Bioactive Macrocyclic Cysteine-rich Peptides: Insights on Sequence, Structure and Application of Cyclotides

A Thesis Submitted to

The University of Trans-Disciplinary Health Sciences and Technology, Bengaluru

for the Award of the Degree of

Doctor of Philosophy

by

Neha Vijay Kalmankar

Under the guidance of

Dr. Radhika Venkatesan

National Centre for Biological Sciences (NCBS), Bengaluru

Tata Institute of Fundamental Research (TIFR), Mumbai

December 2021

Certificate

This is to certify that the work incorporated in this thesis "**Bioactive Macrocyclic Cysteine-Rich Peptides: Insights on Sequence, Structure and Application of Cyclotides**" submitted by **Neha Vijay Kalmankar** was carried out under my supervision during the period of 2016 - 2021 for the degree of Doctor of Philosophy at The University of Trans-disciplinary Health Sciences and Technology, Bengaluru. No part of this thesis has been submitted for a degree or examination at any other university. I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation. In my capacity as supervisor of the candidate's thesis, I certify that the above statements are true to the best of my knowledge.

Research Supervisor

Date: 06-12-2021

Dr. Radhika Venkatesan
Assistant Professor
National Center for Biological Sciences (NCBS),
Tata Institute of Fundamental Research (TIFR),
GKVK Bellary Road, Bangalore - 560065

Research Co-Supervisor

Date: 06-12-2021

Prof. R. Sowdhamini
Professor
National Center for Biological Sciences (NCBS),
Tata Institute of Fundamental Research (TIFR),
GKVK Bellary Road, Bangalore - 560065

Abstract

Cyclotides are a unique class of gene-encoded, ribosomally synthesized, macrocyclic peptides (26-37 residues) produced in several plant species. They form a cyclic cystine knot (CCK) motif comprising of six disulfide bonds and circular backbone, and can be divided into two main subfamilies i.e. Möbius and Bracelet, with Möbius cyclotides containing a *cis*-proline residue in the loop 5 region.

The transcriptome assembly of *Clitoria ternatea* was performed to facilitate gene mining of cyclotide precursors arising in four tissues (pod, stem, leaf and flower). This resulted in the identification of 71 precursor genes, of which 26 are novel cyclotide sequences. Differential expression analysis revealed that numerous cyclotide genes were differentially expressed across the tissues. Moreover, we have also proposed a model of cyclotide biosynthesis, wherein we followed the presence and expression of several oxidative folding and processing enzymes. Additionally, an assortment of cyclotides was detected across five tissues using proteomics method. Notable variations in LC-MS/MS product ion distributions were observed in cyclotides belonging to the two structural subfamilies based on the number and positions of prolines. Distinct MS/MS patterns determined by Xxx-Pro bond fragmentation of prototypical cyclotides was used as a diagnostic to rapidly sequence two novel cyclotides, ctr pep 30 and ctr pep 43.

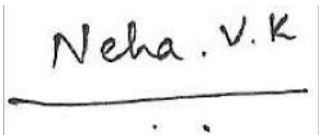
Furthermore, our in-house disulfide bond database - DSDBASE2.0 was updated to include the latest PDB entries (January 2021 release) and multiple tools were incorporated for structure prediction of disulfide-rich peptides (http://caps.ncbs.res.in/dsdbase2). It is now possible to also obtain three-dimensional models of disulfide-rich proteins using an independent algorithm - RANMOD. Finally, we cyclotides' ability as an inhibitor of β-amyloid (Aβ) fibrils was investigated and demonstrated using an animal model of Alzheimer's disease. We describe how these peptides protect against β-amyloid toxicity and oxidative stress using different transgenic *Caenorhabditis elegans* strains.

Collectively, the findings in this thesis provide insights into the diversity and characterization of cyclotides from *Clitoria ternatea* and highlight their future applications as therapeutic lead molecules, especially as an inhibitor of Aβ fibrils.

Declaration

I declare that this thesis entitled "**Bioactive Macrocyclic Cysteine-Rich Peptides: Insights on Sequence, Structure and Application of Cyclotides**" submitted for the award of Doctor of Philosophy degree to The University of Trans-disciplinary Health Sciences and Technology, Bengaluru, is my original work, conducted under the supervision of my guide **Dr. Radhika Venkatesan**, and co-supervisor **Prof. R. Sowdhamini**, at the National Centre for Biological Sciences, Bengaluru. I also wish to inform that no part of the research has been submitted for a degree or examination at any other university. References, help and material obtained from other sources have been duly acknowledged.

I hereby confirm the originality of the work and that there is no plagiarism in any part of the dissertation.



Place: Bengaluru **Signature of the Candidate**
Date: 06-12-2021

Neha Vijay Kalmankar
Reg. No.: 21215030013

Acknowledgement

This thesis marks the end of my PhD journey, which would not have been possible without the professional and personal support from those around me. My colleagues, friends and family have been central in shaping my experience, and I wish to convey my sincere gratitude to each one of them. My heartfelt apologies to those whom I might have forgotten to mention by name.

First of all, I would like to thank my (unofficial) thesis advisor and a venerable scienist, **Prof. P. Balaram** (former Director, Indian Institute of Science, Bangalore, India). I am deeply indebted to him for sparking a great interest in me to pursue research career and for co-opting an engineer into a group of chemists and biologists. I had the privilege of working with him as a junior and senior research fellow in his lab from 2010 to 2015. He encouraged me to register for a PhD program in 2016, and has not stopped steering me through my research ever since. Thank you for believing in me, especially when I lacked it myself. If one day I can manage to become half the scientist and a humble person as him, I will consider myself fortunate.

I would like to thank my primary supervisor, **Dr. Radhika Venkatesan** (NCBS) for giving me the opportunity of working in her lab on chemical ecology. I am grateful for her guidance, our numerous scientific discussions, and for pushing me beyond my comfort and limits to succeed. I owe my deepest gratitude to **Prof. R. Sowdhamini** (NCBS) for being my co-supervisor, always being there for me, and being a wonderful mentor. Thank you for giving me the opportunity to work in your lab and learn bioinformatics from the best. My thesis work would not have been possible without her unwavering kindness, patience and faith, that she shows towards every student who works in her lab. I am glad to have been a joint PhD student of both of them, as this allowed me to hone in diverse sets of skills and expertise over the years.

To all the past and present colleagues at NCBS and IISc, who I am lucky enough to call my friends, thank you all for making this experience an enjoyable one. I thank all the members of the **Interactions Lab** and **CAPS lab** for their camaraderie, both inside and outside the laboratories. In the time I spent in Interactions lab (Dr. Radhika's lab), I have some wonderful memories of all the lab members, namely Rohit, Enakshi, Shiksha, Saikat, Vineet, Ashwathi, Radhika, Anupam, Karan, Rohini, Praveena and Kokila. I thank each one of you for being part of my journey all these years. I would like to thank Saikat for helping me grow butterfly pea plants and with the cyclotide extraction procedures. Special note of thanks to Kruthika, Hrudya, Sanjana, and Priya, the trainees I worked with. Thank

you for your assistance with many of the experiements and our exciting scientific discussions.

I have learnt a lot from every member of the CAPS lab (Prof. Sowdhamini's lab) that includes Teerna, Shafi, Sajeevan, Naseer, Pritha, Nitish, Bhavika, Anamya, Yughandar, Surbhi, Oommen, Gandhimathi, Binnu, Anshul, Snehal, Mahita, Harini, Kaushik, Rithvik, Atul, Meenakshi, Rajas, Adwait, Souradeep, Seshank, Nagarathnam, Sheetal, Vikas, Shailya, Pankaj, Abhishek, and Soumya. A big thanks to each one of you for helping me with several computational aspects of my thesis work. I would like to specially acknowledge Shafi and Teerna from the lab, for their kind friendship, all the non-academic fun and frolic, and all our chat sessions over several cups of tea. I have several fond memories from our numerous gatherings and outings taken as a big group. Special thanks to Murugavel for all his help with DSDBASE/RANMOD related work and all the conversations we shared on both scientific and non-academic fronts. I am grateful to Sajeevan for helping me with the RNA extraction work and our discussions on genomics and transcriptomics related topics. CAPS lab has also been a host to several trainees over the years and I would like to acknowledge all of them for making the lab so energetic and cheerful. Special thanks to Anisha, a trainee who worked on her 6-month dissertation project with me in the lab. The experience I gained by mentoring several Bachelors and Masters students for their dissertations in both the Interactions lab and CAPS lab over the years has helped me learn a lot in the process. So, a big thanks to each one of you for making me a better researcher today.

