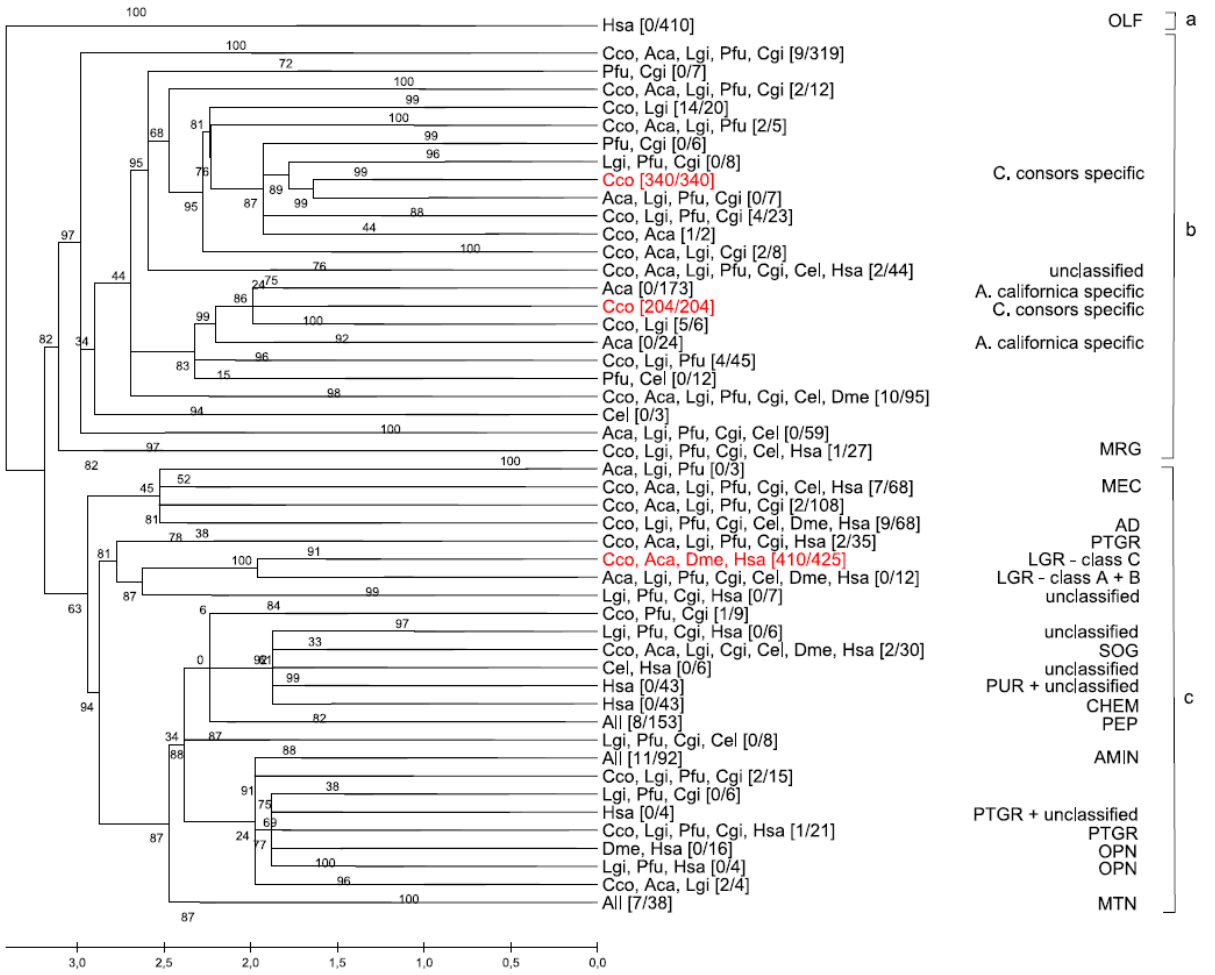
Domeenid PF10551, PF01498, PF00078 ja PF03184 on seotud kordusjärjestustega *C. consors*'i genoomis, kust leiti suurel hulgal seni kirjeldamata, ilmselt molluski-spetsiifilisi korduste motiive, milleks olid põhiliselt DNA- ja retrotransposoonid (Remm jt., 2014).

Huvitaval kombel ei olnud üle-esindatud domeenide hulgas ühtki Pfam-A andmebaasis kirjeldatud konopeptiidi domeeni (PF08088, PF02950, PF05374). Kuna konopeptiidid esinevad vaid koonustigudes, siis oleksid need teoreetiliselt pidanud olema ühed kõige enam *C. consors'*is üle-esindatud domeenid. Kasutatud metoodikaga oli võimalik leida vaid domeene, mille HMM-iga modelleeritud osast vähemalt 50% asub ühes eksonis. Konopeptiidid koosnevad üldjuhul kolmest domeenist (signaaljärjestus, pro-regioon, lõplik peptiid) (Olivera jt., 1999), mis on kodeeritud 1-3 eksoni poolt. Pfam-A konopeptiidide

HMM-id ei modelleeri kahjuks domeene eraldi, vaid sisaldavad kõiki kolme domeeni koos. Seetõttu ei suuda meie kasutatud metoodika neid sellisel kujul genoomist tuvastada, sest ühes ORF-is on liiga väike osa konopeptiidist ning see ei täida HMM-i 50% katvuse kriteeriumi.

3.3 R7TM fülogeneesipuu ja alamrühmad *C. consors'*is

Üheks arvukamaks *C. consors'*is üle-esindatud domeeniks oli 7tm_1, mis on koonusteos ilmselt ohtralt duplitseerunud, täitmaks paremini mõnd koonusteole olulist, spetsiifilist funktsiooni. Leidmaks, millised inimeses kirjeldatud R7TM alamklassid on *C. consors'*is esindatud ning potentsiaalselt uute R7TM perekondade avastamiseks konstrueeriti viie molluski ja kolme referentsorganismi 7tm_1 domeeni järjestustest fülogeneesipuu (Joonis 4), mis võimaldas inimese R7TM-ide alamgruppide teadaolevaid funktsioone üle kanda ka ortoloogsetele *C. consors'*i valkudele ning ühtlasi aitas leida molluskite-spetsiifilisi gruppe.



