Realia et naturalia

DISSERTATIONES
BIOLOGICAE
UNIVERSITATIS
TARTUENSIS
279

SILJA LAHT

Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
279

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
279

SILJA LAHT

Classification and identification of conopeptides using profile hidden Markov models and position-specific scoring matrices

UNIVERSITY OF TARTU
PRESS
1632

Institute of Molecular and Cell Biology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in gene technology) on 18.06.2015 by the Council of the Institute of Molecular and Cell Biology, University of Tartu.

Supervisor: Prof. Maido Remm, PhD
Department of Bioinformatics,
Institute of Molecular and Cell Biology,
University of Tartu, Tartu, Estonia

Opponent: Dr. Helena Safavi-Hemami, PhD
Department of Biology,
University of Utah, Salt Lake City, USA

Commencement: Room No 105, 23B Riia Str, Tartu, on August 27th 2015, at 10.15

Publication of this thesis is granted by the Institute of Molecular and Cell Biology, University of Tartu and by the Graduate School in Biomedicine and Biotechnology created under the auspices of European Social Fund.

ISSN 1024–6479
ISBN 978-9949-32-887-1 (print)
ISBN 978-9949-32-888-8 (pdf)

University of Tartu Press
www.tyk.ee

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

I. Laht S, Koua D, Kaplinski L, Lisacek F, Stöcklin R, Remm M. (2012) Identification and classification of conopeptides using profile Hidden Markov Models. Biochim Biophys Acta. 1824:488–92.

II. Koua D, Brauer A, Laht S, Kaplinski L, Favreau P, Remm M, Lisacek F, Stöcklin R. (2012) ConoDictor: a tool for prediction of conopeptide superfamilies. Nucleic Acids Res. 40(Web Server issue):W238–41.

III. Koua D, Laht S, Kaplinski L, Stöcklin R, Remm M, Favreau P, Lisacek F. (2013) Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies. Biochim Biophys Acta. 1834(4):717–24.

IV. Remm M, Roosaare M, Stockwell TB, Andreson R, Kaplinski L, Laht S, Kõressaar T, Lepamets M, Brauer A, Kukuškina V, Baden-Tillson H, Piquemal D, Puillandre N, Biass D, Hulo N, Favreau P, Brownstein MJ, Ducancel F, Kauferstein S, Mebs D, Ménez A, Stöcklin R. Expansion of G-protein coupled receptor gene families in the genome of the fish-hunting cone snail *Conus consors*. (submitted to BMC Genomics)

The author made the following contributions to these four articles:

Ref. I: designed the experiment, performed the analysis, and wrote the manuscript;

Ref. II: constructed the pHMMs and participated in the writing of the manuscript;

Ref. III: constructed the pHMMs and participated in both the analysis and writing of the manuscript;

Ref. IV: participated in designing the genome sequence analysis, participated in discovering the conopeptides from the assembled genome sequence, performed an analysis of conopeptide gene structures and participated in writing of the manuscript.

LIST OF ABBREVIATIONS

AA – amino acids
mRNA – messenger ribonucleic acid
UTR – untranslated region
kb – kilo base (1000 bases)
SVMs – support vector machines
PsAAC – pseudo-amino acid composition
IDQD – increment of diversity combined with quadratic discriminant
BLAST – basic local alignment search tool
MSA – multiple sequence alignment
pHMM – profile hidden Markov model
TE – transposable element
PSSM – position specific scoring matrix
MSV – multiple segment Viterbi

INTRODUCTION

As biologists, we classify the world around us into nested categories to ease the process of describing all characteristics for every organism. Knowing that a manul is a cat we automatically know that it has all the features of a living organism, an animal, a chordate, and a feline. This is a lot of information. We also classify chemicals and molecules within living organisms in a similar manner. Gene sequencing, together with other technologies, now provide scientists with huge amounts of new data, however, this data is only useful if we are able to put it into a context compatible with existing knowledge. Currently, we possess more knowledge than any one human can possibly comprehend and the pool of unprocessed data is rapidly growing. To combat this problem, scientists now rely on databases and bioinformatics tools to find the answers they need.

This dissertation provides another tool to help scientists draw conclusions from and classify large datasets.

Nature has had many millions of years to find optimal solutions to most problems involving survival and humans have only recently started to understand and use this highly refined knowledge. Scientists are now searching for biologically active components with pharmaceutical potential, and venomous organisms are a rich source of such molecules. Cone snails have very potent venoms that are composed of mixtures of biologically active peptides termed conopeptides. A small fraction of total conopeptide diversity has been described and already some conopeptides are being used as medicines (Han et al. 2008; Lewis et al. 2012). High-throughput genomic, transcriptomic, and peptidomic methods are able to rapidly generate large datasets that contain, in an unclassified form, knowledge about undiscovered conopeptides. However, identifying and classifying these conopeptides can only be made feasible by developing bioinformatics tools specifically designed for this. A well-designed classification tool would allow one to identify and classify conopeptide sequences.

The first part of this thesis provides an overview of cone snails and conopeptides and also of the model based methods for protein homology searches and annotation.

The second part covers the research that was carried out while developing the method and tool for conopeptide classification. This part also discusses the results of applying our approach to describe conopeptide diversity within the genome of the cone snail *Conus consors*.

LITERATURE REVIEW

1.1. Cone snails and their feeding habits

Cone snail is a common name for species classified within the very diverse marine gastropod genus *Conus*. By January 2014, 761 species were described in the literature and new species descriptions are published every year. Cone snails have cone shaped shells that have been valued by collectors for centuries for their beautiful colors and patterns, however, it is the venom produced by the cone snails that has attracted the attention of researchers. Cone snails are predators that feed on worms, other mollusks, and fish. A slow-moving snail requires a good weapon to catch a fast-moving fish, and cone snails have a great weapon – venom that paralyzes their prey within a few seconds. The venoms of cone snails are complex mixtures of small peptides termed conotoxins or conopeptides, which mostly act on different ion channels and immobilize the prey (Olivera 1997; Han et al. 2008).

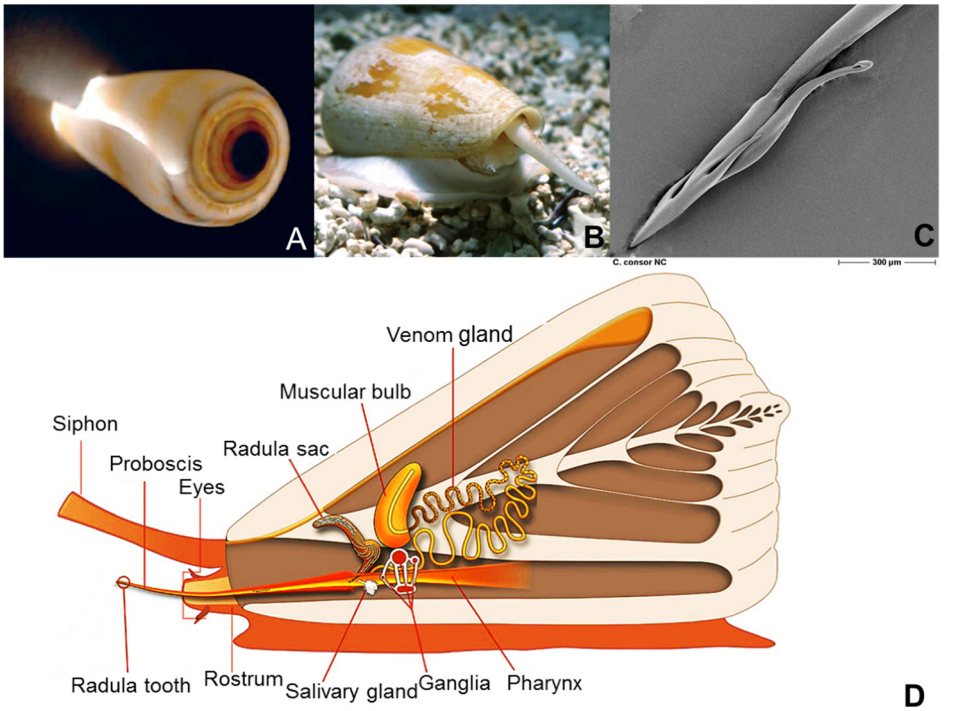

Figure 1. Shell and venomous apparatus of *Conus consors*. A – the shell of *C. consors* (by Thierry Parel), B – a live specimen of *C. consors* (by Thierry Parel). C – a radula tooth of *C. consors* (scanning electron microscopy photo by Dietrich Mebs), D – schematic presentation of the cone snail venom apparatus (by Xavier Sprungli, modified by Dietrich Mebs).

The most common strategy for prey capture among cone snails is the "hook-and-line" approach. The snail fires a hollow venom-filled harpoon, termed the radular tooth, into the fish who is then immobilized and swallowed. Some fish-eating cone snails also use the "net" strategy to capture their prey. They release venom into the water within a school of small fish who then become docile and disoriented. The snail can then easily swallow one or more at a time. Figure 1 shows how a cone snail looks like and the schematics of its venom apparatus.

Cone snails also use their venom for defense when attacked by larger predators such as octopus or fish. Dutertre *et.al* (2014) demonstrated that several *Conus* species produce separate defensive venom that is more complex in composition than the predatory venom. The defense-evoked venom of *Conus geographus* contains high concentrations of paralytic peptides that make it deadly even for humans. The predatory venom is produced in a different part of the venom duct and is more specialized and less potent. The authors speculated that worm-eating cone snails adapted to fish and mollusk diets using the toxins from their defensive venom (Dutertre et al. 2014).

1.2. Conopeptides

Conopeptides are the main active components within all known cone snail venoms. Hundreds of different peptides have been observed in the venom of one snail species (Olivera 2006; Biass et al. 2009) yet only a few of the same conopeptides have been found in more than one species (Mr12.5 from *Conus marmoreus* and Eb12.4 from *Conus eburneus*) (Liu et al. 2010). Thus, the overall number of different conopeptides is estimated to be hundreds of thousands. Three years ago, roughly six thousand conopeptides from over one hundred species had been collected and described within the conopeptide reference database ConoServer (http://www.conoserver.org). This number has increased six times within the last three years (Laht et al. 2012) and is likely to grow even faster with the reducing cost of transcriptome sequencing and advances in proteomics technologies that allow one to analyze snail venom ducts using both technologies simultaneously (Jin et al. 2013; Safavi-Hemami et al. 2014).

Transcriptomic studies of *Conus miles* (Jin et al. 2013), *Conus tribblei* (Barghi et al. 2015), *Conus victoriae* (Robinson et al. 2014), *Conus marmoreus* (Lavergne et al. 2013), and *Conus consors* (Terrat et al. 2012) have found between 53 (*Conus consors*) and 662 (*Conus miles*) different conopeptide transcripts within the venom of one cone snail. This significant difference most probably stems from the different criteria used within each study to report unique conopeptides. In the *C. miles* transcriptome study they used an unassembled transcriptome and reported 495 putative conopeptide transcripts from only one 454 sequencing read (Jin et al. 2013). Most other studies require much more evidence to report a potential new conopeptide.

Conopeptides are synthesized as 60–80 amino acid (AA) long peptide precursors. The N-terminal signal peptide (~ 20 AA) and pro-peptide (20–30 AA) are followed by a mature peptide (10–30 AA) (Figure 2 A). The signal peptide targets the conopeptide secretion while the pro-peptide is required for proper folding. Both the signal and the pro-peptide are cleaved during the maturation process.