I wish to thank Mr. Ravi Kumar for all the assistance with administrative work at The University of Trans-Disciplinary Health Sciences & Technology (TDU) and Ms. Vishalakshi for academics related assistance at NCBS. I thank all the collegues and staff members at the NCBS/IISc, like the Kitchen staff, Instrumentation, IT department, Lab support, Stores, Administration department and canteen facilities for their timely assistance with any related issues. Special thanks to Divya, Chakrapani, Selva, Prashanth, Allwyn, Ranjith, and many more who have professionally maintained all the instruments, greenhouse, and other facilities on the campus. This thesis would not be complete without a mention of the Proteomics Facilities at both NCBS and IISc. I sincerely appreciate all the help I received from Sunitha (MBU dept., IISc) and Chhaya (NCBS) on the mass spectrometry related experiments. I would also like to thank Dr. Ashwini Godbole (TDU) for providing me with the *C. elegans* strains and her suggestions with the *in vivo* experiments.

Even though my PhD journey began in early 2016, my research career started way back in 2010. I want to take this platform to acknowledge several colleagues and friends of mine from my days in IISc. I express my gratitude to all the **PB lab** (Prof. P. Balaram's

lab) members, such as Debarati, Krishnayan, Shashanka, Chiddanand, Jissa, Vasantha, Vidhi, Mousumi, Moumita, Vijaysarathy, Soorej, Mukesh, Kallol, Vaigundan and other past and present members of the lab. I am grateful for all the enjoyable days I have spent with them in IISc.

A special note of thanks to two extraordinary mentors whom I miss and remember very fondly, **late Prof. K. S. Krishnan** (NCBS) and **late Prof. C. Ramakrishnan** (IISc). I am thankful to them for triggering a great curiosity for science and mentoring me in the early years of my career. I got lucky when I was hired by Prof. Balaram's first PhD student (fondly called as KSK), at the same time graduating as Prof. Balaram's last PhD student. Prof. Krishnan's greatest gift to me was to remind me to always be amazed by the wonders of nature and life around us. I am fortunate to have also worked with Prof. C. Ramakrishnan in IISc on computational biology projects. Everything I have learnt on the principles of the Ramachandran map and protein structures is because of him. His unflinching unassuming child-like enthusiasm of doing research is a quality that I would like to possess some day. I want to also thank **Prof. N. V. Joshi** (CES department, IISc) for his outstanding teaching abilities in mathematical ecology and statistics. He has been instrumental in guiding me through statistical analyses and data interpretation in several projects we collaborated on in IISc.

To all my closest friends outside of my workplace - Shweta, Neela, Swathi, Ananya, Shalini, Pradeep, Wisvesh, Devak, Ammara, Rithika and Sumeet, I am immensely grateful for their unwavering love and encouragement. They have patiently listened to me rant about my PhD blues, failed experiments, challenging days, and offered comforting advices that helped me get through it all.

Last but not least, I would like to take this opportunity to recognize the people in my personal life, without whom none of this would have been possible. My family has been my biggest support system throughout my life. I wish to thank my parents, **Dr. Vijay** and **Suhasini Kalmankar**, for giving me the freedom and rectitude to follow my dreams, and I dedicate this thesis to them. I am indebted to their numerous sacrifices, unconditional love and support, and key values they have instilled in me. I am also grateful to my younger brothers, **Nikhil** and **Neeraj**, for patiently enduring me all these years and having my back at all times. I am also thankful to my parents-in-law, **Beena** and **M. I. Joseph**, for encouraging me the past two years and helping me focus on my career. To my husband, **Vinu Joseph**, I am forever grateful for his trust, affection and unflinching companionship. Thank you for sharing your life with me. I can undeniably say that I would not be where I am right now without your love and support. Thank you!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Aβ – β-amyloid
ACN – Acetonitrile
AEP – Asparaginyl endopeptidase
AD – Alzheimer's disease
CCK – Cyclic cystine-knot
CID – Collision induced dissociation
CtAEP – *Clitoria ternatea* asparaginyl endopeptidase
CtPDI – *Clitoria ternatea* protein disulfide isomerase
CTPP – C-terminal propeptide
DCM – Dichloromethane
DNA – Deoxy ribonucleic acid
DTT – Dithiothreitol
Endo Glu-C – Endoproteinase Glu-C
ESI-MS – Electro-spray ionization mass spectrometry
GO – Gene Ontologies
FA – Formic acid
HCD – Higher-energy C-trap dissociation
IAM – Iodoacetamide
kB1 – Kalata B1
LC – Liquid chromatography
MALDI-TOF – matrix assisted laser desorption/ionization - time of flight
MCoTI-I – Momordica cochinchinensis trypsin inhibitor-I
MCoTI-II – Momordica cochinchinensis trypsin inhibitor-II
M_{calc} – Calculated mass
MD – Molecular dynamics simulation
Met-Ox – Methionine-oxidised
MeOH – Methanol
M_{expt} – Experimental mass
MODIP – Modeling of disulfide bridges in proteins
MS – Mass spectrometry
MS/MS – tandem mass spectrometry
NTPP – N-terminal propeptide
NTR – N-terminal repeat
OaAEP – *Oldenlandia affinis* asparaginyl endopeptidase

Oak1 – *Oldenlandia affinis* kalata B1 precursor

PA1b – pea albumin 1b chain

PDB – Protein Data Bank

PAL – Peptidyl asparaginyl ligase

PTM – Post-translational modifications

PxAEP – *Petunia hybrida* asparaginyl endopeptidase

RANMOD – Random conformation to polypeptide backbone

RMSD – Root mean square deviation

RMSF – Root mean square fluctuations

Rg – Radius of gyration

RNA – Ribonucleic acid

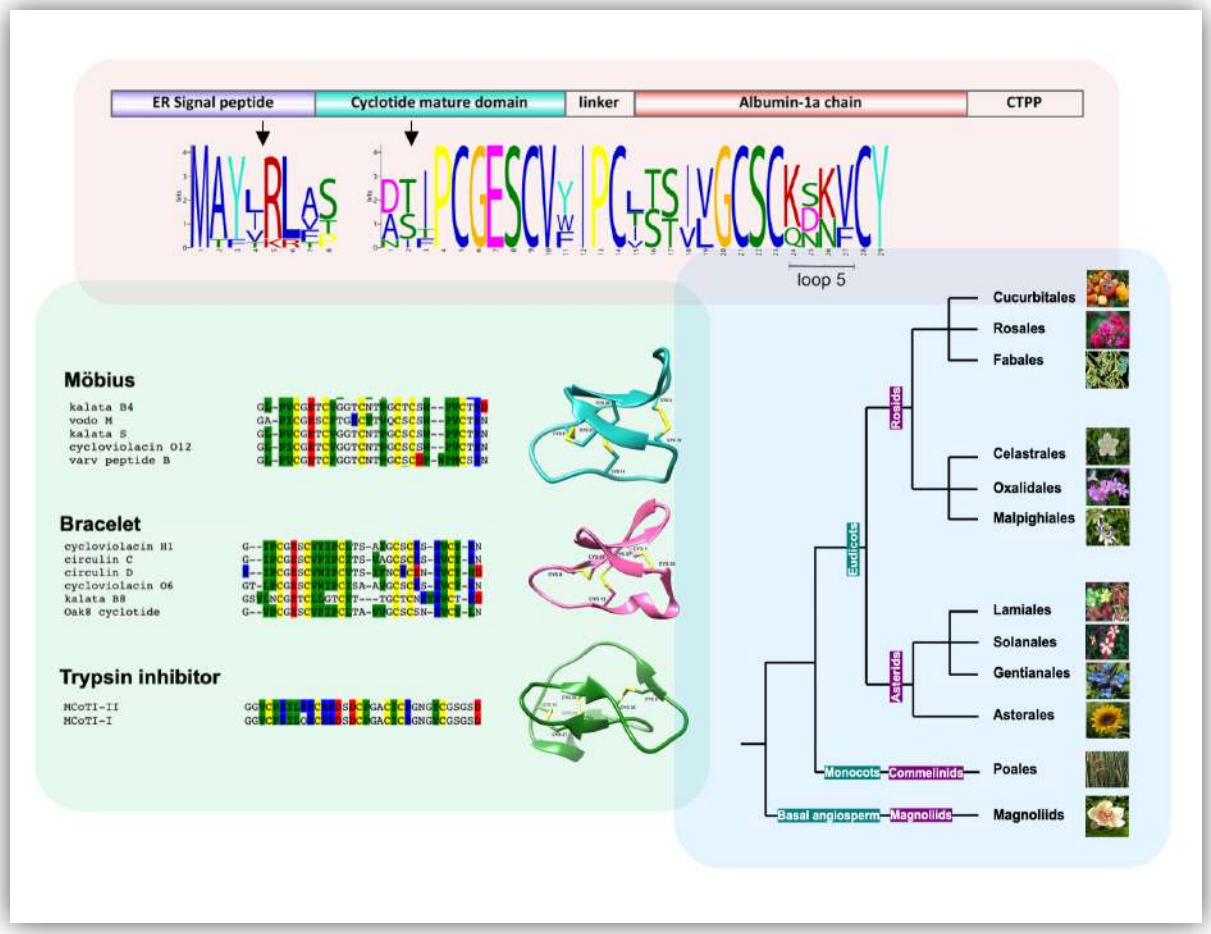
RP-HPLC – Reverse phase high-performance liquid chromatography