Joonis 4. Rodopsiini-sarnaste 7-transmembraansete retseptorite fülogeneesipuu. Puu tegemisel kasutati avatud lugemisraame *C. consors*'i, *A. californica*, *L. gigantea*, *P. fucata ja C. gigas*'e transleeritud genoomidest ning 7tm_1 domeeni sisaldavaid valke *C. elegans*'i, *D. melanogaster*'i *ja H. sapiens*'i proteoomidest. Valgudomeenid leiti HMMER 3.0 paketi alamprogrammi hmmsearch'i abil, puu tehti FastTree programmiga (standardparameetrid). Organismid on puul tähistatud järgmiselt: Cco - *C. consors;* Aca – *A. californica*; Lgi – *L. gigantea*; Pfu – *P. fucata*; Cgi – *C. gigas*; Cel – *C. elegans*; Dme – *D. melanogaster*; Hsa – *H. sapiens*. Numbrid sulgudes [x/y]: x tähistab *C. consors*'i valkude arvu, y valkude koguarvu vastavas puu harus. R7TM inimese alamrühmad on puul tähistatud järgnevalt: OLF – lõhnaretseptorid; MRG – MRG-retseptorid; MEC - melanokortiini ja kannabinoidide retseptorid; AD – adenosiini retseptorid; PTGR – prostaglandiini retseptorid; LGR – LRR-domeene sisaldavad retseptorid; SOG – somatostatiini ja opioidide retseptorid; PUR – puriini retseptorid; CHEM – kemokiinide retseptorid; PEP – erinevate peptiidide retseptorid; AMIN – biogeeniliste amiinide retseptorid; OPN – opsiinid, MTN – melatoniini retseptorid. Nomenklatuur on koostatud (Chabbert jt., 2012) põhjal. Puu jaguneb kolmeks peamiseks haruks: a – inimese OLF retseptorid, b – peamiselt molluski-spetsiifilised retseptorite grupid, c – suurem osa inimeses kirjeldatud R7TM alamklassidest. Kaks suurimat *C. consors*'i-spetsiifilist ning klass C LGR-ide haru on tähistatud punasega – need on kolm *C. consors*'is, võrreldes teiste molluskitega, kõige rohkem laienenud retseptorite gruppi.

Fülogeneesipuu jaguneb laias plaanis kolmeks: eraldiseisev haru (a), kuhu kuuluvad vaid inimese olfaktoorsed retseptorid (OLF); suur haru (b), mis sisaldab peamiselt molluskite valke, kuid lisaks ka mõningaid *C. elegans'*i ja *D. melanogaster'*i ning inimese valke (GPR139 ja MRG-retseptorid) ning haru, milles paikneb enamik senikirjeldatud R7TM alamklassidest (c). On näha, et lõhnaretseptorid moodustavad teistest täiesti eraldiseisva

grupi, mis on esindatud vaid inimeses. Molluskites ega teistes referentsorganismides neid ei ole. Haru (b), kus esinevad peamiselt molluskite valgud, on inimese R7TM alamrühmade hulgast kõige sarnasem MRG-retseptoritele, mis osalevad valu tunnetamises (Grazzini jt., 2004). Seal asuvad ka kaks väga suurt *C. consors'*i-spetsiifilist gruppi. Kuna MRG-retseptorid on ülejäänud harust lahknenud kõige varem, võib eeldada, et aja jooksul on valgud funktsionaalselt divergeerunud ning suurem osa molluskite valkudest harus (b) ei pruugi olla seotud valu tunnetamise, vaid mõne muu funktsiooniga, mida pole inimese R7TM alamklasside hulgas kirjeldatud. Harus (b) asub ka kaks *A. californica'*le spetsiifilist haru (173 ja 24 järjestust) millest suurem koondab endas tõenäoliselt mitmesuguseid kemoretseptoreid, sest 54 neist on rohkem kui 75% identsed *A. californica'*s kirjeldatud kemoretseptoritega (Cummins et al., 2009). Inimese valgu GPR139 täpne funktsioon ei ole teada, varasemalt on leitud, et seda ekspresseeritakse eranditult ajupiirkonnas (Matsuo jt., 2005). Samal harul asuvad ka *C. elegans'*i ja *D. melanogaster'*i mõningad neuropeptiidsete RF-amiidide retseptorid.

Harus (c) on esindatud kõik organismid ja see sisaldab enamikke inimese R7TM alamklassidest. On näha, et *C. consors'*is puuduvad lisaks lõhnaretseptoritele veel ka opsiinid (OPN) ning kemokiinide (CHEM) ja puriinide (PUR) retseptorid. Väga rohkelt on *C. consors'*is esindatud aga LRR-domeeni sisaldavate R7TM retseptorite (LGR) sarnased valgud. Huvitaval kombel on kõik *C. consors'*i LGR-sarnased valgud ühes grupis inimese LDL-A domeeni sisaldavate klass C LGR-idega (LDL-A-LGR), mille hulka kuuluvad relaksiini retseptorid. Inimese klass A ja B LGR-ide grupis *C. consors'*i valke ei ole. Klass C LGR-ide sarnaste valkude hulk on *C. consors'*is, võrreldes teiste organismidega, tohutult laienenud – 425-st klass C LGR-st on 410 pärit *C. consors'*ist.

3.4 7tm_1 domeen võib *C. consors*'is esineda koos LRR ja LDL-A domeenidega

Varasemates töödes on leitud, et R7TM-ides esinevad lisaks 7tm_1 domeenile veel ka LRR, LDL-A ja potentsiaalselt ka CUB-domeenid (Tensen jt., 1994, Kamesh jt., 2008). Teiste domeenide võimalikku koosesinemist 7tm_1-ga vaadeldi ka *C. consors'*is, et kirjeldada koonusteos esinevad kombinatsioonid. Kasutatud metoodika võimaldas kombinatsioone genoomis tuvastada vaid ühe ORF-i piires - juhul, kui domeenid asusid kõik koos ühes või mitmes eksonis, ilma et nende vahel oleks olnud stoppkoodoneid. Kokku leiti genoomist neli erinevat 7tm_1 sisaldavat kombinatsiooni, transkriptoomist kolm (tabel 4, domeenide arv kombinatsioonides võib varieeruda, välja arvatud 7tm_1 puhul, mida on alati üks). Leitud kombinatsioonidest 7tm_1 + LRR sarnaneb inimese klass A ja B LGR-idele, 7tm_1 + LRR +