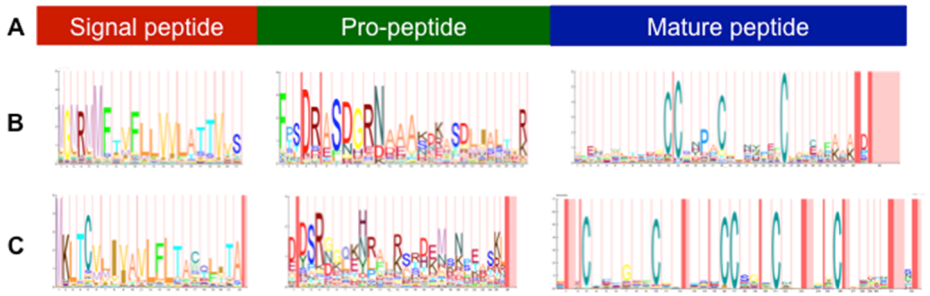

Figure 2. Conopeptide precursor structure. A – The most common structure of conopeptide precursor peptides. B – Sequence logo diagrams of superfamily A. C – Sequence logo diagrams of superfamily O1. Sequence logos were created using LogoMat-M software (Schuster-Böckler et al. 2004).

1.2.1. Conopeptide genes

Conopeptides have mostly been studied at the mRNA and peptide levels, however, little is known about their gene structure. It was discovered already in 1999 that the signal, pro- and mature peptides of O1 superfamily conopeptides are each coded in a different exon separated by long introns (Olivera et al. 1999). Each region of the pre-pro-peptide sequence has diverged at very different rates. This is illustrated by the sequence logos of A and O1 superfamily conopeptides (Figure 2 B and C). While the signal sequence is very highly conserved, even at the nucleotide level within the superfamily (almost no synonymous substitutions), the mature peptide region sequence has a mutation rate that is more than ten times higher (Olivera et al. 1999). This is quite a difference for a translation product that is only ~100 amino acids (AA) long.

For most conopeptide superfamilies (I1, M, O2, O3, P, S, T) the gene structure is similar to the O1 superfamily (Figure 3). The first exon contains a coding sequence for a 5' untranslated region (UTR), signal peptide, and a few codons of a propeptide. The second exon codes for the pro-peptide and the third exon for the mature peptide and 3' UTR. The three exons are separated by introns that are more than three kilobases (kb) long. The I2 superfamily conopeptide genes also have three exons and two introns, however, the order of the functional parts is different. In this case, the pro-peptide comes after the mature peptide in the last exon and is termed a post-peptide. The A superfamily conopeptide genes have two exons and one intron (~1 kb long) that splits the

pro-peptide sequence into two parts. The intron sequences are conserved within superfamilies among different species, just like the signal peptide sequences. (Yuan et al. 2007; Wu et al. 2013).

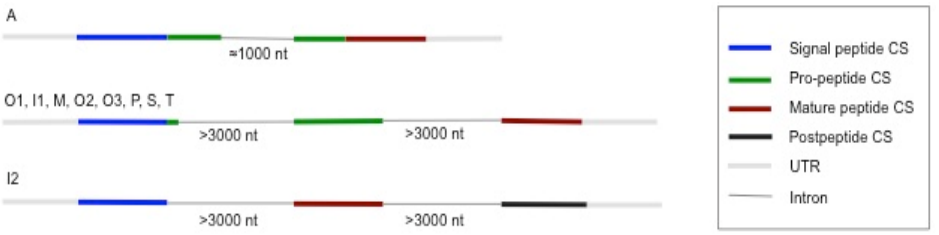

Figure 3. Conopeptide gene structures (Yuan et al. 2007; Wu et al. 2013).

1.2.2. The molecular mechanisms behind known conopeptide diversity

Thousands of peptides with different masses have been detected in the venom of cone snails with modern ultra-sensitive mass spectrometry technologies (Davis et al. 2009; Dutertre et al. 2013; Biass et al. 2015). This high diversity of conopeptides is achieved using several molecular mechanisms.

Each cone snail species expresses roughly one hundred to several hundred conopeptide genes as determined at the transcript level as there are no genome-wide studies of cone snails. Conopeptide genes are thought to be under diversifying selection (Conticello et al. 2001), however, the number of different transcripts is an order of magnitude smaller than the number of different masses detected in the venom. Alternative splicing has been detected for at least one conopeptide (Wu et al. 2013). The number of different conopeptide gene products may also be provided by transcriptomic "messiness" that produces single amino-acid substitutions and alternative peptidase cleavage sites (Jin et al. 2013). The main mechanism for increasing the variability of conopeptides appears to be post-translational modification.

The maturation process of conopeptides includes both cleavage of the signal and pro-peptide and a wide array of post-translational modifications, including the formation of disulfide bridges, C-terminal amidation, and hydroxylation of proline at C-4. To date, 16 different naturally occurring post-translational modifications have been described for conopeptides (Gerwig et al. 2013). The cleavage of the N-terminal pro-peptide is not site-specific and amino acids can also be cleaved from the C-terminus. Both alternative post-translational modifications and alternative cleavage of the pre-pro-peptide have been observed (Dutertre et al. 2013; Lu et al. 2014). For example, the venom of *Conus marmoreus* contains an average of 20 different and a maximum of 72 unique masses per precursor sequence (Dutertre et al. 2013).

1.3. Conopeptide classification

Conopeptides are classified in three different ways – by gene superfamily, by cysteine framework or by pharmacological family (Kaas et al. 2010; Robinson and Norton 2014).

1.3.1. Pharmacological families

The first conopeptides were isolated from the venom of cone snails and described at the peptide level. Thus, the first classification methods adopted were pharmacological family and a cysteine framework.

Pharmacological families are based on the target receptor specificity of the conopeptide and denoted with a Greek letter. In total, 12 pharmacological families have been designated (www.conoserver.org) yet only 167 conopeptides have been assigned to a pharmacological family (Table 1). The main reason for this low number is that the pharmacological family can only be determined with functional experiments and not through the use of bioinformatics methods.

Table 1. Conopeptide pharmacological families (based on Conoserver)

Pharmacological family	Activity	Nr of proteins	Superfamilies	Cysteine frameworks
alpha	Nicotinic acetylcholine receptors (nAChR)	71	A, B3, D, J, L, M, S	I, II, III, IV, VIII, XIV, XX, XXIV
chi	Neuronal noradrenaline transporter	4	T	X
delta	Voltage-gated Na channels (agonist, delay inactivation)	18	O1	VI/VII
epsilon	Presynaptic Ca channels, presynaptic GPCRs	1	T	V
gamma	Neuronal pacemaker cation currents (inward cation current)	4	O1, O2	VI/VII
iota	Voltage-gated Na channels (agonist, no delayed inactivation)	2	I1, M	III, XI
kappa	Voltage-gated K channels (blocker)	11	A, I2, J, M, O1	III, IV, VI/VII, XI, XIV
mu	Voltage-gated Na channels (antagonist, blocker)	25	M, O1, T	III, IV, V, VI/VII
omega	Voltage-gated Ca channels (blocker)	27	O1, O2	VI/VII, XVI, XXVI
rho	Alpha1-adrenoceptors (GPCR)	1	A	I
sigma	Serotonin-gated ion channels (GPCR)	1	S	VIII
tau	Somatostatine receptor	2	T	V

1.3.2. Cysteine frameworks

Classification of conopeptides into cysteine frameworks is based on the cysteine patterns of mature conopeptides. The cysteine frameworks are defined by the number of cysteines and the number of residues (none or at least one) between consecutive cysteines. Currently, researchers have defined 26 cysteine patterns that contain either 4, 6, 8, or 10 cysteine residues, each designated with a roman numeral (Table 2). The cysteine frameworks are not exclusive to a superfamily – one superfamily often contains conopeptides with different cysteine frameworks and different cysteine frameworks can be found in different superfamilies (Kaas et al. 2010). Conopeptides with less than 4 cysteines can not be classified using this method.

Table 2. Cysteine frameworks (based on ConoServer and (Robinson and Norton 2014)).

Cysteine framework	Nr of sequences	Superfamilies	Cysteine framework	Nr of sequences	Superfamilies
I	359	A, M, O1, T	**XV**	27	N, O2, V
II	3	A, M	**XVI**	12	M, O1, T
III	329	M	**XVII**	1	Y
VI	55	A, M	**XVIII**	2	
V	203	T	**XIX**	2	
VI/VII	646	A, H, I1, I3, M, O1, O2, O3	**XX**	21	D
VIII	20	B2, S	**XXI**	5	
IX	35	M, O1, P	**XXII**	10	A, E
X	10	T	**XXIII**	6	K
XI	104	I1, I2, I3	**XXIV**	1	B3
XII	49	I4, O1	**XXV**	1	
XIII	2	G	**XXVI**	1	
XIV	78	A, I2, J, L, M, O1, O2, P			

1.3.3. Conopeptide gene superfamilies

Conopeptides are classified into gene superfamilies based on the similarity of their signal sequence (Table 3). The signal sequences show little homology between superfamilies excepting standard features of a signal sequence, such as having a methionine at the first position and a central hydrophobic region (Kaas et al. 2010) (Figure 1 B and C). Classification into gene superfamilies is supported by

evolutionary evidence that shows that members of different gene superfamilies are genetically and evolutionarily divergent (Puillandre et al. 2012).

There are 26 superfamilies in the conopeptides reference database ConoServer (www.conoserver.org), however, 35 are listed in a recent review about conotoxin superfamilies (Robinson and Norton 2014). The main difference comes from the fact that ConoServer does not classify cysteine-poor (two cysteines or less) conopeptides into superfamilies. They are classified separately into classes, however, differentiation between cysteine-rich and cysteine-poor conopeptides has been shown to have no phylogenetic meaning (Puillandre et al. 2012).

Table 3. Conopeptide gene superfamilies (based on ConoServer and Robinson and Norton 2014).

Super-family	Nr of sequences	Cysteine frameworks	Super-family	Nr of sequences	Cysteine frameworks
A	276	I, II, IV, VI/VII, XIV, XXII	**M**	443	I, II, III, IV, VI/VII, IX, XIV, XVI, –
B(1)	18	–	**N**	4	XV
B2	2	VIII	**O1**	575	I, VI/VII, IX, XII, XIV, XVI
B3	1	XXIV	**O2**	133	VI/VII, XIV, XV, –
C	4	–	**O3**	43	VI/VII
D	28	XX	**P**	12	IX, XIV
E	1	XXII	**Q**	12	IX, XIV
F	2		**S**	21	VIII
G	1	XIII	**T**	234	I, V, X, XVI
H	10	VI/VII	**U**	3	VI/VII
I1	26	VI/VII, XI	**V**	2	XV
I2	60	XI, XIV	**Y**	1	XVII
I3	9	VI/VII, XI	**Con-ikot-ikot**	7	homodimer, XXI
I4	3	XII	**ConoCAP**	1	–
J	30	XIV	**Conopressin/ conophysin**	7	–
K	4	XXIII	**Conkunitzin**	3	Kunitz-fold
L	14	XIV			

Some cysteine-poor conopeptides that were previously classified into families have been reclassified into superfamilies, together with cysteine-rich conopeptides (conomarphins and contryphans have moved into the M and O2 superfamilies, respectively). Other cysteine-poor conopeptides have unique signal sequences and have been placed within their own superfamilies (conantokins and contulakins can now be refered to as superfamily B and C, respectively) (Robinson and Norton 2014).

Every transcriptomic study has revealed conopeptide sequences that cannot be placed into any of the existing superfamilies. Some authors confidently declare new superfamilies (Lavergne et al. 2013) while others assign them to temporary superfamilies (Biggs et al. 2010; Jin et al. 2013). There are currently 13 superfamilies from *Conus californicus* in ConoServer that are termed "divergent". There is no question if the conopeptides within these divergent superfamilies exist, however, it is not clear if a new superfamily should be declared based on a few sequences from a single species. Robinson and Norton (2014) also note that superfamilies identified using only one or two *Conus* species should be considered putative.

1.3.4. Bioinformatics methods for conopeptide classification

Several different bioinformatics methods have been developed or suggested to classify conopeptides into gene superfamilies. The first and the most common method is performing a homology search against previously described conopeptides. This approach is useful when the novel conopeptide sequence contains a signal peptide because signal peptides have high conservation within a superfamily and lack homology between different superfamilies.