RT - Retention time

SFTI – Sunflower trypsin inhibitor

TFA – Trifluoroacetic acid

TI – Trypsin inhibitor


ER Signal peptide
Cyclotide mature domain
linker
Albumin-1a chain
CTPP
loop 5
Möbius
kalata B4
vodo M
kalata S
cycloviolacin O12
varv peptide B
Bracelet
cycloviolacin H1
circulin C
circulin D
cycloviolacin O6
kalata B8
Oak8 cyclotide
Trypsin inhibitor
MCoTI-II
MCoTI-I
Cucurbitales
Rosales
Fabales
Celastrales
Oxalidales
Malpighiales
Lamiales
Solanales
Gentianales
Asterales
Poales
Magnoliids
Rosids
Eudicots
Asterids
Monocots
Commelinids
Basal angiosperm
Magnoliids

1

Introduction

1.1 GENERAL INTRODUCTION

1.1.1 New class of macrocyclic peptides

Cyclotides are one of the largest but mostly underexplored family of circular plant peptides. They are a unique class of macrocyclic peptides (26-37 residues) containing six conserved cysteine residues forming three disulfide bonds, produced by the cyclization of a gene encoded, linear precursor protein (De Veer et al., 2019). The three conserved disulfide bonded framework (Cys I-IV, II-V, III-VI) and head-to-tail cyclic backbone creates a cyclic cystine-knot (CCK) motif, as illustrated in **Figure 1.1** for the classical cyclotide cycloviolacin O2 (Göransson et al., 2009). This CCK motif renders cyclotides with exceptionally stability against enzymatic, chemical and thermal degradation, with their principal function hypothesized to be in plant defense (Colgrave and Craik, 2004). This was established subsequently as they exhibited insecticidal, anthelmintic, antifouling and molluscicidal activities and could interact with and disrupt biological membranes (Gerlach and Mondal, 2012). The signature CCK motif provides them stability that can harness a wide spectrum of bioactivities of medical interest such as cytotoxic, anthelmintic, neurotensin inhibitory, anti-cancerous, anti-HIV and antimicrobial effects (Burman et al.,

2011; Colgrave et al., 2008; Esmaeili et al., 2016; Nguyen et al., 2016; Wang et al., 2008a). Thus, cyclotides have become potential candidates in the development of peptide-based drugs; either as scaffolds to stabilize susceptible peptide sequences or as drugs by their own right.

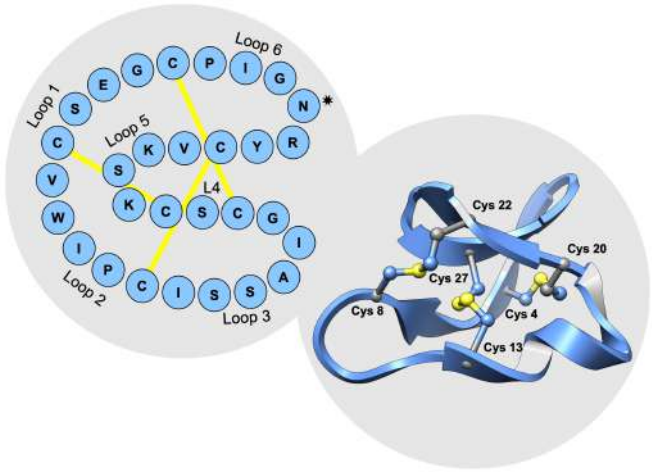


Figure 1.1: Schematic representation of the cyclotide - cycloviolacin O2. (Left) The primary sequence of cycloviolacin O2 (PDB ID 2KNM) is shown using one-letter amino acid codes. The ligation site is located at the Asn of loop 6 and indicated with a star (*). (Right) 3D structure of cycloviolacin O2 shows the cystine-knot motif with the cysteine residues labeled.

1.1.2 History of cyclotides

One of the first ethnopharmacological observations of cyclotides dates back to the 1960s, when Lorents Gran, a Norwegian doctor while serving on a Red Cross relief mission in the Congo in the 1960s, noted that *Oldenlandia affinis* leaves were consumed as herbal tea decoction by Congolese women in Africa during labor resulting in accelerated child birth (Gran, 1973a). The decoction was made by boiling the dry aerial parts of *O. affinis* or "kalata-kalata" (Tsjiluba language). Dr. Gran brought back some of these plants to the University of Oslo and along with his colleagues partially characterized the active agent from this plant as a peptide containing approximately 30 amino acids. This led to the discovery of the prototypic member, kalata B1, as a powerful uterotonic agent (Gran, 1973b). However, the complete primary structure of the peptide was determined only 25 years later using a combination of mass-spectrometry and nuclear magnetic resonance (NMR) methods (Saether et al., 1995).

1.2 DISTRIBUTION IN PLANTS

In the 1990s, several groups independently discovered macrocyclic peptides from

Rubiaceae and Violaceae plant families. From *Viola arvensis* (Violaceae), violapeptide I was discovered that displayed hemolytic activity (Schopke et al., 1993). Circulins A and B was identified from *Chassalia parvifolia* (Rubiaceae) showing anti-HIV properties (Gustafson et al., 1994).