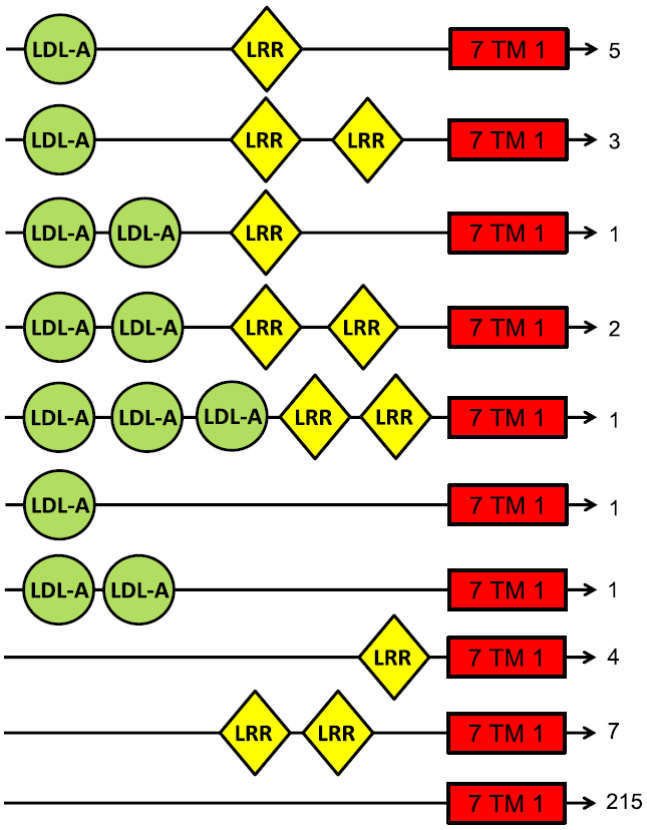
LDL-A inimese klass C LGR-idele. 7tm_1 + LDL-A on varem leittud meripõies *C. intestinalis* (Kamesh jt., 2008) ning 7tm_1 + LRR + LDL-A + CUB merisiilikus *S. purpuratus* (UniProtKB/TrEMBL), kummagi puhul ei ole funktsioon teada.

Tabel 4. *C. consors'*i genoomist ja kudede transkriptoomidest leitud 7tm_1 domeeni kombinatsioonid teiste domeenidega.

	7tm1 + LRR	7tm1 + LDL-A	7tm1 + LRR + LDL-A	7tm1 + LRR + LDL-A + CUB	Kokku
Genoom	191	13	159	3	366
Mürgijuha	1	0	0	0	1
Närviganglion	3	0	0	0	3
Süljenääre	1	0	0	0	1
Lõhnaelund	12	4	12	0	28
Mantel	1	0	1	0	2
Peajätke	2	0	0	0	2
Mürgipõis	1	0	0	0	1

7tm_1 kombinatsioone sisaldavaid transkripte genoomsetele ORF-idele vastandades oli võimalik määrata ka erinevate kombinatsioonide jaotus 7tm_1 fülogeneesipuul. Kõik kombinatsioonid, mis transkriptoomis esinesid, asusid fülogeneesipuul inimese klass C LGR-ide harus. Kuna antud metoodikaga sai tuvastada vaid kombinatsioone, milles domeenide vahel ei olnud intronitest tingitud stoppkoodoneid, siis on tõenäoline, et suur osa *C. consors'*i 7tm_1 domeenidest klass C LGR-ide harus, mida ei leitud koos teiste domeenidega või mis leiti kombinatsioonis ainult LDL-A või LRR domeenidega, on tegelikult siiski kombinatsioonis nii LRR kui ka LDL-A domeenidega.

LDL-A-LGR retseptoritel on oma kindel struktuur, mis on üldiselt määratud domeenide omavahelise järjestusega valgus. Käesolevas töös analüüsiti, kui hästi langeb kokku *C. consors'*is leiduvate potentsiaalsete LDL-A-LGR-ide domeenijärjestus kirjanduse andmetega, et kinnitada võimalike leidude usaldusväärsust. Domeenide omavahelise paigutuse kindlaksmääramiseks kasutati hmmsearch'i abil erinevate kudede transkriptides ennustatud domeenide asukohti. Joonisel 5 on ülevaade erinevatest 7tm_1 domeeni kombinatsioonidest, mis leiti *C. consors'*i lõhnaelundi transkriptoomist.



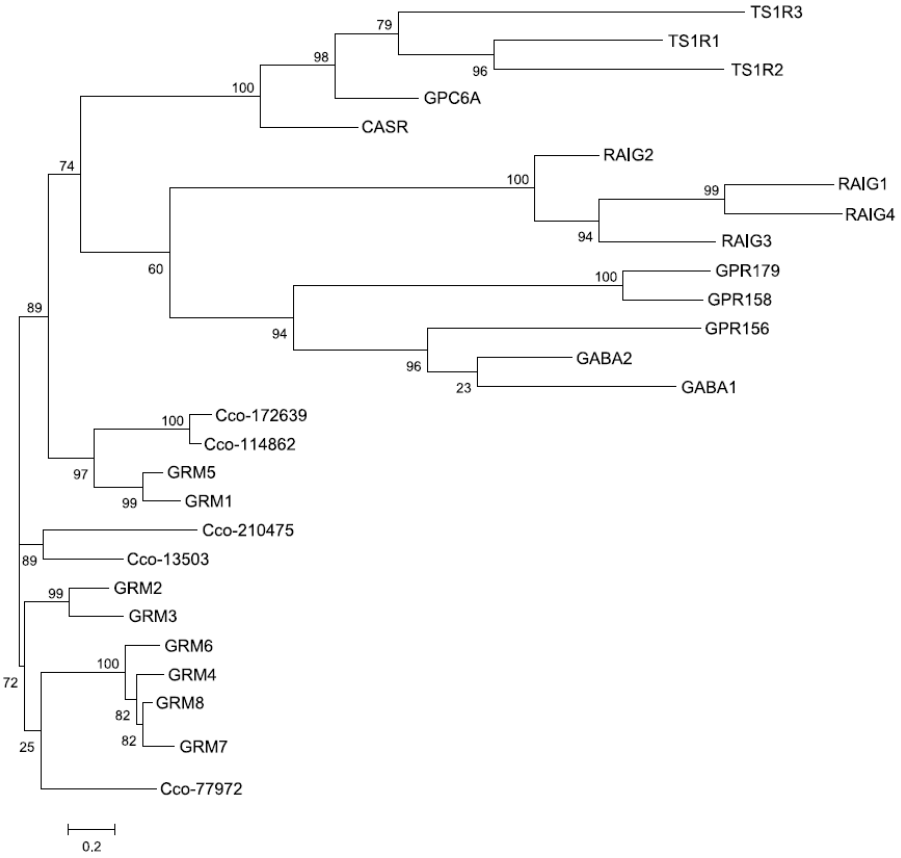
Joonis 5. 7tm_1 erinevad kombinatsioonid teiste domeenidega *C. consors*'i lõhnaelundi transkriptoomi põhjal. Esinemiste arv on välja toodud kombinatsioonist paremal. LDL-A-LGR retseptorid sisaldavad tavapäraselt üht 7tm_1 domeeni ning üht või mitut LRR- ja LDL-A domeeni. Inimeses kuuluvad LDL-A-LGR-ide hulka relaksiini retseptorid. *C. consors*'i puhul on nende ehituses 7tm_1 domeene alati üks, LRR domeene 1-2 ja LDL-A domeene 1-3 ning nende järjestus vastab varem kirjeldatud LDL-A-LGR-ide üldstruktuurile (Tensen jt., 1994, Kamesh jt., 2008, van der Westhuizen jt., 2008) – 7tm_1 domeen ankurdub membraani ning rakuvälised LRR-sillad ühendavad seda eespool paiknevate LDL-A domeenidega, mis osalevad ligandi sidumisel.