ConoServer includes a conopeptide sequence analysis tool called ConoPrec (http://www.conoserver.org/?page=conoprec) that accepts both nucleotide and protein sequences. First, it finds the signal sequence cleavage site with the SignalP algorithm and then uses the defined signal sequence to search for similar signal sequences among the conopeptide sequences already in the ConoServer database. The similarity cut-off for superfamily designation is 90%, and when this is not exceeded, the user will obtain the maximum percentage of identity within each superfamily for further evaluation. In addition to gene superfamily prediction, the ConoPrec tool also provides information on propeptide cleavage locations and possible post-translational modifications.

When the precursor sequence of a conopeptide is not known and the classification has to be made based on a mature peptide sequence alone, a simple homology search is often not sufficient. Use of support vector machines (SVMs) along with pseudo-amino acid composition (PsAAC) has been shown to be effective for the classification of mature conotoxins. The sensitivity of this approach was shown to be in the range of 84.0–94.1% and the specificity between 80.0–95.5% for the superfamilies tested (A, M, O, T) (Mondal et al. 2006). An algorithm called increment of diversity combined with quadratic discriminant

(IDQD) has also been suggested for conopeptide gene superfamily classification (Lin and Li 2007). This algorithm only uses the mature peptide sequence and provided an overall sensitivity of 88% and a specificity of 91% for the superfamilies tested (Lin and Li 2007). Although these methods show rather good performance for conopeptide superfamily prediction, they have not been widely adopted by other conopeptide researchers. The most probable reason for this is the lack of usability.

If one's goal is to discover drug candidates for either further development or as research tools, classification into gene superfamilies provides a limited amount of information. Of more interest is discovering the most likely targets of the newly discovered conopeptides. This task is much more complicated than defining the gene superfamily because there is no obvious sequence similarity between the mature conopeptide sequences within each pharmacological family. To the best of my knowledge, no tool is able to predict the conotoxin pharmacological family based on its sequence data. The only attempt to distinguish pharmacological families was carried out by Lin and Li using the IDQD method. The amount of data available for testing was very limited and they only attempted to determine the pharmacological families within the O-superfamily (omega, delta, and other). This method provided a sensitivity of 72% and a specificity of 78% (Lin and Li 2007). Another tool, called iCTX-Type, predicts the type of ion channel that the conopeptide targets (K-channel, Ca-channel or Na-channel) using only the sequence of the mature peptide. This tool uses a statistical prediction method that applies pseudo amino acid composition to build vectors from the protein sequences followed by classification using support vector machines. When tested on an independent dataset, the iCTX-Type had a success rate of 91.7%, 91.9%, and 90.5% for K-, Na-, and Ca-conotoxins, respectively (Ding et al. 2014). The authors have also set up a web server for the iCTX-Type tool to make it convenient for experimental scientists. While the identification of the target ion-channel type is not as specific as prediction of pharmacological families, it is a big step forward in our ability to predict conopeptide function and significantly narrows the number of laboratory tests required to determine the function of novel conopeptides.

1.4. Models for protein homology searches and annotation

Proteins, RNAs, DNA repeats, and other biological sequences can usually be organized into families of related sequences. Similarity at the sequence level very often indicates similarity in function as well. However, functional studies are often time-consuming, resource intensive, and complicated. The ability to predict function based on sequence data can aid and simplify this process. The most popular tool for similarity searches is BLAST, or Basic Local Alignment Search Tool, that works by performing pairwise similarity searches (Altschul et al. 1990). BLAST is relatively fast and has modified versions for specific applications. It also has an easy-to-use web interface and is kept up to date.

While it is often very useful, pairwise comparison does not provide much information about which residues in the sequence are more conserved or where insertions or deletions are allowed. This information can be revealed when sequences of the same protein family are aligned into multiple sequence alignment (MSA). The conserved motifs can be described as sequence patterns using a qualitative consensus sequence (regular expressions). This is useful for short motifs such as the active sites within enzymes that are highly conserved. Pattern matching is fast but we lose information on the relative frequency of each allowed amino acid at a given position and it is very difficult to write long patterns. The information seen in the MSA can also be "saved" into profiles or models in the form of position-specific scores for residues and position-specific penalties for gaps or deletions (Gribskov et al. 1987; Eddy 1998; Sigrist et al. 2002). Two different methods for protein modeling that were adopted in the studies presented herein are described below in more detail.

1.4.1. Profile hidden Markov models (pHMMs)

Profile HMMs are statistical models of multiple sequence alignments (MSA). A pHMM can incorporate more information about similar sequences than any single representative sequence. A simple schematic representation of a pHMM is provided in Figure 4. While generating a pHMM, probabilities for three different states are calculated for each position in the MSA. A 'match' state models the distribution of amino acid or nucleotide residues allowed in that position. For a nucleotide sequence a 'match' state has four emission probabilities and for a protein sequence 20 emission probabilities – one for each possible residue in that position. An 'insert' state models the probability of inserting one or more residues between that position and the next and also has the emission probabilities. A 'delete' state models the probability of deleting the consensus residue (Eddy 1998). All scoring parameters are set based on probability theory, including the gap and insertion scores. This makes pHMM-based methods amenable to automation and therefore easy to apply in large-scale analysis (HMMER userguide, http://hmmer.janelia.org).

The most popular and best-supported tool for using pHMMs in sequence analysis is HMMER (http://hmmer.janelia.org). Until recently, HMMER processed sequences about 100 times slower than BLAST (Altschul et al. 1997). In the era of huge data, speed outweighs sensitivity in most cases. The latest version of HMMER, HMMER3 uses an acceleration heuristic algorithm called multiple segment Viterbi (MSV) filter that makes the searches as fast as BLAST. Despite the added filtering steps, HMMER is still as sensitive as before. The speed makes it applicable for large-scale similarity searches with both single sequences (including iterative searching) and multiple sequence alignment based pHMMs (Eddy 2011). The increased speed has also enabled the development of a web-based tool for HMMER (Finn et al. 2011)

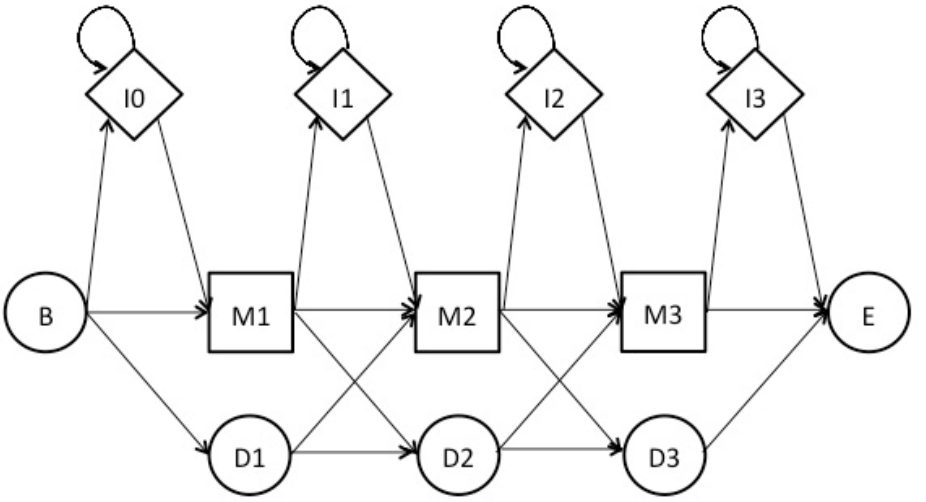

Figure 4. Schematic representation of a profile hidden Markov model (pHMM). Three match states (squares labeled M1–M3) represent three columns in a multiple sequence alignment (MSA). Insert states (diamonds labeled I0–I3), delete states (circles labeled D1–D3), begin (B) and end (E) states are also included. Arrows indicate state transition probabilities.

pHMMs are widely used in computational biology. The first and most common application is to model protein domain families in the search for homologous proteins. Pfam is a database of curated protein families where each family is defined by two alignments and a pHMM. A curator selects a seed alignment of a representative sequence in the family and builds a profile HMM from it. The profile HMM is searched against the full sequence database and all protein sequences that exceed a given gathering threshold are included in the full alignment of the protein family. The gathering thresholds for each family are set by the curator to exclude false positive hits (Finn et al. 2014).

In metagenomics, a pHMM based tool HMM-FRAME helps by accounting for sequencing errors thereby providing annotations for a larger portion of the sequencing reads (Zhang et al. 2012). Profile HMMs have also been used to detect viral sequences (Skewes-Cox et al. 2014) and antibiotic resistance functions (Gibson et al. 2015) from metagenomic data.

In addition to protein sequences, pHMMs can be applied to nucleotide sequences. The sensitivity of finding remote homologues has been used to annotate repeats within genomes. The Dfam database of transposable elements (TEs) is built using pHMMs and the models can be used to annotate all TEs in a genome sequence (Wheeler et al. 2013).

Profile hidden Markov models have also been used to classify HIV strains (Dwivedi and Sengupta 2012).

1.4.2. Position specific scoring matrices (PSSMs)

Position specific scoring matrices (also called generalized profiles) are another way of describing the multiple sequence alignment of proteins for similarity searches. PSSMs provide numerical weights for each possible match or mismatch between a sequence residue and a profile position. By using an amino acid substitution matrix, appropriate weights can be assigned to residues not observed at a given alignment position. PSSMs also enable one to model insertions and deletions by applying position-specific penalties (Gribskov et al. 1987; Sigrist et al. 2002)

Generalized profiles are used within the PROSITE database (Sigrist et al. 2010). The developers of PROSITE have also developed and maintain tools for the generation of PSSMs generation and for performing database searches. For the PROSITE profiles annotated multiple sequence alignments are used. Curators, who are experts in their field, perform the annotations manually. The curators also set the cut-off scores to decide when a sequence is considered to be similar enough to be classified into a given protein family (Sigrist et al. 2010). The manual curation has both advantages and disadvantages at the same time. Expert involvement guarantees that profiles that are best able to distinguish between true positives and negatives. The biggest disadvantage is that expert curation takes a lot of time.

PSSMs are not well suited to model patterns with variable length or positional dependencies, or patterns that contain insertions or deletions.

PSSMs have been used to build specialized databases of proteins, e.g., PeroxiBase (Koua et al. 2009). PSSMs also display good performance for the prediction of membrane transport proteins and their substrate specificity (Mishra et al. 2014).

AIMS OF THE STUDY

Our workgroup was one of the partners in the project Cone Snail Genome for Health (CONCO). The aim of the CONCO project was to discover novel conopeptides and possible drug candidates from the venom of *Conus consors*. The project involved several groups that performed different tasks including proteomic analysis of milked and dissected venom, functional analysis of venom components, and sequencing of the transcriptomes of several tissues together with the entire genome. The problem we set out to solve was very practical – to develop a method for discovering and classifying conopeptides using both the venom duct transcriptome and the *Conus consors* genome.

The aims of this study were:

a) Develop a method to identify and classify conopeptides.
b) Identify and describe the conopeptide genes within the genome of *Conus consors.*

RESULTS AND DISCUSSION

2.1. pHMMs for conopeptide identification and classification (ref. I)

As described in section 2.3.4, there are several methods that can be used to identify and classify conopeptides. However, none of them is especially amenable for high-throughput data analysis and most are limited to only some of the known superfamilies. Our aim was to build a set of models that could be used to annotate all conopeptide superfamilies described to date even from partial sequences obtained using mass-spectrometry and short-read sequencing data.