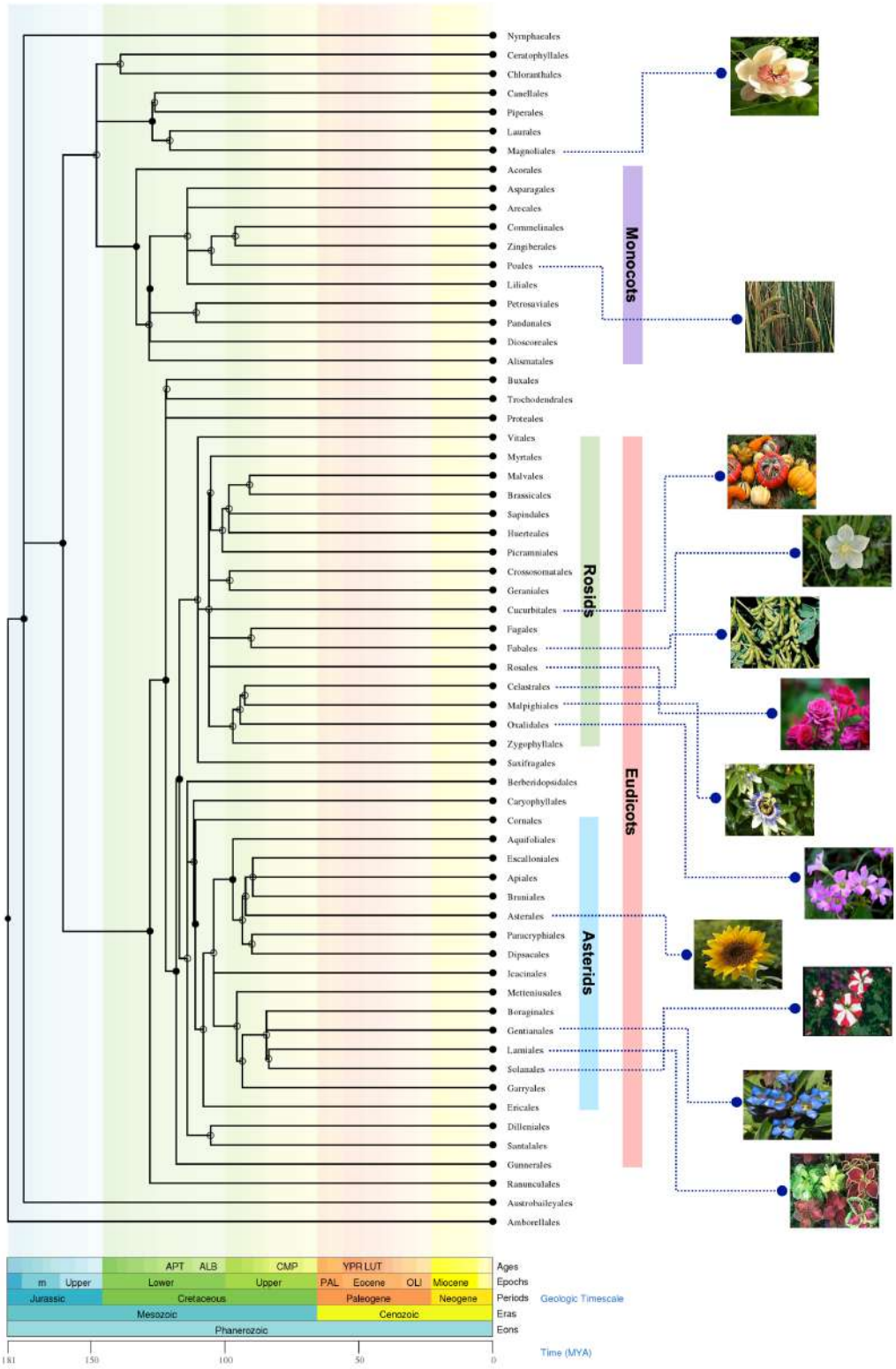


Figure 1.2: Distribution of cyclotides and 'cyclotide-like' sequences in Angiosperms.

The occurrences of cyclotides in angiosperms differ between the basal angiosperms, monocots and eudicots. **Figure 1.2** summarizes the distribution of known cyclotides and the newly discovered 'cyclotide-like' sequences in the plant kingdom. 'Acyclotides' i.e. sequences that lack the cyclic backbone are prevalent in monocots and eudicots. However, true cyclotides have been found only in eudicots so far (Mulvenna et al., 2006; Zhang et al., 2015). No cyclotide-like sequences have yet been found in the basal angiosperms. To date, >600 cyclotides have been characterized from eudicot families: Rubiaceae (coffee), Violaceae (violet), Fabaceae (legume), Solanaceae (potato) and Cucurbitaceae (cucurbit) and acyclotides have been discovered in the monocot family of Poaceae. It is estimated that there could be approximately over 160 different cyclotides produced in a single plant (Gruber et al., 2008). This high diversity and abundance might be indicative of their role in plant defense and given their insecticidal property, they are highly desirable candidates in agriculture. However, most of the interest in cyclotide research is focused on the pharmaceutical applications due to their exceptional stability and unique structural framework.

1.3 CYCLOTIDE BIOSYNTHESIS

Cyclotides are gene-encoded and ribosomally synthesized cyclic proteins that undergo post-translational modification in the plant cell. The cyclotide gene architecture from five different plant families is illustrated in **Figure 1.3**. Typically, the precursor genes start with endoplasmic reticulum (ER) signal peptide, followed by an N-terminal propeptide (NTPP) region, an N-terminal repeat (NTR) region, the cyclotide mature domain, ending with C-terminal propeptide (CTPP). In some cases, more than one NTR, mature cyclotide domain and CTR repeat regions are found. The ER signal peptide indicates that the precursor follows the secretory pathway and presumably where the disulfide bridge formation and protein folding occur. After being processed in the ER, the mature cyclotides are excised and cyclized by a specialized enzyme called asparaginyl endopeptidase (AEP) (Gillon et al., 2008; Saska et al., 2007). **Figure 1.4** shows a schematic illustration of the mechanism of macrocyclization reaction catalysed by plant AEPs. However, before the cyclization step in cyclotide maturation, the leader peptide must be cleaved from mature peptide, and the enzyme (s) involved in this step remain poorly understood.