3.5 Maitseretseptorite analüüs *C. consors*'is

Viie erineva põhimaitse tunnetamises osalevad erinevad retseptorite tüübid. Siiani on täpsemalt kirjeldatud imetajate Taste1 retseptorid (T1R), mis osalevad umami ja magusa maitse (Nelson jt., 2001) tunnetamises ning mõru maitse retseptorid Taste2 (T2R) (Adler jt., 2000). Mõlemate puhul on tegu GPCR-idega. Mõru ja soolase maitse retseptoreid ei ole siiani suudetud täpselt kindlaks määrata, tõenäoliselt ei kuulu nad GPCR-ide hulka. Käesoleva töö käigus uuriti, kas *C. consors* võiks tunda umamit, magusat ning mõru maitset imetajatega sarnaselt.

Nii T1R kui ka T2R on kirjeldatud Pfam-A andmebaasis, T1R vastavalt glutamaadi perekonna mudeli 7tm_3 ja T2R neile spetsiifilise mudeli TAS2R abil. HMMER3 paketi abil leiti *C. consors*'ist kokku 20 7tm_3 domeeni, kuid mitte ühtki TAS2R domeeni sisaldavat ORF-i. 7tm_3 domeenide täpsemaks analüüsiks ehitati fülogeneesipuu (joonis 6), mis sisaldas

*H. sapiens'*i proteoomist pärit valke ning *C. consors'*i ORF-e kus tuvastati 7tm_3 domeen. 20-st *C. consors'*i ORF-ist sai puus kasutada vaid viit, sest ülejäänutel puudus eesmine või tagumine osa domeenist ja järjestus ei sisaldanud piisavalt infot kvaliteetse fülogeneesipuu tegemiseks. Tõenäoliselt oli puudumise põhjuseks see, et mõningad koonusteo 7tm_3 domeenidest asusid erinevates eksonites. Antud metoodikaga sai leida täies pikkuses vaid sellised 7tm_3 domeenid, mis olid kodeeritud ühes eksonis. Fülogeneesipuul ei paikne ükski *C. consors'*ist pärit valk T1R harus, vaid on lähedasemad metabotroopilistele glutamaadi retseptoritele. Selle alusel võib öelda, et *C. consors* ei tunne mõru ega tõenäoliselt ka umami ega magusat maitset imetajatega sarnaste mehhanismide abil.

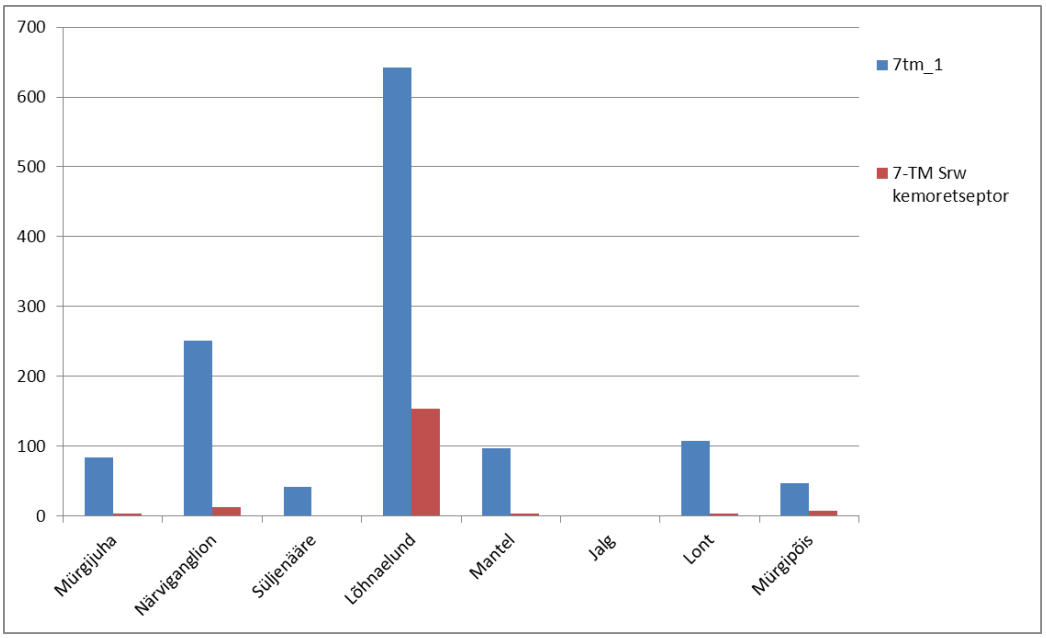


Joonis 6. *H. sapiens'*i ja *C. consors'*i glutamaadi perekonna GPCR-ide fülogeneesipuu. Joonis kujutab *H. sapiens'*i valkudest ja *C. consors'*i ORF-idest, mis sisaldasid Pfam-A domeeni 7tm_3, ehitatud fülogeneesipuud. 7tm_3 domeen kirjeldab kõiki glutamaadi perekonna valke. Puu on loodud FastTree programmiga. Inimese valgud on tähistatud lühenditega, *C. consors'*i valgud eesliitega Cco ja sellele järgneva ORF-i numbriga. Fülogeneesipuu jaguneb kolmeks suuremaks haruks: metabotroopilised glutamaadi retseptorid (GRM), maitseretseptoreid (TS1R) sisaldav haru ning $GABA_B$ ja RAIG-retseptorite haru. Kõik *C. consors'i* täispikad 7tm_3 järjestused on GRM-ide harus.