Based on the available conopeptide sequences 62 pHMMs were built for the 24 conopeptide superfamilies described at the time of the study (Table 1 in reference I). Separate models were made for each functional part of the conopeptide for each superfamily, when possible. Five disulfide-poor conopeptide superfamilies had only been described at the mature peptide level by that time, so no signal and propeptide model could be constructed for these groups. The rationale behind constructing three separate models for each functional part was to create models that are more versatile and that can be used together with proteomic and genomic data where the three functional parts are not sequential.

16 out of the 24 conopeptide superfamilies contained less than 10 sequences at the time of analysis. Therefore, we needed to determine how many sequences are required to train pHMM models that possess sufficient sensitivity and specificity. For this we trained three superfamilies that contain over 100 conopeptide precursor sequences with 2, 3, 5, 10, 20, 30, 40, 50, 60, 70, and 80 sequences. For highly conserved signal peptide models, a sensitivity of 100% was achieved already with two sequences. The maximum sensitivity of propeptide models was reached using between 10–20 training sequences. For the mature peptide models, more than 30 sequences were required to obtain maximum sensitivity. The specificity was always nearly 100% with the exception of the O1 superfamily whose mature peptide model incorrectly identified sequences from the I3 superfamily, which has the same cysteine framework. From these results we conclude that the specificity of the conopeptide models trained on less than ten sequences can be trusted, however, the sensitivity would improve if more sequences were included. When full precursor sequences for superfamilies with only mature peptides become available, the sensitivity will increase significantly because both the propeptide and especially the signal peptide models are more sensitive.

We also determined if the conopeptide pHMMs are specific for conopeptides by searching the entire UniProtKB/SwissProt protein database using each of the 62 pHMMs. In total, we found only 111 false-positive predictions and 57 pHMMs yielded no false-positive matches. The I1, I3, O1, O2 superfamily mature peptide models had several and the P superfamily mature peptide model had only one false-positive match. All of these false positives were cysteine-rich peptides that

contain a knottin domain that has a similar structure to conotoxins within the O and I superfamilies. Overall, the specificity of these models is high and the few false-positives found were easily excluded using manual analysis.

To see how well the conopeptide pHMMs work in practice, we performed two experiments. First the pHMMs we used to classify a set of 53 novel conopeptide precursors found within the venom duct transcriptome of *Conus consors* (Terrat et al. 2012). Only propeptide and mature peptide pHMMs were used to classify this set of conopeptides because the presence of signal sequence guarantees 100% sensitivity and specificity. The propeptide models gave neither false positives nor false negatives on this set of conopeptides. The mature peptide pHMMs were able to correctly classify 79% of the sequences in the test set. Two sequences from the O superfamily were incorrectly classified into the I3 superfamily (see the explanation above). The lower sensitivity of mature peptide models can be explained by the high variability of mature peptides within a given superfamily. This test also shows that it is important to include all conopeptide superfamilies, including the ones that only contain a few sequences. By discovering and describing new members of smaller superfamilies we were able to reduce the bias that has been introduced due to the superfamily-specific identification of conotoxins (Luo et al. 2006; Yuan et al. 2007; Liu et al. 2009; Liu et al. 2010).

The second test was performed on a set of 2410 putative conotoxins identified from the venom-duct transcriptome of *Conus bullatus* using homology search with BLASTX (Hu et al. 2011). The authors were able to classify 543 (23%) of the conotoxin contigs into a superfamily based on the similarity of the signal sequences. We were able to classify 1188 (49%) of these putative conotoxins using our pHMMs. Many of the 2410 putative conopeptide contigs did not contain full precursors and lacked the entire signal sequence. This is illustrated by the fact that 766 conotoxin contigs were classified without the signal peptide model. These results indicate that the three models per superfamily approach is justified and the possibility to classify conopeptides without a signal sequence is also very useful.

In my opinion, the pHMMs are a useful and relatively easy to use tool for conopeptide identification from large data sets. The models are applicable for both transcriptome and genome data analysis (see section 4.3).

Since our work was published, other researchers have started using pHMMs to identify and classify conopeptides. Robinson *et al*. used pHMMs together with BLASTX to search for conopeptides from the transcriptome of *Conus victoriae* with good results. They described 114 different conopeptides from 20 superfamilies, the largest number of conopeptides discovered in one study (Robinson et al. 2014). The ConoSorter tool developed by the researchers behind the ConoServer database uses pHMMs together with regular expression to identify conopeptides (Lavergne et al. 2013). Compared with our pHMMs, ConoSorter is less specific. A direct comparison between ConoSorter and our pHMMs was performed against the UniprotKB/Swiss-Prot database and revealed specificities

of 99.94% and 99.98%, with 738 and 111 false positives, respectively (Lavergne et al. 2013). This comparison is not entirely fair because the number of proteins in the UniprotKB/Swiss-Prot database was 540251 during the ConoSorter study and 531473 during the testing of our pHMMs. However, it is unlikely that the 1.6% increase in the size of the database size would result in an almost seven-fold increase in the number of false positives.

2.2. PSSMs and pHMMs complement each other for conopeptide classification (ref II, III)

Profile HMMs are a good tool for conopeptide classification, however, there is room for improvement. For this we chose to apply position specific scoring matrices (PSSMs) to improve the classification.

In this study we constructed models for 14 gene superfamilies with at least three precursor sequences available at the time of analysis. The A, O1 and O2 superfamilies were split into two subsets based on the number of cysteines in the mature peptide (A_4, A_6, O1_6, O1_8, O2_6, O2_8). The smaller superfamilies were excluded, because the PSSMs built using only one or two sequences cannot add sensitivity or specificity compared with homology searches.

As before, separate models for each part of the precursor sequence were built. The construction of PSSMs takes somewhat more effort than pHMM training and profiles need to be calibrated against a large database and cutoffs tuned manually to avoid false positive matches. In addition, "compete lines" were added to the profiles for a competition step where the highest scoring profile can be chosen when more than one profile has a match to the sequence. Altogether, we generated and tested 97 models (47 pHMMs and 50 PSSMs). We tested the ability of each model to classify conopeptides using a test set consisting from previously described conopeptides that were not used during the construction of the models. As expected, the signal-based models performed well, however, both propeptide and mature peptide models were not far behind. Combining the classification results from these two sets of models significantly increased the number of correct classifications for each superfamily. The PSSM approach had more predictive power for highly variable motifs. In the T and M superfamilies, PSSMs for mature peptides displayed 75% and 73% sensitivity, however, use of the pHMMs alone provided only 39% and 36% sensitivity, respectively. The PSSMs failed to classify D superfamily mature peptides, yet pHMMs were able to classify D superfamily conopeptides correctly without any difficulties. The combined prediction sensitivity was 91% with an accuracy of 92% using only the mature peptide models. Combined classification performs better than all other previously developed methods for mature peptide classification. For the SVMs, an accuracy of 88% was achieved (Mondal et al. 2006) and for IDQD an overall sensitivity of 88% was achieved (Lin and Li 2007).

Another advantage of model-based methods is their usability. To make it accessible for non-bioinformaticians we built a web-based tool we call ConoDictor that is designed for conopeptide classification. ConoDictor takes amino acid sequences as an input and users can either choose to use the pHMMs and PSSMs we have built or upload their own models. The results are represented to the user with various levels of detail starting with an overview of the combined results and going down to individual positions, scores, and e-values for each PSSM and pHMM match found. The results can either be viewed online or downloaded in a spreadsheet or text format for further analysis.

ConoDictor can be used to discover sequences from full transcriptome data, although the PSSM search is somewhat slow. With large datasets it may be more practical to only use the pHMMs. ConoDictor can also be used to classify sequences identified as putative conopeptides with other methods. One application could be classification of peptides discovered from the venom of cone snails using mass spectrometry data.

Because this tool is web-based it should be easier for other researchers to use our models. However, in the field of conopeptide research, which has possible applications in pharmacology, other researchers may not be very eager to upload their newly discovered sequences to a competitor's server. A competing tool, ConoSorter, has been designed especially for the analysis of transcriptome data (Lavergne et al. 2013). It is available as downloadable software that can be installed locally thereby eliminating the concern for data security. Similar to ConoDictor, ConoSorter also combines two identification/classification schemes by combining patterns (regular expressions) with pHMMs and has separate models for each functional part of the precursor sequence. The superfamilies were divided into clusters of closely related sequences in order to establish subsets that best describe each superfamily which resulted in 777 models for classification (Lavergne et al. 2013). The choice of identification/classification approach depends on both the data and the goals of the study. ConoDictor is more conservative and provides fewer false positive results while ConoSorter should be more capable of discovering novel conopeptide superfamilies.

2.3. Conopeptides in the genome of *Conus consors* (ref IV)

Prior to constructing the conopeptide pHMMs, we intended to use them to find and classify conopeptides from the genome of *Conus consors*. The genome was sequenced using Roche 454 and Illumina paired end technologies with approximately 19x coverage. The size of the *C. consors'* genome is approximately 3 GB. Approximately 20% of the genome contains low-complexity (mononucleotide, dinucleotide, trinucleotide and tetranucleotide) repeats and in total 49% of the genome consists of repeats. The high repeat content and the lack of

a reference genome complicated the assembly process. The assembly of conopeptide coding genes was especially challenging due to a relatively high amount of simple repeats in the introns and high similarity between the signal sequences of different conopeptides from the same superfamily.

To identify conopeptide genes in the assembled genome, we used the known conopeptide sequences available in the UniProtKB/Swiss-Prot database (975 peptides from more than 30 different superfamilies), the 64 conopeptide pHMMs described in Ref. I, peptide sequences from the *C. consors* proteome sequencing (126 peptides from 8 different superfamilies) (Violette et al. 2012), and conopeptide precursor sequences predicted from the transcriptome data (135 distinct precursor sequences from 23 different superfamilies). We used four different approaches to both maximise the number of conopeptide discoveries and gather more evidence for each potential conopeptide gene identified. Conopeptides from the transcriptome were predicted using the first three methods.

The fragmented assembly made it difficult to determine the exact number of conopeptide genes. We aligned the fragments of putative conopeptide genes together with both similar transcripts and previously described conopeptide precursor sequences into multiple sequence alignments to estimate the number of different genes. This approach allowed us to identify 27 genes of previously described conopeptides, 141 novel conopeptide genes, and 46 dubious conopeptide genes that were only found in the genome and not from the transcriptome or proteome. In total, we found 214 putative conopeptide genes from 21 gene superfamilies (table 4). There are probably more conopeptide genes in the genome of *C. consors* than we report, however, we decided to only report the genes we are most confident are conopeptides. We cannot distinguish between pseudogenes so not all reported conopeptide genes are necessarily expressed as active conopeptides.

The largest conopeptide gene superfamilies in the *C. consors* genome are A (29 genes), M (29 genes, including 19 conomarphin-like genes) and O1 (33 genes). These superfamilies are also the most abundant in the transcriptome and injectable venom both by the number of different conopeptides and by their expression levels (Terrat et al. 2012; Violette et al. 2012).

We identified the exon-intron structure of 15 conopeptide genes from 13 superfamilies (Table 5 and figure 5). The gene structures of representatives from the I1, M, O1, O2 and S gene superfamilies are the same as previously described (Yuan et al. 2007; Wu et al. 2013) and is considered the classical gene structure of conopeptides where each part of the precursor is encoded in a separate exon with a few amino acids of the propeptide sometimes coded by the signal and/or mature peptide exon. We were unable to determine the length of the introns except for the intron between the S superfamily conopetpide signal and propeptide exons, which is 995 nt long. The structure of the one A superfamily conopeptide gene we were able to reconstruct is also the same as previously described with a single intron between the signal peptide exon and

the exon that encodes both the pro- and mature peptides (Olivera 1997; Yuan et al. 2007). The gene structure of the P superfamily conopeptide described by Wu et al. (2013) contained three exons and two introns. We found a gene with two exons and one intron. The first exon contains the sequence for the signal peptide and most of the propeptide and the second exon encodes the reminder of the propeptide and the mature peptide.