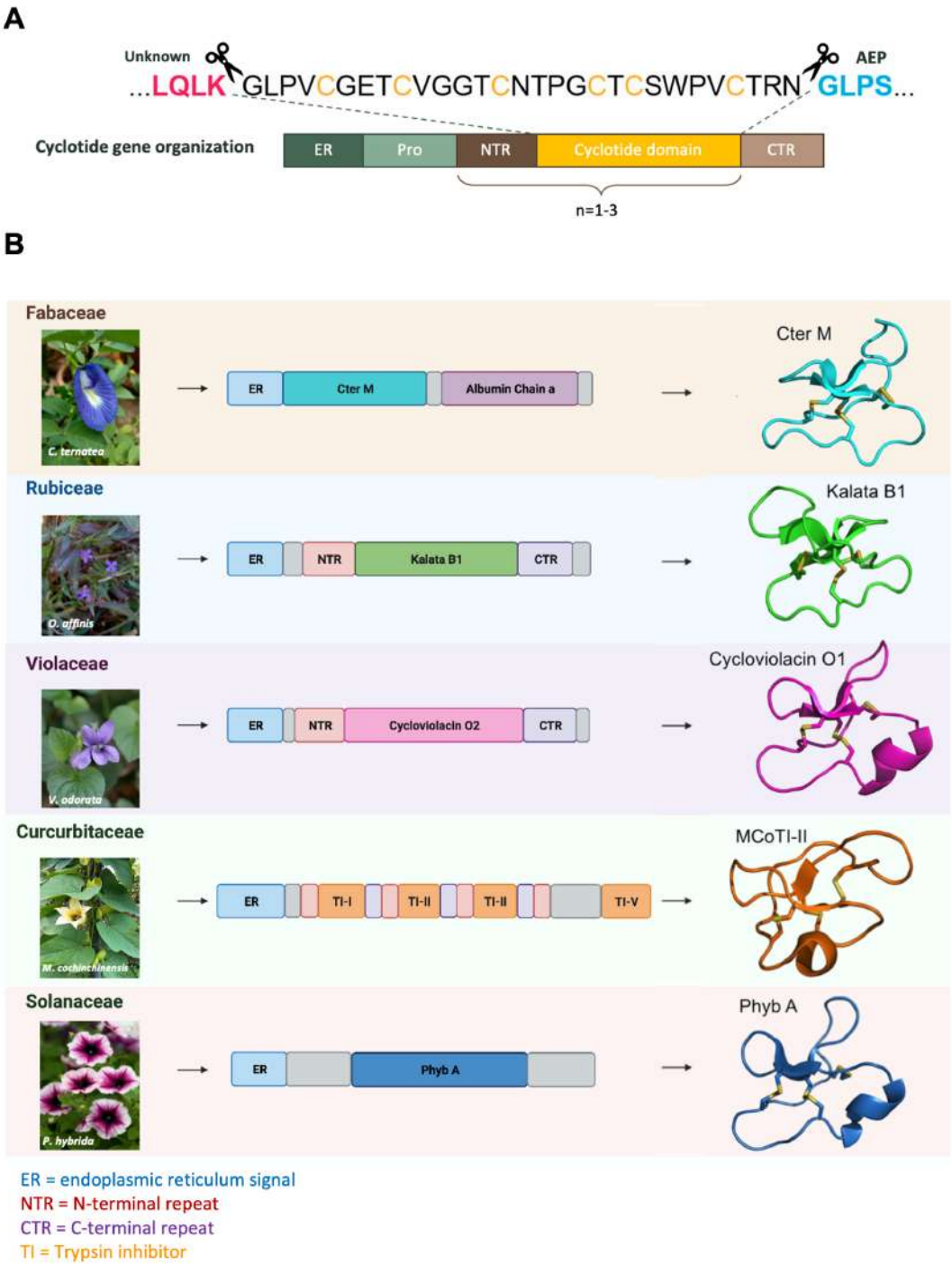


Figure 1.3: Genetic architecture of cyclotide genes from different plant families. (A) Gene architecture of cyclotides in general (B) Violaceae and Rubiaceae precursors contain: ER signal peptide (blue), NTPP (grey), NTR (red), mature cyclotide domain, CTR (purple) region and CTPP (grey). In contrast, in Fabaceae precursor genes, the ER signal peptide is immediately followed by the cyclotide domain, a short peptide linker and the albumin 1a-chain. Cucurbitaceae precursors comprise of an ER signal peptide followed by NTPP region, the NTR region and the cyclotide domain. This cluster is repeated a couple of times with multiple cyclotide domains. In Solanaceae, cyclotide precursors contain an ER sequence, NTPP region followed by a mature cyclotide domain and a short CTPP region. This plant family precursors generally lack the N- and C- terminal repeat regions.

1.3.1 Asparaginyl endopeptidases

Asparaginyl endopeptidases (AEPs, also known as vacuolar processing enzymes or legumains) are primarily known to cleave peptide bonds through hydrolysis mechanism. The first allusion of an unconventional ligation property for AEPs came in 2001 where it was reported to act on the conserved Asx residue in the P1 position in loop 6 region by transpeptidation (Jennings et al., 2001). Since then, several independent studies have confirmed this unique ligating property of AEPs. A study in 2007 showed that transient expression of Oak1 (kalata B1 precursor) in non-native cyclotide producing tobacco plant *Nicotiana benthamiana* led to the production of cylic kB1 and consequently when AEP was inhibited, the production of cyclic kB1 decreased (Saska et al., 2007). Another study in 2008 showed that mutation of the highly conserved Asn to Ala in Oak1, cyclic kB1 could not be detected in transformed *N. tabacum* or *Arabidopsis thaliana* (Gillon et al., 2008). However, the first direct evidence for this came in 2014 by Tam and group who set to identify the specific peptide ligase responsible for circularisation of cyclotides in *C. ternatea* (Nguyen et al., 2014). This enzyme was named butelase 1 and belongs to the legumain family of cysteine proteases. Several AEPs have been characterized so far since the first discovery of butelase-1, such as *Oldenlandia affinis* OaAEP1b (Harris et al., 2015), *Viola yedoensis* VyPALs (Haywood et al., 2018), *Petunia x hybrida* PxAEP3b (Jackson et al., 2018), and *Helianthus annuus* HaAEP1 (Hemu et al., 2019). However, it should be noted that ligase type AEPs are not as frequently occurring as the the native peptidase type.

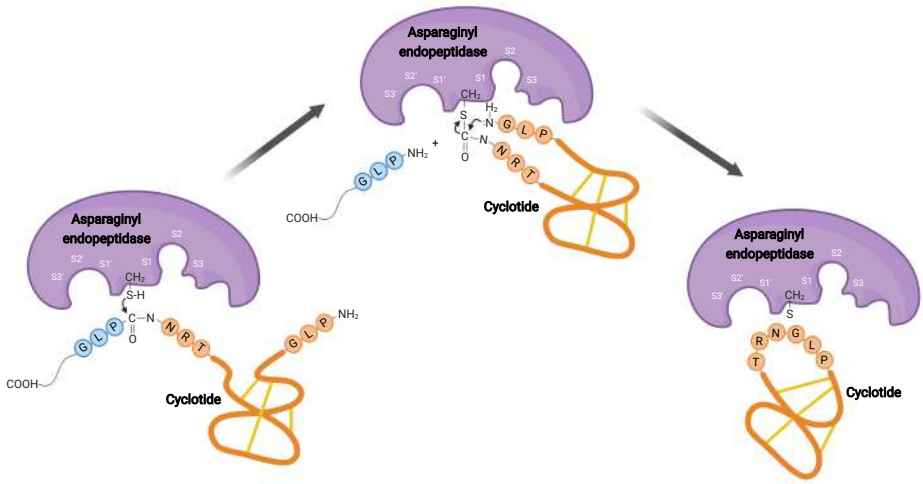


Figure 1.4: Scheme representing the proposed mechanism of cyclotide formation by asparaginyl endopeptidase. The predicted mechanism of macrocyclization involves a cleavage-coupled transpeptidation reaction involving an acyl-transfer step from the acyl-AEP intermediate to the N-terminal residue of the cyclotide domain. AEP recognises and cleaves a peptide substrate containing Asn to form an acyl-enzyme intermediate via a tetrahedral intermediate. The CTPP region and mature cyclotide are highlighted in blue orange respectively.

It is known that the cyclization of cyclotides happens via a transpeptidation reaction. It is believed that this transpeptidation occurs simultaneously with the cleavage of the CTPP peptide from the precursor protein. The transpeptidation reaction takes place by an acyl-transfer from the acyl-AEP intermediate to the N-terminal residue of the cyclotide domain (**Figure 1.4**). For such precise cleavage and ligase mechanisms to co-occur, it neccesitates the presence of key recognition sites. The precursor protein containing the mature cyclotide domain is recognized by the enzyme asparaginyl endopeptidase (AEP) at C-terminal processing site (TRNGLP). The enzyme cleaves on the carboxyl side of the Asn residue, releasing the CTPP and forming an acyl-enzyme intermediate with the cyclotide domain. N-terminal tripeptide motif residues in the cyclotide domain (Gly-Leu-Pro) swing around to occupy the newly vacated S1', S2' and S3' binding subsites. Nucleophilic attack by the amine group from the P1' Gly residue forms the peptide bind, and hence the cyclic peptide backbone. The final cyclic peptide is then released from the enzyme active-site. Several studies have noted the importance of the residues following the conserved Asn in loop 6 for positive cyclization of cyclotides. For example, truncation of the GLP tripeptide motif in the Oak1 precursor, cyclic products of kB1 could not be detected (Conlan et al., 2012; Gillon et al., 2008). Further truncation to GL significantly reduced cyclic kB1 production (Conlan et al., 2012). Significant increase in the production of cyclic kB1 was observed when *O. affinis* AEP1b (OaAEP1b) was transiently co-expressed in *N. benthamiana,* with the native Oak1 precursor containing the GLP motif following the Asn residue (Poon et al., 2018). Furthermore, most of the cyclotide precursors in *C. ternatea* contain a conserved HV motif following the Asn residue. Transient co-expression of butelase-1 and Oak1-HV led to a significantly increased production of cyclic kB1 (Poon et al., 2018). A previous study had already demonstrated the importance of HV motif as a recognition site for AEP in *C. ternatea* by treating Oak1-HV precursor with *C. ternatea* extract and detecting the production of cyclic kB1 (Nguyen et al., 2014). The same study showed a number of synthesized precursors containing the HV motif could be cyclized by butelase-1.