3.6 *C. consors*'i lõhnaelundis leidub kõige enam unikaalseid R7TM ja Srw kemoretseptorite transkripte

Kaht tüüpi retseptorid – R7TM ja 7-TM Srw kemoretseptorid – olid *C. consors*'i genoomis teiste molluskitega võrreldes oluliselt rohkem esindatud. 7-TM Srw kemoretseptoreid on varasemalt kirjeldatud *C. elegans*'is, mis, sarnaselt koonustigudele, ei kasuta nägemis- ega ka kuulmismeelt, vaid tunnetab keskkonda sealt pärinevaid keemilisi signaale vastu võttes. Srw retseptorite puhul on näidatud sarnasust järjestuste tasemel GPCR-idega ja ühtset päritolu erinevate neuropeptiidide retseptoritega ning arvatakse, et Srw ligandideks on mitmesugused peptiidid (Thomas ja Robertson, 2008).

Koetranskriptoomide abil oli võimalik määrata R7TM ja 7-TM Srw retseptorite jaotus – millises koes transkribeeritakse rohkem, millises vähem erinevaid retseptoreid, lähtudes neid sisaldavate erinevate transkriptide arvust, mis oli määratud programmi hmmsearch abil. Mõnevõrra oodatult olid mõlemad retseptoritüübid selgelt kõige rohkem esindatud koonusteo lõhnaelundis ja vähemal määral närviganglionides (joonis 7).



Joonis 7. Unikaalsete 7tm_1 ja 7-TM Srw kemoretseptori domeenide jaotus *C. consors*'i kudedes. Vertikaalteljel on välja toodud vastavat tüüpi domeenide esinemissagedus. *C. consors*'i transkriptoom sisaldas kaheksast erinevast koest pärit transkripte, mille hulgast otsiti programmi hmmsearch abil sellised, mis sisaldasid 7tm_1 või 7-TM Srw kemoretseptori domeene. Kudedes on erinevate transkriptide koguarv küllaltki sarnane, seega on domeenide esindatus kudede vahel võrreldav. Selgelt on näha, et 7tm_1 (sinine) ja eriti 7-TM Srw kemoretseptori (punane) domeene sisaldavaid transkripte leidub eriti rohkelt koonusteo lõhnaelundis ja vähemal määral närviganglionides. Teistes kudedes on neid vähem, jalas peaaegu üldse mitte (vaid üks 7tm_1 domeeni sisaldav transkript).

4. Arutelu

C. consors'ist leiti 12 tüüpi domeene, mis olid teiste molluskitega võrreldes oluliselt üle-esindatud. Laias plaanis jagunesid need kahte gruppi – kordusjärjestustega seotud domeenid ning mitmesugused 7TM retseptoritega seotud domeenid. Viimaste hulgas olid LDL-A, kaht tüüpi LRR-id ning 7tm_1 ja 7-TM Srw. Kaks nendest domeenidest – 7tm_1 ja 7-TM Srw – on retseptorid, mis annab alust arvata, et võrreldes teiste antud töös analüüsitud molluskitega on mõni nende retseptorite alamklass *C. consors*'is oluliselt laienenud, et täita mõnd koonusteole väga olulist funktsiooni. LRR ja LDL-A domeenide rohkust on keeruline täpselt analüüsida, sest need on küllaltki laialdaselt kasutatud bioloogilised „ehituskivid“ ja esinevad väga erinevates valkudes.

Teadaolevalt ei ole varem molluskite genoomides valgudomeenide repertuaari kirjeldatud, mistõttu pole teada, milliseid domeene seal peaks kindlasti leiduma ja milliseid mitte. Huvitava leiuna meie analüüsis võib välja tuua SCAN domeeni (PF02023), mida on siiani kirjeldatud vaid selgroogsetes (Sander jt., 2003; Emerson ja Thomas, 2011). SCAN on umbes 80 aa pikkune, ohtralt leutsiini sisaldav konserveerunud valgumotiiv, mis paikneb tsink-sõrme motiivi sisaldavate valkude N-terminuses ja osaleb valk-valk interaktsioonides. Käesolevas töös leiti SCAN domeene ka kõigi molluskite genoomides, välja arvatud *A. californica*. Fülogeneetilisel analüüsil moodustasid molluskite SCAN-domeenid omaette grupi, eraldudes selgesti inimese omadest, mis näitab, et tõenäoliselt ei ole tegu kontaminatsiooniga, vaid tõepoolest molluskites eraldi evolutsioneerunud domeeniga.

7tm_1 domeenide fülogeneetilisel analüüsil selgus, et võttes aluseks R7TM alamklassid inimeses, on koonusteos teiste R7TM alamklassidega võrreldes väga ohtralt esindatud LDL-A-LGR-id, mis sarnanevad inimese relaksiini retseptoritele (klass C LGR, 425-st valgust 410 *C. consors*'ist), ning MRG-retseptoritega kaugelt seotud retseptorid (kaks suurt *C. consors*'i-spetsiifilist gruppi, vastavalt 340 ja 204 valku). Mõningad inimeses esinevaid R7TM alamklasse koonusteost ei leitud: polnud opsiine (OPN), lõhnaretseptoreid (OLF), kemokiinide retseptoreid (CHEM) ega puriinide retseptoreid (PUR). Suure hulga klass C LGR-sarnaste retseptorite kõrval puudusid *C. consors*'is klass A ja B LGR-sarnased retseptorid, mis olid teistes molluskites olemas. LDL-A-LGR-ide väga suur hulk koonusteos võib anda vastuse ka LRR ja LDL-A domeenide üle-esindatuse kohta: LDL-A-LGR-i moodustab domeenide tasemel üks 7tm_1, mitu LRR-i ja üks (inimese puhul, teistes organismides võib arv erineda) LDL-A domeen. Potentsiaalsete koonusteo LDL-A-LGR-ide struktuuri määramiseks uuriti 7tm_1 võimalikku koosesinemist koos teiste domeenidega nii