Table 4. Number of conopeptides found from the genome using both superfamily and direct evidence approaches.

Superfamily	UniProt genes (present in our datasets)	Novel genes	Dubious genes	Total
A	8	13	8	29
B	1	3	1	5
C	0	2	3	5
Conkunitzin	0	6	0	6
ConoCAP	0	3	0	3
Conodipine	1	3	2	6
Conophysin	0	8	4	12
Conoporin	1	13	0	14
I1	0	5	0	5
I2	0	3	0	3
I3	0	2	1	3
J	0	4	1	5
K	0	3	2	5
M	8	18	3	29
O1	8	18	7	33
O2	0	8	4	12
O3	0	7	2	9
P	0	5	1	6
S	0	9	1	10
T	0	7	5	12
V	0	1	1	2
	27	**141**	**46**	**214**

We reported the gene structure of six gene superfamilies that were previously not known. The I3 superfamily representative was found to have the classical structure while the C and conkunitzin superfamilies have two exons and one intron. The first exon codes for the signal peptide and part of the propeptide and the second exon codes for the other part of propeptide and mature peptide. Genes from the B, J and conodipine superfamilies do not have any introns and the full precursor sequences are encoded by one exon. While the B superfamily conopeptide has a signal, pro- and mature peptide, the single conodipine gene did not contain a propeptide and the two genes from the J superfamily have a post-peptide in addition to the first three classical domains.

Table 5. Gene structures of conopeptide superfamilies

Superfamily	Gene structure (previously reported)	Gene structure (this study)
A	2 exons, 1 intron	2 exons, 1 intron
B	–	1 exon, no introns
C	–	2 exons, 1 intron
Conkunitzin	–	2 exons, 1 intron
Conodipine	–	1 exon, no introns
I1	3 exons, 2 introns	3 exons, 2 introns
I2	3 exons, 2 introns	–
I3	–	3 exons, 2 introns
J	–	1 exon, no introns
M	3 exons, 2 introns	3 exons, 2 introns
O1	3 exons, 2 introns	3 exons, 2 introns
O2	3 exons, 2 introns	3 exons, 2 introns
P	3 exons, 2 introns	2 exons, 1 intron
S	3 exons, 2 introns	3 exons, 2 introns

It has been suggested that the three exon-two intron gene structure is the original conopeptide gene structure and intron loss has occurred in the evolutionally younger A superfamily (Yuan et al. 2007; Wu et al. 2013). This suggests that the B and J superfamilies that have lost both introns are even younger than the A superfamily. The J and B superfamilies are closely related, however, the C, P and A superfamilies with one intron are phylogenetically distant (Puillandre et al. 2012). This suggests there has been separate events of intron loss. The observed intron loss seems to be an ongoing process when one considers the P superfamily, however, more information is required to confirm this.

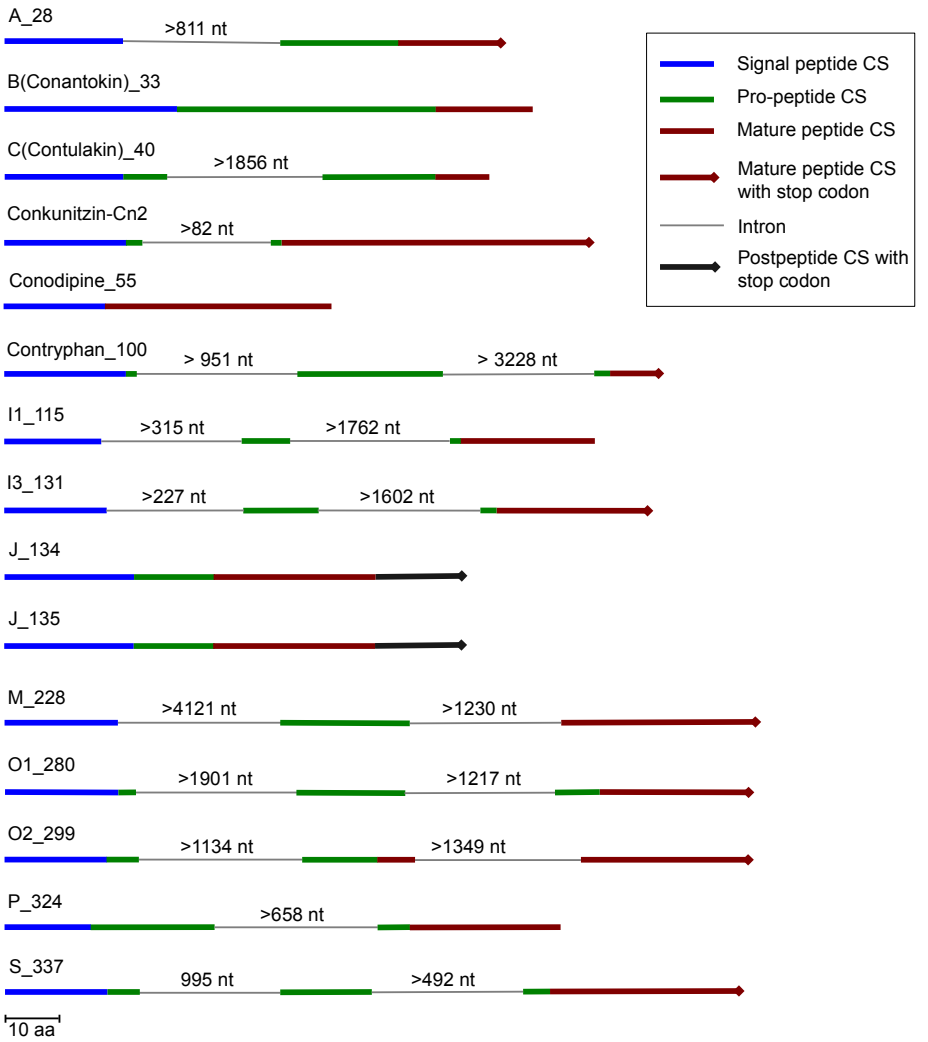

Figure 5. The gene structures discovered from the genome of *Conus consors* (Remm et.al, submitted). CS – coding sequence.

CONCLUSIONS

We developed a specific and sensitive method that is able to identify and classify novel conopeptide genesdiscovery and classification. This method uses both pHMMs and PSSMs and by combining the results from both types of model we were able to increase the sensitivity of identification. These two model sets can either be used together or separately. To make our work more accessible to other researchers, we developed a web-based tool we call ConoDictor that uses both pHMMs and PSSMs to classify conopeptide genes.

In total, we discovered 214 genes that encode conopeptide sequences within the genome of *Conus consors*, and 187 of these conopeptide sequences are novel. We also described the gene exon-intron structure for 13 conopeptide gene superfamilies. For six of these superfamilies, the gene structure was previously not known and for the P superfamily we described a structure that is different from the one reported previously.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Barghi N, Concepcion GP, Olivera BM, Lluisma AO. 2015. High Conopeptide Diversity in Conus tribblei Revealed Through Analysis of Venom Duct Transcriptome Using Two High-Throughput Sequencing Platforms. Mar Biotechnol (NY). 17:81–98.

Biass D, Violette A, Hulo N, Lisacek F, Favreau P, Stocklin R. 2015. Uncovering intense protein diversification in a cone snail venom gland using an integrative venomics approach. J Proteome Res. 14:628–638.

Biass D, Dutertre S, Gerbault A, Menou JL, Offord R, Favreau P, Stocklin R. 2009. Comparative proteomic study of the venom of the piscivorous cone snail Conus consors. J Proteomics. 72:210–218.

Biggs JS, Watkins M, Puillandre N, Ownby JP, Lopez-Vera E, Christensen S, Moreno KJ, Bernaldez J, Licea-Navarro A, Corneli PS et al. . 2010. Evolution of Conus peptide toxins: analysis of Conus californicus Reeve, 1844. Mol Phylogenet Evol. 56:1–12.

Conticello SG, Gilad Y, Avidan N, Ben-Asher E, Levy Z, Fainzilber M. 2001. Mechanisms for evolving hypervariability: the case of conopeptides. Mol Biol Evol. 18:120–131.

Davis J, Jones A, Lewis RJ. 2009. Remarkable inter- and intra-species complexity of conotoxins revealed by LC/MS. Peptides. 30:1222–1227.

Ding H, Deng EZ, Yuan LF, Liu L, Lin H, Chen W, Chou KC. 2014. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. Biomed Res Int. 2014:286419.

Dutertre S, Jin AH, Kaas Q, Jones A, Alewood PF, Lewis RJ. 2013. Deep venomics reveals the mechanism for expanded peptide diversity in cone snail venom. Mol Cell Proteomics. 12:312–329.

Dutertre S, Jin AH, Vetter I, Hamilton B, Sunagar K, Lavergne V, Dutertre V, Fry BG, Antunes A, Venter DJ et al. . 2014. Evolution of separate predation- and defence-evoked venoms in carnivorous cone snails. Nat Commun. 5:3521.

Dwivedi SK, Sengupta S. 2012. Classification of HIV-1 sequences using profile Hidden Markov Models. PLoS One. 7:e36566.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol. 7:e1002195.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics. 14:755–763.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29–37.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al. . 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–30.

Gerwig GJ, Hocking HG, Stocklin R, Kamerling JP, Boelens R. 2013. Glycosylation of conotoxins. Mar Drugs. 11:623–642.

Gibson MK, Forsberg KJ, Dantas G. 2015. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 9:207–216.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A. 84:4355–4358.
Han TS, Teichert RW, Olivera BM, Bulaj G. 2008. Conus venoms – a rich source of peptide-based therapeutics. Curr Pharm Des. 14:2462–2479.
Hu H, Bandyopadhyay PK, Olivera BM, Yandell M. 2011. Characterization of the Conus bullatus genome and its venom-duct transcriptome. BMC Genomics. 12:60-2164-12-60.
Jin AH, Dutertre S, Kaas Q, Lavergne V, Kubala P, Lewis RJ, Alewood PF. 2013. Transcriptomic messiness in the venom duct of Conus miles contributes to conotoxin diversity. Mol Cell Proteomics. 12:3824–3833.
Kaas Q, Westermann JC, Craik DJ. 2010. Conopeptide characterization and classifications: an analysis using ConoServer. Toxicon. 55:1491–1509.
Koua D, Cerutti L, Falquet L, Sigrist CJ, Theiler G, Hulo N, Dunand C. 2009. PeroxiBase: a database with new tools for peroxidase family classification. Nucleic Acids Res. 37:D261–6.
Laht S, Koua D, Kaplinski L, Lisacek F, Stocklin R, Remm M. 2012. Identification and classification of conopeptides using profile Hidden Markov Models. Biochim Biophys Acta. 1824:488–492.
Lavergne V, Dutertre S, Jin AH, Lewis RJ, Taft RJ, Alewood PF. 2013. Systematic interrogation of the Conus marmoreus venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene superfamilies. BMC Genomics. 14:708-2164-14-708.
Lewis RJ, Dutertre S, Vetter I, Christie MJ. 2012. Conus venom peptide pharmacology. Pharmacol Rev. 64:259–298.
Lin H, Li QZ. 2007. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun. 354:548–551.
Liu Z, Yu Z, Liu N, Zhao C, Hu J, Dai Q. 2010. cDNA cloning of conotoxins with framework XII from several Conus species. Acta Biochim Biophys Sin (Shanghai). 42:656–661.
Liu Z, Xu N, Hu J, Zhao C, Yu Z, Dai Q. 2009. Identification of novel I-superfamily conopeptides from several clades of Conus species found in the South China Sea. Peptides. 30:1782–1787.
Lu A, Yang L, Xu S, Wang C. 2014. Various conotoxin diversifications revealed by a venomic study of Conus flavidus. Mol Cell Proteomics. 13:105–118.
Luo S, Zhangsun D, Wu Y, Zhu X, Xie L, Hu Y, Zhang J, Zhao X. 2006. Identification and molecular diversity of T-superfamily conotoxins from Conus lividus and Conus litteratus. Chem Biol Drug Des. 68:97–106.
Mishra NK, Chang J, Zhao PX. 2014. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. PLoS One. 9:e100278.
Mondal S, Bhavna R, Mohan Babu R, Ramakumar S. 2006. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol. 243:252–260.
Olivera BM. 2006. Conus peptides: biodiversity-based discovery and exogenomics. J Biol Chem. 281:31173–31177.
Olivera BM. 1997. E.E. Just Lecture, 1996. Conus venom peptides, receptor and ion channel targets, and drug design: 50 million years of neuropharmacology. Mol Biol Cell. 8:2101–2109.