1.3.2 Protein disulfide isomerase

After synthesis, the linear cyclotide precursor protein undergoes posttranslational processing steps, including the splicing and folding of the mature cyclotide domain. Precise disulfide connectivity is essential for native cyclotide structure and function in the cell i.e. oxidation of the six cysteine residues forming the disulfide knot must occur in a specific order. Incorrect bond formation will result in the precursor not being cyclized or detected as a misfolded protein needing destruction in the cell. It is believed that the process of oxidation of Cys residues and formation of disulfide bridges is facilitated by a protein

disulfide isomerase, known to be involved in the folding of several disulfide-rich proteins. PDIs are generally localized to the ER lumen, and have dual functionality; to form the initial disulphide bridges and to reshuffle the bonds into the final, functional form. This hypothesis is supported by *in vitro* experiments where the oxidative folding of cyclotide kalata B1 from *Oldenlandia affinis* (Rubiaceae) was achieved using OaPDI (Gruber et al., 2007). However, not a single *in vivo* evidence of the interactions of PDI family members and cyclotide precursors is currently known.

1.4 SEQUENCE AND STRUCTURE OF CYCLOTIDES

Cyclotides were initially divided into two main subfamilies i.e. Möbius and Bracelet, based on the presence or absence of a *cis*-proline residue in loop 5 region. This cis-Pro residue is believed to creates a twist in the peptide backbone in Möbius cyclotides structures (Rosengren et al., 2003). In addition to this theoretical twist, both the subfamilies can be distinguished by the differences in loop size and sequence variation (Burman et al., 2015) (**Figure 1.5**). Cyclotide loop sizes in general vary among different sequences. However, it is observed that loop 1 usually contains 3-6 amino acids, loop 2 contains 4-8 amino acids, loop 3 is composed of 3-7 amino acids, loop 4 comprises a single residue, loop 5 is 4-5 residues, and loop 6 is 5-8 residues long.

Several positions in the loop regions are conserved and thus are likely important for cyclotide structure and function. For instance, loops 1 and 4 form the core of the cystine knot and therefore, high sequence conservation is observed in these loops. Furthermore, loops 2 and 3 of Möbius cyclotides contain mainly hydrophilic residues, whereas Bracelet cyclotides contain mostly hydrophobic residues in those loop regions. Additionally, Bracelet cyclotides are noted to be more positively charged due to the presence of Lys or Arg residues in loops 5 and 6 as compared to the Möbius cyclotides which are mostly neutral in global net charge (Chen et al., 2005; Wang et al., 2008b). In general, ~60% of the residues in the Möbius cyclotides are highly conserved i.e. one conserved residue in at least five of the loop regions, compared to Bracelet cyclotides which are more flexible (Pelegrini et al., 2007). As the asparaginyl endopeptidases (AEPs) cleave the linear precursors by recognising the Asx residue in loop 6 and further facilitate circularisation, this position is also conserved by accommodating only Asn or Asp residue (Nguyen et al., 2014). In addition to the six conserved cysteine residues and the C-terminal loop 6 conserved Asx residue, Glu in loop 1 is also conserved in the Bracelet and Möbius subfamily members (Wang et al., 2011) (**Figure 1.5**). This residue forms a hydrogen bonding network with residues in loop 3 and hence, is likely essential for stabilising the

cyclotide framework. Furthermore, Bracelet subfamily members possess a highly conserved Pro residue in loop 2 (Wang et al., 2008b). All of these sequence features become imperative for the cyclotide to adopt a compact structure enough to create the smallest possible ring that enables the third disulfide bonds to pass through the first two disulfide bonds.

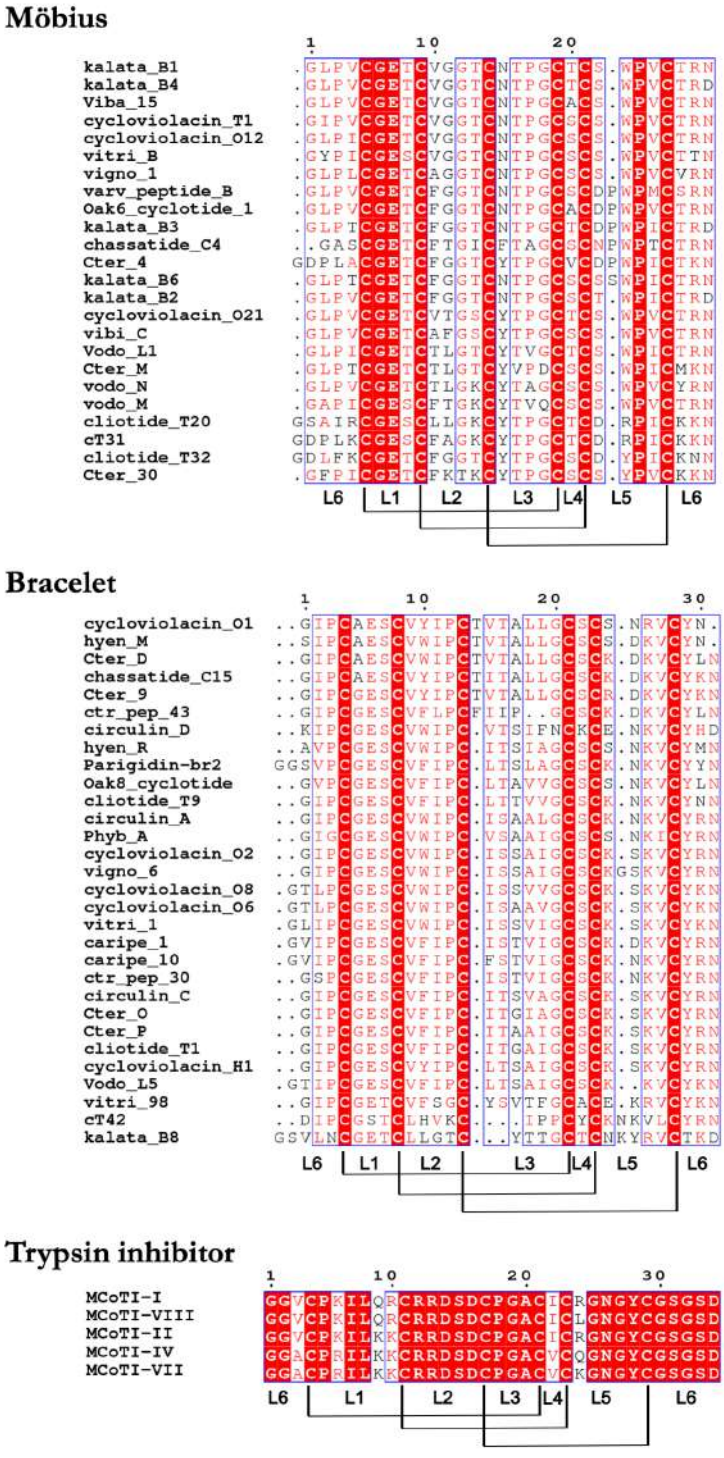


Figure 1.5: Sequence comparison of Möbius, Bracelet and Trypsin inhibitor cyclotides. Sequences were taken from Cybase and were aligned within their respective subfamily. All

six loop regions and disulfide connectivities are marked. Fully conserved positions along the sequences are highlighted in red blocks.

A third subfamily, the trypsin inhibitor (TI) subfamily, was later defined as a cyclotide to classify *Momordica cochinchinensis* trypsin inhibitors (Hernandez et al., 2000), e.g. MCoTI-I and MCoTI-II (**Figure 1.5**). The TI subfamily members do not share sequence similarity to either Möbius and Bracelets, and have been identified only in the Cucurbitaceae plant family. However, due to their cyclized backbone and conserved CCK motif, they have been included in the cyclotide family. TI cyclotide members are relatively more hydrophilic than the Möbius and Bracelet cyclotides, and this is due to the presence of more positively-charged residues and fewer hydrophobic residues (Cascales et al., 2011).