*C. consors'*i transkriptoomis kui ka genoomis ning leiti, et 7tm_1 võib esineda koos LRR ja/või LDL-A domeenidega, mis kinnitab LDL-A-LGR-ide esindatust koonusteos. Ühtlasi võib LRR ja LDL-A domeenide üle-esindatuse põhjuseks olla just neid sisaldavate LDL-A-LGR perekonna retseptorite laienemine koonusteos. Koonusteo transkriptoomis on esindatud samad kombinatsioonid mis genoomiski, välja arvatud CUB-domeeni sisaldav variant. On võimalus, et seda antud kudedes ei transkribeerita või transkribeeritakse väga vähesel määral. Tegu võib olla ka pseudogeeniga, mida reaalselt ei transkribeeritagi. Kombinatsioonide jaotuses kudede vahel tuleb väga selgelt esile, et valdav enamik kombinatsioone sisaldavatest transkriptidest pärineb koonusteo lõhnaelundist, teistes kudedes on neid väga vähe.

Lisaks fülogeneetilisele analüüsile määrati ka unikaalsete 7tm_1 ja 7-TM Srw transkriptide jaotus koonusteo kudedes, kasutades transkriptoomist saadud andmeid. Analüüs näitas, et 7tm_1 esineb kõigis transkriptoomis esindatud kaheksas koes, jalas aga väga vähesel määral (vaid üks transkript). Selgelt kõige rohkem, nii suhteliselt kui ka absoluutarvult, oli neid koonusteo lõhnaelundis, kust leiti kõige rohkem ka Srw-tüüpi 7-TM kemoretseptoreid. Kuna koonusteo nägemismeel on vähearenenud, võib järeldada, et keskkonna tunnetamisel ja jahipidamisel on lõhnaelund ülimalt oluline, sisaldades suurt hulka erinevaid retseptoreid. Trankriptoomi abil oli võimalik määrata ka fülogeneesipuul analüüsitud 7tm_1 domeenide koelist päritolu – suurem osa *C. consors'*is laienenud 7tm_1 gruppidest on pärit just koonusteo lõhnaelundist.

Kokkuvõtlikult saab öelda, et *C. consors'*is on oluliselt laienenud kolm retseptorite gruppi, millest suurim (410 järjestust) on väga sarnane inimese klass C LGR-idega ning kaks ülejäänut (204 ning 340 järjestust) on kaudselt seotud inimese MRG ning *C. elegans'*i ja *D. melanogaster'*i RF-amiidsete neuropeptiidide retseptoritega ning lähemalt *A. californica* kemoretseptoritega (Cummins jt., 2009). Enamikku laienenud gruppidest pärit järjestustest transkribeeritakse peamiselt koonusteo lõhnaelundis, eriti just LDL-A-LGR-sid. Nii MRG kui ka inimese klass C LGR-ide teadaolevateks ligandideks on peptiidid. MRG-l on nendeks β-alaniin ja RF-neuropeptiidid (Dong jt., 2001, Han jt., 2002) ning klass C LGR-idel relaksiin (van der Westhuizen jt., 2008). *A. californica* puhul on näidatud RF-neuropeptiidide alla kuuluva tetrapeptiidi – FMRF-amiidi – laialdast levikut üle terve närvisüsteemi (Lopez-Vera jt., 2008). Peptiidsed ligandid on ka 7-TM Srw kemoretseptoritel (Thomas ja Robertson, 2008).

Siit võib järeldada, et *C. consors'*i-spetsiifiliste R7TM retseptorigruppide bioloogiliseks funktsiooniks võiks olla eelkõige peptiidsete signaalide vastuvõtmine keskkonnast. On teada,

et veekeskkonnas leidub tavaliselt ohtralt nii vabu aminohappeid kui ka peptiide, mida erinevad vees elavad organismid tunnetavad. Konnakullese olfaktoorsete retseptoritega tehtud katsed (Hassenklöver jt., 2012) on näidanud eelkõige vabade aminohapete olulisust signaalmolekulidena, kuid teiste organismide ning retseptoritüüpide puhul võivad tähtsad olla ka peptiidsed signaalid. Seega võib oletada, et *C. consors*'is laienenud retseptorite gruppe kasutab tigu näiteks peptiidsete signaalide abil saagi detekteerimiseks või liigikaaslaste tuvastamiseks.

Kokkuvõte

Käesolevas magistritöös uuriti, mis eristab koonustigu *C. consors* teistest varem sekveneeritud molluskitest (*A. californica, P. fucata, C. gigas ja L. gigantea*) valgudomeenide tasemel. Uuringus keskenduti just *C. consors'*is, võrreldes nelja ülejäänud molluskiga, üle-esindatud domeenidele. Tulemuste tõlgendamiseks kasutati kolme levinud referentsorganismi – *H. sapiens, D. melanogaster* ja *C. elegans*. Kasutati Pfam-A domeeniandmebaasi ning HMMER 3.0 tarkvara domeenide leidmiseks transleeritud genoomidest.

Oluliselt üle-esindatud oli kaks domeenide gruppi – kordusjärjestustega seotud domeenid ning retseptorite domeenid. Edasises analüüsis keskenduti retseptorite domeenidele, eelkõige just rodopsiini-sarnasete 7-transmembraansete retseptorite grupile, mida kirjeldab Pfam-A domeen 7tm_1. Fülogeneetilisel analüüsil eristus kolm *C. consors'*ile omast 7tm_1 gruppi, milledest suurim (410 järjestust) on väga sarnane inimese relaksiini retseptoritele ning teised (204 ja 340 järjestust) sarnanevad *A. californica* kemoretseptoritega ning on kaudselt seotud ka *C. elegans'*i ja *D. melanogaster'*i RF-neuropeptiidide retseptorite ja inimese MRG-retseptoritega. Fülogeneetiline sarnasus viitab sellele, et kolme *C. consors'*i-spetsiifilise grupi ligandideks võiksid olla peptiidid. Koespetsiifiliste transkriptoomide abil selgitati välja, et kõnealuseid retseptorigruppe transkribeeritakse eelkõige koonusteo lõhnaelundis, seega on tõenäoline, et koonusteo jaoks on väga oluline keskkonnast pärit peptiidsete signaalide tunnetamine.