Olivera BM, Walker C, Cartier GE, Hooper D, Santos AD, Schoenfeld R, Shetty R, Watkins M, Bandyopadhyay P, Hillyard DR. 1999. Speciation of cone snails and interspecific hyperdivergence of their venom peptides. Potential evolutionary significance of introns. Ann N Y Acad Sci. 870:223–237.

Puillandre N, Koua D, Favreau P, Olivera BM, Stocklin R. 2012. Molecular phylogeny, classification and evolution of conopeptides. J Mol Evol. 74:297–309.

Robinson SD, Norton RS. 2014. Conotoxin gene superfamilies. Mar Drugs. 12:6058–6101.

Robinson SD, Safavi-Hemami H, McIntosh LD, Purcell AW, Norton RS, Papenfuss AT. 2014. Diversity of conotoxin gene superfamilies in the venomous snail, Conus victoriae. PLoS One. 9:e87648.

Safavi-Hemami H, Hu H, Gorasia DG, Bandyopadhyay PK, Veith PD, Young ND, Reynolds EC, Yandell M, Olivera BM, Purcell AW. 2014. Combined proteomic and transcriptomic interrogation of the venom gland of Conus geographus uncovers novel components and functional compartmentalization. Mol Cell Proteomics. 13:938–953.

Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. 2010. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38:D161–6.

Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3:265–274.

Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL. 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. PLoS One. 9:e105067.

Schuster-Böckler B., Schultz J., Rahmann S. 2004. HMM Logos for visualization of protein families. *BMC Bioinformatics* ;5;7

Terrat Y, Biass D, Dutertre S, Favreau P, Remm M, Stocklin R, Piquemal D, Ducancel F. 2012. High-resolution picture of a venom gland transcriptome: case study with the marine snail Conus consors. Toxicon. 59:34–46.

Violette A, Biass D, Dutertre S, Koua D, Piquemal D, Pierrat F, Stocklin R, Favreau P. 2012. Large-scale discovery of conopeptides and conoproteins in the injectable venom of a fish-hunting cone snail using a combined proteomic and transcriptomic approach. J Proteomics. 75:5215–5225.

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 41:D70–82.

Wu Y, Wang L, Zhou M, You Y, Zhu X, Qiang Y, Qin M, Luo S, Ren Z, Xu A. 2013. Molecular evolution and diversity of Conus peptide toxins, as revealed by gene structure and intron sequence analyses. PLoS One. 8:e82495.

Yuan DD, Han YH, Wang CG, Chi CW. 2007. From the identification of gene organization of alpha conotoxins to the cloning of novel toxins. Toxicon. 49:1135–1149.

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. 490:49–54.

SUMMARY IN ESTONIAN

Konopeptiidide klassifitseermine ja kindlakstegemine varjatud Markovi mudelite ja positsioonispetsiifiliste skoorimaatriksite abil

Asjade ja elusorganismide liigitamine aitab meil kiirelt teada saada ja edasi anda palju informatsiooni ühe konkreetse isendi kohta. Seetõttu on klassifitseerimine väga oluline ka molekulaarbioloogias, kus uue geeni või valgu liigitamine annab kätte niidiotsad, mille järgi tema funktsioon ja omadused kindlaks teha.

Minu uurimisobjektideks olid konopeptiidid. Konopeptiidid on soojades meredes elavate koonustigude (*Conus sp*.) mürgis leiduvad lühikesed valgud. Koonusteod on kiskjad ja toituvad sõltuvalt liigist kas ussikestest, teistest molluskitest või kaladest. Koonusteod tulistavad oma saaki mürgiga täidetud harpuuniga, mürk muudab saaklooma liikumatuks ja tigu saab ta rahulikult tervelt alla neelata. Mürgi uimastavava ja halvava toime annavadki erinevad konopeptiidid, mis toimivad põhiliselt närvi- ja lihasrakkudes olevatele ioonkanalitele. Ühe teo mürgis on leitud umbes 1000 erinevat peptiidi. Teod kasutavad mürki ka enesekaitseks ja mõned suuremad liigid on ohtlikud isegi inimestele. Konopeptiidid sünteesitakse eellasvalkudena, millel on signaaljärjestus rakkudest välja mürgitorusse transportimiseks ja propeptiid, mis kaitseb tigu ennast mürgi toime eest ning aitab mürgipeptiidi õigesti modifitseerida. Konopeptiidi eellasvalke lõigatakse mitmest erinevast kohast ja lisaks modifitseeritakse sageli osasid aminohappeid vajaliku aktiivsuse saavutamiseks. Selline posttranslatsiooniline muutmine on ka üks vahenditest, millega teod saavutavad erinevate konopeptiidide suure hulga oma mürgis.

Konopeptiidid on signaaljärjestuse sarnasuse alusel liigitatud superperekondadesse. Signaaljärjestused on väga konserveerunud, propeptiidid veidi varieeruvamad, aga mürgipeptiidide puhul on ka ühe perekonna siseselt muutlikkus väga suur.

Teadlased uurivad konopeptiide lootusega leida nende hulgast uusi ravimikandidaate. Konopeptiidid on väga spetsiifilised närvirakkudes leiduvate ioonkanalite modulaatorid ja omavad seetõttu suurt potentsiaali näiteks valuvaigistite või lihastelõdvestajatena.

Antud uurimistöö esimeseks eesmärgiks oli välja töötada meetod, mille abil saaks transkriptoomi, genoomi või peptidoomi järjestuste hulgast välja otsida ja klassifitseerida konopeptiidid. Selleks kasutasime me profiil-HMM'id ja PSSM'id.

pHMM'ide tugevateks külgedeks on kiirus ja skooride määramine tõenäosuste alusel, tänu millele on neid lihtne kasutada suurte andmehulkade puhul.

Et saavutada suurem tundlikkus, tegime me mudelid iga perekonna kõigi kolme funktsionaalse osa jaoks eraldi.

Kuueteistkümnes konopeptiidi superperekonnas kahekümne neljast oli analüüside tegemise hetkel vähem kui 10 kirjeldatud järjestust. Kolme suurema

perekonna abil testisime, kui palju järjestusi on vaja piisavalt tundlike pHMM'ide treenimiseks. Signaaljärjestuste mudelid olid 100% tundlikud juba ainult ühe-kahe järjestusega treenides, propeptiidide mudelite puhul oli vaja 10–20 ja mürgipeptiidide puhul rohkem kui 30 näidisjärjestust, et saavutada maksimaalne tundlikkus. pHMM'ide spetsiifilisus UniprotKB/Swiss-Prot valkude andmebaasi peal testides oli 99.98%, kogu. Rohkem kui pool miljonit järjestust sisaldavast andmebaasist leidsid meie konopeptiidimudelid üles ainult 111 valku, mis ei olnud konopeptiidid.

Osade perekondade klassifitseerimisel oli pHMM'ide tundlikkus madal ja selle parandamiseks võtsime lisaks pHMM'idele kasutusele ka PSSMid. pHMM'ide ja PSSM'ide kombineerimisega saavutasime 91% tundlikkuse eriti varieeruvate mürgipeptiidide klassifitseerimisel, mis on varasematest meetoditest parem tulemus. See on oluline seetõttu, et näiteks peptidoomi sekveneerimisel on olemas ainult lühikesed mürgipeptiidid või propeptiidid eraldi, mitte terve konopeptiidi eellasjärjestus ühes tükis. Sama olukord esineb ka konopeptiidide otsimisel genoomist, sest konopeptiidide geenidel on enamasti iga funktsionaalne osa kodeeritud erinevas eksonis.

Selle töö teiseks eesmärgiks oli otsida ja kirjeldada konopeptiide koonusteo *Conus consors*'i genoomist. Konopeptiidide leidmiseks genoomist otisimise sarnasust konopeptiidide pHMM'idele ja ka varem kirjeldatud konopeptiididele ja *C. consors'*i mürgist leitud peptiididele. Me leidsime *Conus consors*'i genoomist 214 konopeptiidi, millest 187 olid sellised järjestused, mida pole varem kirjeldatud.

Meid huvitas ka konopeptiidide geenistruktuur, mida on varem vähe uuritud. *Conus consors*'i genoom on umbes sama suur kui inimesel ja sisaldab palju kordusjärjestusi ja teistest geenidest enam esines lihtsaid kordusi just konopeptiidigeenide ümbruses. Sellepärast õnnestus meil kokku panna vaid 15 konopeptiidi geeni 13-st erinevast superperekonnast ja kirjeldada nende geenide ekson-intron struktuur. Kuue perekonna geenistruktuur oli varem teadmata.

Oma tööga oleme andnud väikese panuse looduse tohutu mitmekesisuse kirjeldamisse.

ACKNOWLEDGEMENTS

I am very grateful to my supervisor Maido, who welcomed me to the bioinformatics workgroup despite my lack of knowledge in programming and statistics and who assigned me to a project that was both interesting and productive. He was always available for guidance and discussions and at the same time left enough freedom to make my own choices.

My colleagues from the bioinformatics department gave a lot of support and many ideas. I also received much needed help with writing scripts and programs from Lauris and Reidar. The lunchtime discussions with topics from military strategies and weapons, different cultures and world politics to gardening have widened my worldview considerably.

Age, Triinu and Ulvi – in addition to professional help you have been very pleasant roommates in the office and on the road.

I am very glad to have met all the great people involved in the CONCO project. The CONCO meetings were all productive, in wonderful locations and really fun. My special thanks go to Dominique, Reto, Yves, Daniel and Nicolas.

My weekly visits to Tartu have been really enjoyable thanks to my dear friends Reidar, Helena, Annaliisa, Elin, Reedik and Merike who have welcomed me to their homes, cooked me gourmet dinners and breakfasts accompanied with most pleasant company ☺. Good nutrition and friends' support have boosted my productivity.

I could not have completed this work without the patience of my family. For six years I have abandoned them for two days a week. Special thanks go to my parents who have always provided the back-end support of logistics and childcare when needed.

PUBLICATIONS

CURRICULUM VITAE

Name: Silja Laht
Date of birth: 05.02.1978
E-mail: siljal@ut.ee

Current position:
Researcher at University of Tartu, Institute of Molecular and Cell Biology

Education:
University of Tartu, BSc in molecular biology and genetics, 2000
University of Tartu, MSc in molecular biology and genetics 2001

Languages:
Estonian – mother language, English –fluent in speech and writing, Russian – basic level

Career:

2012–...	University of Tartu, Institute of Molecular and Cell Biology; Researcher
2009–2012	Estonian Biocentre; Extraordinary Researcher
2009–2011	Cellin Technologies Ltd; Researcher
2005–2008	Inbio OÜ; Researcher
2002–2005	FIT Biotech Oyj Plc; Researcher

Main research interests:
Modeling of conopeptides with different methods. Analysis of the genome and transcriptome data of cone snail *Conus consors* with main focus on finding and classifying conopeptides.

Publications:
Koua, D; Laht, S; Kaplinski, L; Stöcklin, R; Remm, M; Favreau, P; Lisacek, F (2013). Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies. Biochimica et Biophysica Acta-Proteins and Proteomics, 1834(4), 717–724.