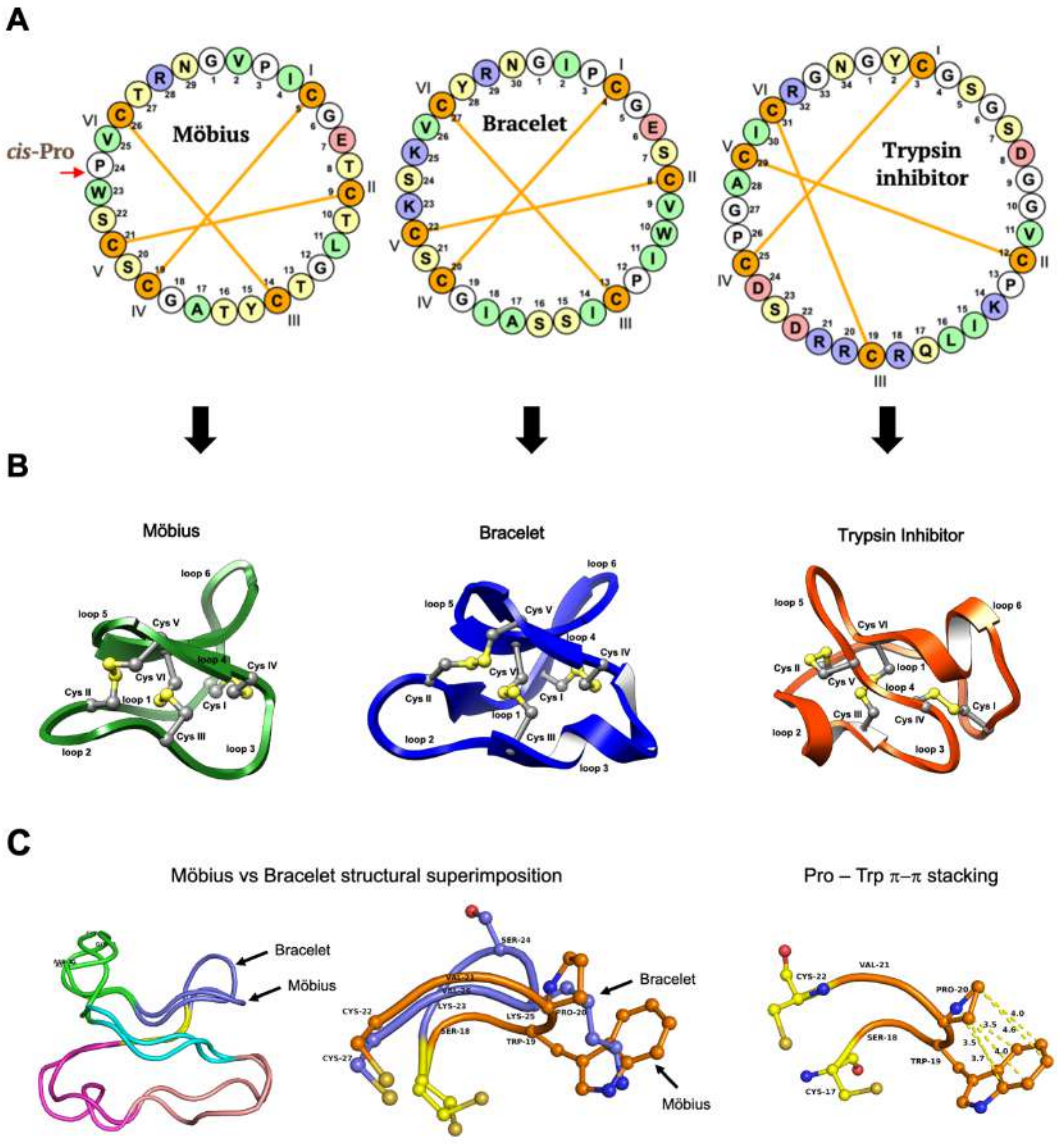


Figure 1.6: Sequence and structure differences between Möbius, Bracelet and Trypsin Inhibitor subfamilies of cyclotides. (A) Cyclic sequence wheel representation of the sequence differences of cyclotides between three subfamilies. Disulfide connectivity are

indicated by orange lines. (B) Three dimensional structures of cyclotides. A ribbon diagram of solution structures elucidated by NMR spectroscopy are shown for Möbius type Cter M (PDB ID 2LAM), bracelet type cycloviolacin O2 (PDB ID 2KNM) and trypsin inhibitor cyclotide MCoTI-I (PDB ID 5WOV). The loops are numbered in arabic numericals and cysteine residues are marked with roman letters. (C) Structural superimposition of Möbius type varv F (PDB ID 3E4H) and bracelet type cycloviolacin O2 (PDB ID 2KNM) structures. The Pro-Trp Ch/π interaction is illustrated.

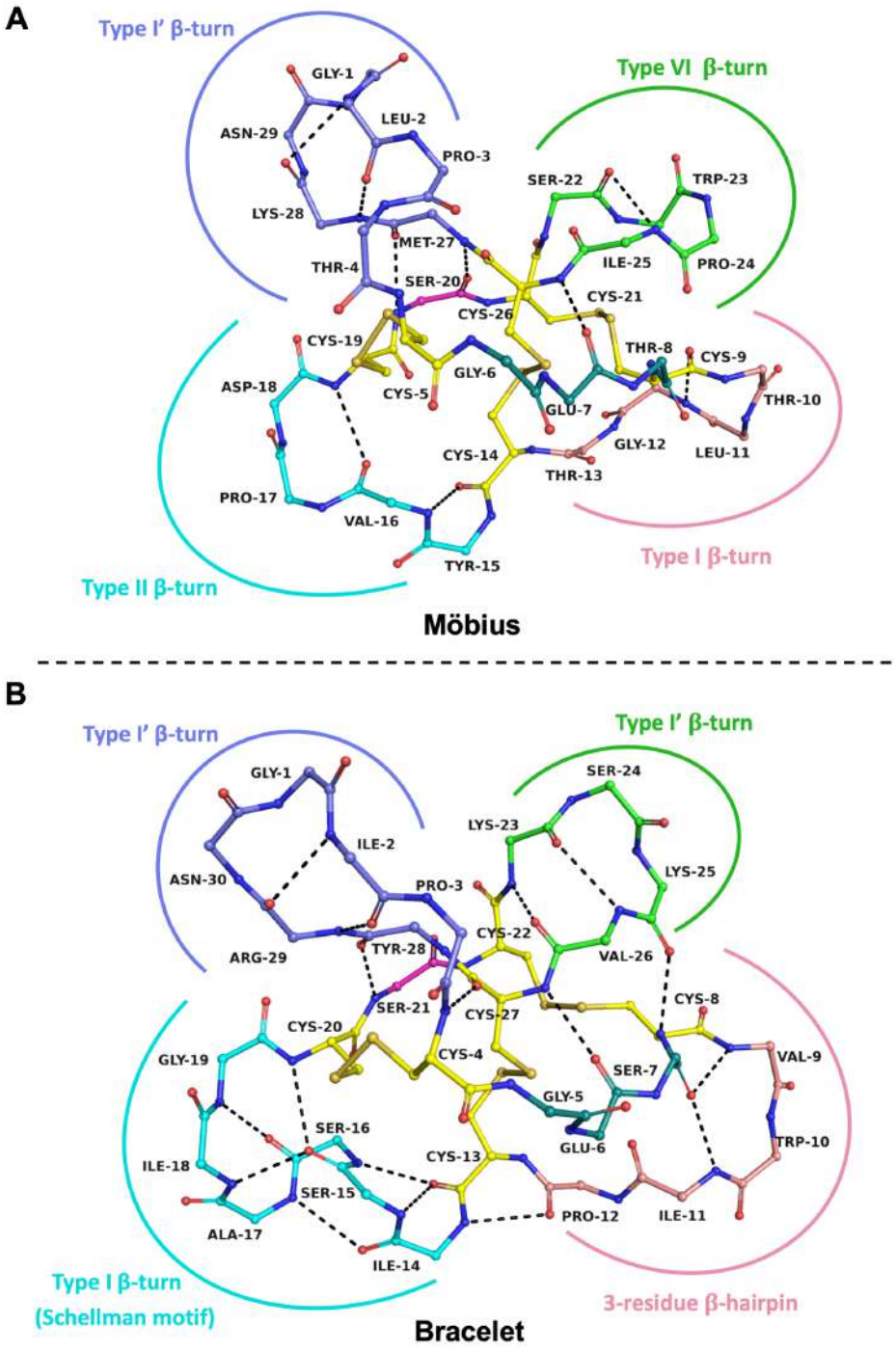


Figure 1.7: Three-dimensional structure comparison between (A) Möbius-type Cter M (PDB ID 2LAM) featuring a type VI β-turn in loop 5 region and (B) Bracelet-type cycloviolacin O2 (PDB ID 2KNM), featuring type I' β-turn in loop 5 region. Six conserved

Cys residues involved in disulfide bond formation are marked in yellow ball-and-stick. Loop 1 is highlighted in teal blue, loop 2 in peach, loop 3 in cyan, loop 4 in magenta, loop 5 in green and loop 6 in purple.