Kokkuvõttes võib öelda, et domeenilist ülesehitust analüüsides on võimalik leida meid huvitavaid organismi-spetsiifilisi valkude gruppe ning fülogeneetilise analüüsi abil kindlaks määrata nende võimalikud funktsioonid. Käesolevate tulemuste põhjal oleks võimalik edasine laboratoorne analüüs, et oletusi kinnitada, sest seniajani ei ole kirjeldatud peptiidsete signaalide suurt olulisust vesikeskkonnas ja ühtlasi pole täpselt teada, kuidas ja mille põhjal täpselt koonustigu oma saagi leiab.

Summary

Märt Roosaare

Protein domain analysis in the cone snail *Conus consors*

C. consors is a venomous, predatory sea snail that uses envenomed harpoons to paralyse and kill prey. Each species produces its own cocktail of neurotoxic conopeptides, so the total number of different peptides is estimated to be over 500 000, which makes cone snails very attractive as potential leads for new pharmaceuticals. However, no cone snail species has yet been sequenced with high coverage, so we are still lacking genomic information about cone snails and even mollusks in general, as just four species have been sequenced so far (*A. californica*, *L. gigantea*, *P. fucata* and *C. gigas*). As *Mollusca* is a huge and diverse phylum, this information is vital in order to gain a deeper understanding about them in general.

In order to shed more light on what makes the cone snail different from other marine mollusks, this master's thesis focuses on the protein domain analysis, looking for the ones that are over-represented in *C. consors* compared to the average of other four sequenced mollusks. We also used three reference organisms – *H. sapiens*, *D. melanogaster* and *C. elegans* – in order to interpret the results. As for the domains, we used Pfam-A database and HMMER 3.0 software to detect them from the translated genomes.

Twelve domain types were significantly over-represented in *C. consors* and they could be divided into two groups – domains related to repeats and the ones related to receptors. Present work focused on the receptors, concentrating on the 7tm_1 domain which represents rhodopsin-like GPCR-s. Phylogenetic analysis of the 7tm_1 revealed three *C. consors*-specific groups: biggest one (410 sequences) was very similar to human relaxin receptors, two other ones (204 and 340 sequences) are similar to *A. californica* chemoreceptors and the RF-amide neuropeptide receptors of *C. elegans* and *D. melanogaster*. These phylogenetic relationships indicate that *C. consors*-specific receptor groups may have peptides as ligands. Using tissue transcriptomes, we found that the three groups are mainly transcribed in osphradium of the cone snail. All in all, it seems that *C. consors* is very well adapted to detect peptide signals from seawater.

To sum up, we can say that domain analysis can be used to find organism-specific groups. On top of that, phylogenetic analyses help to understand which functions these groups may serve. Current results could be tested in a wet lab in order to challenge the speculations as the

importance of peptides as waterborne signal molecules has not been described very well. Moreover, it is not fully known how the cone snails detect their prey or conspecifics.

Viited:

Adler, E., Hoon, M. A., Mueller, K. L., Chandrashekar, J., Ryba, N. J., Zuker, C. S. (2000) A novel family of mammalian taste receptors. Cell. 100(6):693-702.

Altschul, S. F., Gish, W. (1996) Local alignment statistics. Methods Enzymol. 266:460-80.

Bockaert, J., Pin, J. P. (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. EMBO J. 18, 1723–1729.

Chabbert, M., Castel, H., Pele, J., Deville, J., Legendre, R., Rodien, P. (2012) Evolution of class A G-protein-coupled receptors: implications for molecular modeling. Curr. Med. Chem. 19(8):1110-8.

Cummins, S. F., Erpenbeck, D., Zou, Z., Claudianos, C., Moroz, L. L., Nagle, G. T., Degnan, B. M. (2009) Candidate chemoreceptor subfamilies differentially expressed in the chemosensory organs of the mollusc Aplysia. BMC Biol. 7:28.

Davies, M. N., Secker, A., Freitas, A. A., Mendao, M., Timmis, J., Flower, D. R. (2007) On the hierarchical classification of G protein-coupled receptors. Bioinformatics, 23, 3113–3118.

Dong, X., Han, S., Zylka, M. J., Simon, M. I., Anderson, D. J. (2001) A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. Cell. 106(5):619-32.

Eddy, S. R. (2011). Accelerated profile HMM searches. PLoS Comp. Biol., 7:e1002195.

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32 (5): 1792–1797.

Emerson, R. O., Thomas, J. H. (2011) Gypsy and the Birth of the SCAN Domain. J. Virol. 85(22): 12043–12052.

Fredriksson, R., Lagerström, M. C., Lundin, L. G., Schiöth, H. B. (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. Mol. Pharmacol. 63, 1256–1272.

Fredriksson, R., Schiöth, H. B. (2005a) The Repertoire of G-Protein–Coupled Receptors in Fully Sequenced Genomes. Mol. Pharmacol. 67:1414–1425.

Gloriam, D. E., Fredriksson, R., Schiöth, H. B. (2007) The G protein-coupled receptor subset of the rat genome. BMC Genomics; 8: 338.

Granier, S., Kobilka, B. (2012) A new era of GPCR structural and chemical biology. Nature Chemical Biology. 8, 670–673.

Grazzini, E., Puma, C., Roy, M. O., Yu, X. H., O'Donnell, D., Schmidt, R., Dautrey, S., Ducharme, J., Perkins, M., Panetta, R., Laird, J. M., Ahmad, S., Lembo, P. M. (2004) Sensory neuron-specific receptor activation elicits central and peripheral nociceptive effects in rats. Proc. Natl. Acad. Sci. USA. 101(18):7175-80.

Han, S.K., Dong, X., Hwang, J.I, Zylka, M.J., Anderson, D.J., Simon, M.I. (2002) Orphan G protein-coupled receptors MrgA1 and MrgC11 are distinctively activated by RF-amide-related peptides through the Galpha q/11 pathway. Proc Natl Acad Sci U S A. 99(23):14740-5.