Koua, D.; Brauer, A.; Laht, S.; Kaplinksi, L.; Favreau, P.; Remm, M.; Lisacek, F.; Stöcklin, R. (2012). ConoDictor: a tool for prediction of conopeptide superfamilies. Nucleic Acids Research, 40(W1), W238 - W241.

Laht S, Koua D, Kaplinski L, Lisacek F, Stöcklin R, Remm M (2012). Identification and classification of conopeptides using profile Hidden Markov Models. Biochimica et Biophysica Acta- Proteins and Proteomics, 1824, 488–492.

Laht, Silja; Meerits, Kati; Altroff, Harri; Faust, Helena; Tsaney, Robert; Kogerman, Priit; Järvekülg, Lilian; Paalme, Viiu; Valkna, Andres; Timmusk, Sirje.

(2008). Generation and characterization of a single-chain Fv antibody against G, a hedgehog signaling pathway transcription factor. Hybridoma, 167–174.
Laht, Silja; Karp, Helen; Kotka, Pille; Järviste, Aiki; Alamäe, Tiina. (2002). Cloning and characterization of glucokinase from a methylotrophic yeast Hansenula polymorpha: different effects on glucose repression in H. polymorpha and Saccharomyces cerevisiae. Gene, 296(1–2), 195–203.

Inventions:

Selection system containing non-antibiotic resistance selection marker.; Owner: FIT Biotech OYJ PLC; Authors: Andres Männik, Tanel Tenson, Maarja Adojaan, Urve Toots, Mart Ustav, Silja Laht; Priority number: 20031319; Priority date: 15.09.2003

ELULOOKIRJELDUS

Nimi: Silja Laht
Sünniaeg: 05.02.1978
Aadress: Bioinformaatika õppetool,
Molekulaar- ja Rakubioloogia Instituut, Riia 23B, 51010, Tartu.
Telefon: 737 4046
E-mail: siljal@ut.ee

Töökoht:
Tartu Ülikooli Molekulaar- ja rakubioloogia instituudi teadur

Haridus:
Tartu Ülikool, bakalaurusekraad molekulaarbioloogia ja geneetika erialal, 2000 a.
Tartu Ülikool, magistrikraad molekulaarbioloogia ja geneetika erialal, 2001 a.

Keelteoskus:
Eesti keel – emakeel, inglise keel – vabalt kõnes ja kirjas, vene keel – algtase

Teenistuskäik

2012–	Tartu Ülikool, Molekulaar- ja rakubioloogia instituut; Teadur
2011–	LabToWellness OÜ; Arendusjuht
2009–2012	Eesti Biokeskus; Erakorraline teadur
2009–2011	OÜ Cellin Technologies; Teadur
2005–2008	InBio OÜ; Teadur
2002–2005	FIT Biotech Oyj Plc Eesti filiaal; Teadur

Peamised uurimisvaldkonnad:
Konopeptiidide modelleerimine erinevate mudelitega. Koonusteo *Conus consors* genoomi ja transkriptoomi analüüs ja konopeptiidide otsimine ning iseloomustamine järjestuseandmetest.

Publikatsioonid:
Koua, D; Laht, S; Kaplinski, L; Stöcklin, R; Remm, M; Favreau, P; Lisacek, F (2013). Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies. Biochimica et Biophysica Acta-Proteins and Proteomics, 1834(4), 717–724.

Koua, D.; Brauer, A.; Laht, S.; Kaplinksi, L.; Favreau, P.; Remm, M.; Lisacek, F.; Stöcklin, R. (2012). ConoDictor: a tool for prediction of conopeptide superfamilies. Nucleic Acids Research, 40(W1), W238–W241.

Laht S, Koua D, Kaplinski L, Lisacek F, Stöcklin R, Remm M (2012). Identification and classification of conopeptides using profile Hidden Markov Models. Biochimica et Biophysica Acta- Proteins and Proteomics, 1824, 488–492.

Laht, Silja; Meerits, Kati; Altroff, Harri; Faust, Helena; Tsaney, Robert; Kogerman, Priit; Järvekülg, Lilian; Paalme, Viiu; Valkna, Andres; Timmusk, Sirje. (2008). Generation and characterization of a single-chain Fv antibody against G, a hedgehog signaling pathway transcription factor. Hybridoma, 167–174.

Laht, Silja; Karp, Helen; Kotka, Pille; Järviste, Aiki; Alamäe, Tiina. (2002). Cloning and characterization of glucokinase from a methylotrophic yeast Hansenula polymorpha: different effects on glucose repression in H. polymorpha and Saccharomyces cerevisiae. Gene, 296(1–2), 195–203.

Patentsed leiutised

Selection system containing non- antibiotic resistance selection marker. Omanik: FIT Biotech OYJ PLC; Autorid: Andres Männik, Tanel Tenson, Maarja Adojaan, Urve Toots, Mart Ustav, Silja Laht; Prioriteedinumber: 20031319; Prioriteedikuupäev: 15.09.2003

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida*. Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.
13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand.** The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro*. Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous grassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan.** Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.

41. **Sulev Ingerpuu.** Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.
42. **Veljo Kisand.** Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa.** Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa.** Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik.** Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo.** Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo.** Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots.** Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero.** Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees.** Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks.** Cholecystokinin (CCK) — induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and serotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna.** Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro.** Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane.** Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm.** Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg.** Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild.** The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu.** Studies of the TOL plasmid transcription factor XylS. Tartu 2000. 88 p.

61. **Dina Lepik.** Modulation of viral DNA replication by tumor suppressor protein p53. Tartu 2000. 106 p.
62. **Kai Vellak.** Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu 2000. 122 p.
63. **Jonne Kotta.** Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu 2000. 160 p.
64. **Georg Martin.** Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000. 139 p.
65. **Silvia Sepp.** Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira.** On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000. 96 p.
67. **Priit Zingel.** The role of planktonic ciliates in lake ecosystems. Tartu 2001. 111 p.
68. **Tiit Teder.** Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu 2001. 122 p.
69. **Hannes Kollist.** Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu 2001. 80 p.
70. **Reet Marits.** Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu 2001. 112 p.
71. **Vallo Tilgar.** Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major*, breeding in Nothern temperate forests. Tartu, 2002. 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002. 108 p.
73. **Liina Eek-Piirsoo.** The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002. 74 p.
74. **Krõõt Aasamaa.** Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002. 110 p.
75. **Nele Ingerpuu.** Bryophyte diversity and vascular plants. Tartu, 2002. 112 p.
76. **Neeme Tõnisson.** Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002. 124 p.
77. **Margus Pensa.** Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003. 110 p.
78. **Asko Lõhmus.** Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003. 168 p.
79. **Viljar Jaks.** p53 — a switch in cellular circuit. Tartu, 2003. 160 p.
80. **Jaana Männik.** Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003. 140 p.
81. **Marek Sammul.** Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003. 159 p

82. **Ivar Ilves.** Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003. 89 p.
83. **Andres Männik.** Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003. 109 p.
84. **Ivika Ostonen.** Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003. 158 p.
85. **Gudrun Veldre.** Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003. 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004. 159 p.
87. **Aare Abroi.** The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004. 135 p.
88. **Tiina Kahre.** Cystic fibrosis in Estonia. Tartu, 2004. 116 p.
89. **Helen Orav-Kotta.** Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004. 117 p.
90. **Maarja Öpik.** Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004. 175 p.
91. **Kadri Tali.** Species structure of *Neotinea ustulata*. Tartu, 2004. 109 p.
92. **Kristiina Tambets.** Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylo-geographic approach. Tartu, 2004. 163 p.
93. **Arvi Jõers.** Regulation of p53-dependent transcription. Tartu, 2004. 103 p.
94. **Lilian Kadaja.** Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004. 103 p.
95. **Jaak Truu.** Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004. 128 p.
96. **Maire Peters.** Natural horizontal transfer of the *pheBA* operon. Tartu, 2004. 105 p.
97. **Ülo Maiväli.** Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004. 130 p.
98. **Merit Otsus.** Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004. 103 p.
99. **Mikk Heidemaa.** Systematic studies on sawflies of the genera *Dolerus*, *Empria*, and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004. 167 p.
100. **Ilmar Tõnno.** The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004. 111 p.
101. **Lauri Saks.** Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004. 144 p.
102. **Siiri Rootsi.** Human Y-chromosomal variation in European populations. Tartu, 2004. 142 p.

103. **Eve Vedler.** Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.
104. **Andres Tover.** Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 126 p.
105. **Helen Udras.** Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005. 100 p.
106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005. 162 p.
107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005. 162 p.
108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005. 156 p.
109. **Kaasik, Krista.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005. 121 p.
110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005. 112 p.
111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005. 103 p.
112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005. 112 p.
113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 106 p.
114. **Mait Metspalu.** Through the course of prehistory in india: tracing the mtDNA trail. Tartu, 2005. 138 p.
115. **Elin Lõhmussaar.** The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006. 124 p.
116. **Priit Kupper.** Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006. 126 p.
117. **Heili Ilves.** Stress-induced transposition of Tn*4652* in *Pseudomonas Putida*. Tartu, 2006. 120 p.
118. **Silja Kuusk.** Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006. 126 p.
119. **Kersti Püssa.** Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006. 90 p.
120. **Lea Tummeleht.** Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006. 94 p.
121. **Toomas Esperk.** Larval instar as a key element of insect growth schedules. Tartu, 2006. 186 p.
122. **Harri Valdmann.** Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.

123. **Priit Jõers.** Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli.** Gata3 and Gata2 in inner ear development. Tartu, 2007. 123 p.
125. **Kai Rünk.** Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007. 143 p.
126. **Aveliina Helm.** Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007. 89 p.
127. **Leho Tedersoo.** Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007. 233 p.
128. **Marko Mägi.** The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007. 135 p.
129. **Valeria Lulla.** Replication strategies and applications of Semliki Forest virus. Tartu, 2007. 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007. 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007. 171 p.
132. **Tatjana Krama.** Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007. 112 p.
133. **Signe Saumaa.** The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007. 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007. 96 p.
135. **Priit Kilgas.** Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007. 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007. 95 p.
137. **Kärt Padari.** Protein transduction mechanisms of transportans. Tartu, 2008. 128 p.
138. **Siiri-Lii Sandre.** Selective forces on larval colouration in a moth. Tartu, 2008. 125 p.
139. **Ülle Jõgar.** Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008. 99 p.
140. **Lauri Laanisto.** Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008. 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008. 105 p.
142. **Birgot Paavel.** Bio-optical properties of turbid lakes. Tartu, 2008. 175 p.

143. **Kaire Torn.** Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg.** Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd.** Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag.** Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.
147. **Ulvi Karu.** Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm.** Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks.** Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu.** Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš.** Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko.** Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast.** Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats.** Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova.** The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik.** Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2009, 112 p.
158. **Kaja Kiiver.** Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja.** Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast.** Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.
161. **Ain Vellak.** Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.

162. **Triinu Remmel.** Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.
163. **Jaana Salujõe.** Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.
164. **Ele Vahtmäe.** Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.
165. **Liisa Metsamaa.** Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.
166. **Pille Säälik.** The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.
167. **Lauri Peil.** Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.
168. **Lea Hallik.** Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.
169. **Mariliis Tark.** Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.
170. **Riinu Rannap.** Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.
171. **Maarja Adojaan.** Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.
172. **Signe Altmäe.** Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.
173. **Triin Suvi.** Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.
174. **Velda Lauringson.** Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.
175. **Eero Talts.** Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.
176. **Mari Nelis.** Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.
177. **Kaarel Krjutškov.** Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.
178. **Egle Köster.** Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.
179. **Erki Õunap.** Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.
180. **Merike Jõesaar.** Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.
181. **Kristjan Herkül.** Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.

182. **Arto Pulk.** Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü.** Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla.** Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey.** The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.
186. **Katrin Kepp.** Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber.** The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro.** The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold.** Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert.** Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu.** Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik.** ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber.** Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper.** Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak.** Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo.** Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel.** Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus.** Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius.** Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv.** Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.
201. **Kristjan Välk.** Gene expression profiling and genome-wide association studies of non-small cell lung cancer. Tartu, 2011, 98 p.