Hassenklöver, T., Pallesen, L.P., Schild, D., Manzini, I. (2012) Amino acid- vs. peptide-odorants: responses of individual olfactory receptor neurons in an aquatic species. PLoS One. 7(12):e53097

Herbert, T.E., Bouvier, M. (1998) Structural and functional aspects of G protein-coupled receptor oligomerization. Biochem. Cell Biol., 76, 1–11.

Henikoff, S. (1996) Scores for sequence searches and alignments. Curr. Opin. Struct. Biol., 6, 353-360.

Higgins, D. G., Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 73(1):237-44.

Hu, H., Bandyopadhyay, P. K., Olivera, B. M., Yandell, M. (2011) Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. BMC Genomics. 12: 60.

Kamesh, N., Aradhyam, G. K., Manoj, N. (2008) The repertoire of G protein-coupled receptors in the sea squirt Ciona intestinalis. BMC Evolutionary Biology, 8:129.

Klabunde, T., Hessler, G. (2002) Drug design strategies for targeting G-protein-coupled receptors. Chembiochem, 3, 928–944.

Lagerström, M. C., Schiöth, H. B. (2008) Structural diversity of G protein-coupled receptors and significance for drug discovery. Nat. Rev. Drug Discov.7(4):339-57.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C. jt. (2001) Initial sequencing and analysis of the human genome. Nature.409(6822):860-921.

López-Vera, E., Aguilar, M.B., Heimer de la Cotera, E.P. (2008) FMRFamide and related peptides in the phylum mollusca. Peptides. 29(2):310-7.

Matsuo, A., Matsumoto, S., Nagano, M., Masumoto, K. H., Takasaki, J., Matsumoto, M., Kobori, M., Katoh, M., Shigeyoshi, Y. (2005). Molecular cloning and characterization of a novel Gq-coupled orphan receptor GPRg1 exclusively expressed in the central nervous system. Biochem. Biophys. Res. Commun. 331(1):363-9.

Nelson, G., Hoon, M. A., Chandrashekar, J., Zhang, Y., Ryba, N. J., Zuker, C. S. (2001). Mammalian sweet taste receptors. Cell. 106(3):381-90.

Nordström, K.J., Sällman Almén, M., Edstam, M. M., Fredriksson, R., Schiöth, H.B. (2011) Independent HHsearch, Needleman--Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. Mol Biol Evol. 28(9):2471-80.

Nygaard, R., Frimurer, T. M., Holst, B., Rosenkilde, M. M., Schwartz, T. W. (2009) Ligand binding and micro-switches in 7TM receptor structures. Trends Pharmacol. Sci. (5):249-59.

Olivera, B. M., Walker, C., Cartier, G. E., Hooper, D., Santos, A. D., Schoenfeld, R., Shetty, R., Watkins, M., Bandyopadhyay, P., Hillyard, D. R. (1999) Speciation of cone snails and interspecific hyperdivergence of their venom peptides. Potential evolutionary significance of introns. Ann. N. Y. Acad. Sci.870:223-37.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., Miyano, M. (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. Science 289: 739–745.

Ponting, C. P., Russell, R. R. (2002) The natural history of protein domains. Annu. Rev. Biophys. Biomol. Struct. 31:45-71.

Price, M., Dehal, P., Arkin, A. (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26: 1641-1650.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., Finn, R. D. (2012) The Pfam protein families database. Nucleic Acids Research. Database Issue 40:D290-D301

Remm, M., Stockwell, T., Andreson, R., Brauer, A. jt. (2014) The genome of the fish-hunting cone snail *Conus consors*. Manuskript edastatud publitseerimiseks.

Rasmussen, S. G., DeVree, B. T., Zou, Y., Kruse, A. C., Chung, K. Y., Kobilka, T. S., Thian, F. S., Chae, P. S., Pardon, E., Calinski, D., Mathiesen, J. M., Shah, S. T., Lyons, J. A., Caffrey, M., Gellman, S. H., Steyaert, J., Skiniotis, G., Weis, W. I., Sunahara, R. K., Kobilka, B. K. (2011) Crystal structure of the β2 adrenergic receptor-Gs protein complex. Nature 477(7366):549-55.

Sander, T. L., Stringer, K. F., Maki, J. L., Szauter, P., Stone, J. R., Collins, T. (2003) The SCAN domain defines a large family of zinc finger transcription factors. Gene. 310:29-38.

Schiöth, H.B., Fredriksson ,R. (2005b) The GRAFS classification system of G-protein coupled receptors in comparative perspective. Gen Comp Endocrinol. 142(1-2):94-101.

Simakov, O., Marletaz, F., Cho, S. J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D. H. et al. (2013) Insights into bilaterian evolution from three spiralian genomes. Nature. 493(7433):526-31.

Sonnhammer, E. L. L., Eddy, S. R., Durbin, R. (1997) Pfam: a comprehensive database of protein families based on seed alignments: Proteins 28:405-420.

Spengler, H. A., Kohn, A. J. (1995) Comparative external morphology of the *Conus* osphradium (Mollusca: Gastropoda). Journal of Zoology, 235: 439–453.

Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E. et al. (2012) Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA Res. 19(2):117-30.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S.(2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28(10):2731-9.

Tensen, C. P., van Kesteren, E. R., Planta, R. J., Cox, K. J., Burke, J. F., van Heerikhuizen, H., Vreugdenhil, E. (1994) A G protein-coupled receptor with low density lipoprotein-binding motifs suggests a role for lipoproteins in G-linked signal transduction. Proc. Natl. Acad. Sci. USA. 91(11):4816-20.

Thomas, J.H., Robertson, H. M. (2008) The Caenorhabditis chemoreceptor gene families. BMC Biol. 6:42.

Waterhouse, A.M., Procter, J. B., Martin, D. M. A., Clamp, M., Barton, G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

van der Westhuizen, E. T., Halls, M. L., Samuel, C. S., Bathgate, R. A., Unemori, E. N., Sutton, S. W., Summers, R. J. (2008) Relaxin family peptide receptors - from orphans to therapeutic targets. Drug Discov. Today. 13(15-16):640-51.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H. et al. (2012) The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 490(7418):49-54.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina _____

(*autori nimi*)

(sünnikuupäev: _____)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
_____
_____
_____,

(*lõputöö pealkiri*)

mille juhendaja on _____,

(*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, _____ (*kuupäev*)