202. **Arno Põllumäe.** Spatio-temporal patterns of native and invasive zooplankton species under changing climate and eutrophication conditions. Tartu, 2011, 153 p.
203. **Egle Tammeleht.** Brown bear (*Ursus arctos*) population structure, demographic processes and variations in diet in northern Eurasia. Tartu, 2011, 143 p.
205. **Teele Jairus.** Species composition and host preference among ectomycorrhizal fungi in Australian and African ecosystems. Tartu, 2011, 106 p.
206. **Kessy Abarenkov.** PlutoF – cloud database and computing services supporting biological research. Tartu, 2011, 125 p.
207. **Marina Grigorova.** Fine-scale genetic variation of follicle-stimulating hormone beta-subunit coding gene (*FSHB*) and its association with reproductive health. Tartu, 2011, 184 p.
208. **Anu Tiitsaar.** The effects of predation risk and habitat history on butterfly communities. Tartu, 2011, 97 p.
209. **Elin Sild.** Oxidative defences in immunoecological context: validation and application of assays for nitric oxide production and oxidative burst in a wild passerine. Tartu, 2011, 105 p.
210. **Irja Saar**. The taxonomy and phylogeny of the genera *Cystoderma* and *Cystodermella* (Agaricales, Fungi). Tartu, 2012, 167 p.
211. **Pauli Saag.** Natural variation in plumage bacterial assemblages in two wild breeding passerines. Tartu, 2012, 113 p.
212. **Aleksei Lulla.** Alphaviral nonstructural protease and its polyprotein substrate: arrangements for the perfect marriage. Tartu, 2012, 143 p.
213. **Mari Järve.** Different genetic perspectives on human history in Europe and the Caucasus: the stories told by uniparental and autosomal markers. Tartu, 2012, 119 p.
214. **Ott Scheler**. The application of tmRNA as a marker molecule in bacterial diagnostics using microarray and biosensor technology. Tartu, 2012, 93 p.
215. **Anna Balikova**. Studies on the functions of tumor-associated mucin-like leukosialin (CD43) in human cancer cells. Tartu, 2012, 129 p.
216. **Triinu Kõressaar.** Improvement of PCR primer design for detection of prokaryotic species. Tartu, 2012, 83 p.
217. **Tuul Sepp.** Hematological health state indices of greenfinches: sources of individual variation and responses to immune system manipulation. Tartu, 2012, 117 p.
218. **Rya Ero.** Modifier view of the bacterial ribosome. Tartu, 2012, 146 p.
219. **Mohammad Bahram.** Biogeography of ectomycorrhizal fungi across different spatial scales. Tartu, 2012, 165 p.
220. **Annely Lorents.** Overcoming the plasma membrane barrier: uptake of amphipathic cell-penetrating peptides induces influx of calcium ions and downstream responses. Tartu, 2012, 113 p.

221. **Katrin Männik.** Exploring the genomics of cognitive impairment: whole-genome SNP genotyping experience in Estonian patients and general population. Tartu, 2012, 171 p.
222. **Marko Prous.** Taxonomy and phylogeny of the sawfly genus *Empria* (Hymenoptera, Tenthredinidae). Tartu, 2012, 192 p.
223. **Triinu Visnapuu.** Levansucrases encoded in the genome of *Pseudomonas syringae* pv. tomato DC3000: heterologous expression, biochemical characterization, mutational analysis and spectrum of polymerization products. Tartu, 2012, 160 p.
224. **Nele Tamberg.** Studies on Semliki Forest virus replication and pathogenesis. Tartu, 2012, 109 p.
225. **Tõnu Esko.** Novel applications of SNP array data in the analysis of the genetic structure of Europeans and in genetic association studies. Tartu, 2012, 149 p.
226. **Timo Arula.** Ecology of early life-history stages of herring *Clupea harengus membras* in the northeastern Baltic Sea. Tartu, 2012, 143 p.
227. **Inga Hiiesalu.** Belowground plant diversity and coexistence patterns in grassland ecosystems. Tartu, 2012, 130 p.
228. **Kadri Koorem.** The influence of abiotic and biotic factors on small-scale plant community patterns and regeneration in boreonemoral forest. Tartu, 2012, 114 p.
229. **Liis Andresen.** Regulation of virulence in plant-pathogenic pectobacteria. Tartu, 2012, 122 p.
230. **Kaupo Kohv.** The direct and indirect effects of management on boreal forest structure and field layer vegetation. Tartu, 2012, 124 p.
231. **Mart Jüssi.** Living on an edge: landlocked seals in changing climate. Tartu, 2012, 114 p.
232. **Riina Klais.** Phytoplankton trends in the Baltic Sea. Tartu, 2012, 136 p.
233. **Rauno Veeroja.** Effects of winter weather, population density and timing of reproduction on life-history traits and population dynamics of moose (*Alces alces*) in Estonia. Tartu, 2012, 92 p.
234. **Marju Keis.** Brown bear (*Ursus arctos*) phylogeography in northern Eurasia. Tartu, 2013, 142 p.
235. **Sergei Põlme.** Biogeography and ecology of *alnus*- associated ectomycorrhizal fungi – from regional to global scale. Tartu, 2013, 90 p.
236. **Liis Uusküla.** Placental gene expression in normal and complicated pregnancy. Tartu, 2013, 173 p.
237. **Marko Lõoke.** Studies on DNA replication initiation in *Saccharomyces cerevisiae.* Tartu, 2013, 112 p.
238. **Anne Aan.** Light- and nitrogen-use and biomass allocation along productivity gradients in multilayer plant communities. Tartu, 2013, 127 p.
239. **Heidi Tamm.** Comprehending phylogenetic diversity – case studies in three groups of ascomycetes. Tartu, 2013, 136 p.

240. **Liina Kangur.** High-Pressure Spectroscopy Study of Chromophore-Binding Hydrogen Bonds in Light-Harvesting Complexes of Photosynthetic Bacteria. Tartu, 2013, 150 p.
241. **Margus Leppik.** Substrate specificity of the multisite specific pseudouridine synthase RluD. Tartu, 2013, 111 p.
242. **Lauris Kaplinski.** The application of oligonucleotide hybridization model for PCR and microarray optimization. Tartu, 2013, 103 p.
243. **Merli Pärnoja.** Patterns of macrophyte distribution and productivity in coastal ecosystems: effect of abiotic and biotic forcing. Tartu, 2013, 155 p.
244. **Tõnu Margus.** Distribution and phylogeny of the bacterial translational GTPases and the Mqsr/YgiT regulatory system. Tartu, 2013, 126 p.
245. **Pille Mänd**. Light use capacity and carbon and nitrogen budget of plants: remote assessment and physiological determinants. Tartu, 2013, 128 p.
246. **Mario Plaas.** Animal model of Wolfram Syndrome in mice: behavioural, biochemical and psychopharmacological characterization. Tartu, 2013, 144 p.
247. **Georgi Hudjašov.** Maps of mitochondrial DNA, Y-chromosome and tyrosinase variation in Eurasian and Oceanian populations. Tartu, 2013, 115 p.
248. **Mari Lepik**. Plasticity to light in herbaceous plants and its importance for community structure and diversity. Tartu, 2013, 102 p.
249. **Ede Leppik**. Diversity of lichens in semi-natural habitats of Estonia. Tartu, 2013, 151 p.
250. **Ülle Saks.** Arbuscular mycorrhizal fungal diversity patterns in boreonemoral forest ecosystems. Tartu, 2013, 151 p.
251. **Eneli Oitmaa**. Development of arrayed primer extension microarray assays for molecular diagnostic applications. Tartu, 2013, 147 p.
252. **Jekaterina Jutkina.** The horizontal gene pool for aromatics degradation: bacterial catabolic plasmids of the Baltic Sea aquatic system. Tartu, 2013, 121 p.
253. **Helen Vellau.** Reaction norms for size and age at maturity in insects: rules and exceptions. Tartu, 2014, 132 p.
254. **Randel Kreitsberg.** Using biomarkers in assessment of environmental contamination in fish – new perspectives. Tartu, 2014, 107 p.
255. **Krista Takkis.** Changes in plant species richness and population performance in response to habitat loss and fragmentation.Tartu, 2014, 141 p.
256. **Liina Nagirnaja.** Global and fine-scale genetic determinants of recurrent pregnancy loss. Tartu, 2014, 211 p.
257. **Triin Triisberg.** Factors influencing the re-vegetation of abandoned extracted peatlands in Estonia. Tartu, 2014, 133 p.
258. **Villu Soon.** A phylogenetic revision of the *Chrysis ignita* species group (Hymenoptera: Chrysididae) with emphasis on the northern European fauna. Tartu, 2014, 211 p.

259. **Andrei Nikonov.** RNA-Dependent RNA Polymerase Activity as a Basis for the Detection of Positive-Strand RNA Viruses by Vertebrate Host Cells. Tartu, 2014, 207 p.
260. **Eele Õunapuu-Pikas**. Spatio-temporal variability of leaf hydraulic conductance in woody plants: ecophysiological consequences. Tartu, 2014, 135 p.
261. **Marju Männiste**. Physiological ecology of greenfinches: information content of feathers in relation to immune function and behavior. Tartu, 2014, 121 p.
262. **Katre Kets.** Effects of elevated concentrations of CO_2 and O_3 on leaf photosynthetic parameters in *Populus tremuloides*: diurnal, seasonal and interannual patterns. Tartu, 2014, 115 p.
263. **Külli Lokko.** Seasonal and spatial variability of zoopsammon communities in relation to environmental parameters. Tartu, 2014, 129 p.
264. **Olga Žilina**. Chromosomal microarray analysis as diagnostic tool: Estonian experience. Tartu, 2014, 152 p.
265. **Kertu Lõhmus**. Colonisation ecology of forest-dwelling vascular plants and the conservation value of rural manor parks. Tartu, 2014, 111 p.
266. **Anu Aun.** Mitochondria as integral modulators of cellular signaling. Tartu, 2014, 167 p.
267. **Chandana Basu Mallick.** Genetics of adaptive traits and gender-specific demographic processes in South Asian populations. Tartu, 2014, 160 p.
268. **Riin Tamme.** The relationship between small-scale environmental heterogeneity and plant species diversity. Tartu, 2014, 130 p.
269. **Liina Remm.** Impacts of forest drainage on biodiversity and habitat quality: implications for sustainable management and conservation. Tartu, 2015, 126 p.
270. **Tiina Talve.** Genetic diversity and taxonomy within the genus *Rhinanthus.* Tartu, 2015, 106 p.
271. **Mehis Rohtla**. Otolith sclerochronological studies on migrations, spawning habitat preferences and age of freshwater fishes inhabiting the Baltic Sea. Tartu, 2015, 137 p.
272. **Alexey Reshchikov.** The world fauna of the genus *Lathrolestes* (Hymenoptera, Ichneumonidae). Tartu, 2015, 247 p.
273. **Martin Pook.** Studies on artificial and extracellular matrix protein-rich surfaces as regulators of cell growth and differentiation. Tartu, 2015, 142 p.
274. **Mai Kukumägi.** Factors affecting soil respiration and its components in silver birch and Norway spruce stands. Tartu, 2015, 155 p.
275. **Helen Karu.** Development of ecosystems under human activity in the North-East Estonian industrial region: forests on post-mining sites and bogs. Tartu, 2015, 152 p.
276. **Hedi Peterson.** Exploiting high-throughput data for establishing relationships between genes. Tartu, 2015, 186 p.

277. **Priit Adler.** Analysis and visualisation of large scale microarray data, Tartu, 2015, 126 p.
278. **Aigar Niglas.** Effects of environmental factors on gas exchange in deciduous trees: focus on photosynthetic water-use efficiency. Tartu, 2015, 152 p.