All the three structural subfamilies have the same CCK motif but differ from each other based on the composition of amino acid residues (**Figure 1.6A & B)**. Additionally, there are few numbers of cyclotides classified as 'hybrid' cyclotides that are an amalgamation of both Möbius and Bracelet type sequence characteristics (Daly et al., 2006; Mulvenna et al., 2005). The presence of a *cis*-proline in loop 5 in the Möbius group is the characteristic difference with the Bracelet group of sequences. The Xxx-Pro bond between an aromatic residue (n-1 residue) and proline residue (n^{th} residue) creates a Ch/π interaction which constrains in twist-like conformation in the peptide backbone for the Möbius-type sequences (**Figure 1.6C**). Several three-dimensional structures have been reported for cyclotides, but mainly have been reported using solution nuclear magnetic resonance spectroscopy. They have all been characterized to possess a similar fold and well-defined structural features. Most cyclotides have a triple-stranded β-sheet and several turns (**Figure 1.6B**). Analysis of the structure of Cter M (PDB ID 2LAM), which belongs to the Möbius subfamily, reveals a type I' β-turn in loop 6, a type I β-turn in loop 2, a type II β-turn in loop 3, and a type VI β-turn in loop 5. A β-hairpin motif is seen in residues 20–27, and this is typical of several cystine knot proteins (**Figure 1.7A**). On the other hand, the structure of cycloviolacin O2 (PDB ID 2KNM) belonging to Bracelet subfamily shows that loop 6 adopts a type I' β-turn, loop 2 adopts a three-residue β-hairpin, a type I β-turn in loop 3 which also forms a Schellman motif, and a type I' β-turn in loop 5 (**Figure 1.7B**).

1.4.1 *C. ternatea* cyclotide diversity

C. ternatea is the only member of the Fabaceae plant family known to synthesize copious amounts of cyclotides across all its tissues. Through proteomic and transcriptomic analyses, ~110 cyclotides have been discovered from *C. ternatea* (**Table 1.1**). Unlike its counterparts, the Fabaceae cyclotide genes are embedded in the pea-like albumin-1 gene and occur with a single cyclotide domain. This architecture contains the ER signal peptide, followed by the cyclotide domain which seems to have usurped the albumin-1b chain, followed by a short ~ 10 aa intervening linker region and ~ 50 aa long albumin-1a chain (**Figure 1.3B**). Additionally, *C. ternatea* cyclotide precursors contain single cyclotide domains, similar to the *Petunia x hybrida* (Solanaceae) precursor sequences, whereas cyclotide precursors from the Violaceae, Rubiaceae and Cucurbitaceae families possess multiple cyclotide domains. Currently, this extensive radiation of albumin-1 multigene family in *C. ternatea* could be explained by two evolutionary tenets i.e. the cyclotide coding region was horizontally

transferred to the albumin gene or the albumin 1b-chain slowly evolved to become the cyclotide domain (Camarero, 2011). While both these are valid explanations, the latter seems more plausible as *C. ternatea* cyclotides were found to be similar to the albumin-1 b-chain (PA1b) of *Pisum sativum* (Poth et al., 2011) in sequence and structure.

Table 1.1: Full list of *C. ternatea* cyclotide sequences.

Name	**Alternative name**	**I**	**Loop 1**	**II**	**Loop 2**	**III**	**Loop 3**	**IV**	**L4**	**V**	**Loop 5**	**VI**	**Loop 6**
cliotide T1		C	GES	C	VFIP	C	ITGAIG	C	S	C	KSKV	C	YRNGIP
cliotide T2		C	GES	C	VQGE	C	YTPG	C	S	C	DWPI	C	KKNGEFLK
cliotide T4	Cter P	C	GES	C	VFIP	C	ITAAIG	C	S	C	KSKV	C	YRNGIP
cliotide T6		C	GES	C	VYIP	C	LTTIVG	C	S	C	KNSV	C	YSNSIP
cliotide T8		C	GES	C	VFIP	C	ISSVVG	C	S	C	KSKV	C	YNNGIP
cliotide T9		C	GES	C	VFIP	C	LTTVVG	C	S	C	KNKV	C	YNNGIP
cliotide T10	Cter 27	C	GES	C	VYIP	C	TVTALLG	C	S	C	KDKV	C	YKNGIP
cliotide T11	Cter 21	C	GES	C	VFIP	C	TITALLG	C	S	C	KDKV	C	YKNGIP
cliotide T12		C	GES	C	VFIP	C	ITGAIG	C	S	C	KSKV	C	YRDGIP
cliotide T13	Cter 23	C	GES	C	VWIP	C	VSSIVG	C	S	C	QNKV	C	YQNDTTP
cliotide T14		C	GES	C	VWIP	C	ISSILG	C	S	C	KDKV	C	YHNDTIP
cliotide T15	Cter 24	C	GET	C	FKTK	C	YTKG	C	S	C	SYPV	C	KRNGLPI
cliotide T17		C	GES	C	VFIP	C	ITGIAG	C	S	C	KNKV	C	YLNGTVP
cliotide T18	Cter 6	C	GET	C	FTGT	C	YTPG	C	T	C	SYPV	C	KKNGLPI
cliotide T19a	Cter 26	C	GES	C	LLGK	C	YTPG	C	T	C	SRPI	C	KKNGSVIK
cliotide T19b		C	GES	C	LLGK	C	YTPG	C	T	C	SRPI	C	KKNGSVIK
cliotide T20		C	GES	C	LLGK	C	YTPG	C	T	C	DRPI	C	KKNGSAIR
cliotide T21	Cter 17	C	AET	C	VHSP	C	IGP	C	Y	C	KHGLI	C	YRNDLQ
cliotide T22[a]		C	GES	C	VWIP	C	TITALVG	C	A	C	HEKV	C	YKSARIP
cliotide T23*		C	GES	C	VFIP	C	TVTALLG	C	S	C	KDKV	C	YKNGFP
cliotide T25*		C	GER	C	LLGR	C	HRPG	C	T	C	VRRI	C	RRNGSIR
cliotide T26*		C	GES	C	VYIP	C	ITALLG	C	S	C	SNQI	C	SKNGFI
cliotide T27*		C	GES	C	VFIP	C	ITGAIG	C	S	C	KSKV	C	YRNGVIP
cliotide T30*		C	GES	C	FAGK	C	YTPG	C	T	C	SRPI	C	KKNGDPLK
cliotide T31*		C	GES	C	FAGK	C	YTPG	C	T	C	DRPI	C	KKNGDPLK
cliotide T32*		C	GET	C	FGGT	C	YTPG	C	S	C	DYPI	C	KNNGDLFK
cliotide T37[a]		C	LET	C	VILP	C	FSSVAG	C	Y	C	HGST	C	MRGVDGF
cliotide T38[a]		C	GES	C	VWIP	C	FTSAFG	C	Y	C	QSKV	C	YHSKIP
cliotide T42*		C	GST	C	LHVK	C	IPP	C	Y	C	KNKVL	C	YRNDIP
cliotide T47[β]		C	GES	C	VYLP	C	LTTIVG	C	S	C	KNNV	C	YTNXIP
cliotide T48[β]		C	GES	C	VFLP	C	FIIPG	C	S	C	KDKV	C	YLNX
cliotide T49[β]		C	GET	C	VLGT	C	YTPD	C	S	C	KAVV	C	XNSAF
cliotide T50[βγ]		C	DQT	C	LMQGK	C	YRSG	C	T	C	DRPX	–	GVSWI
cliotide T51[βγ]		C	GET	C	FMGS	C	YTPG	C	S	C	DAVX	–	GVPL
cliotide T52[βγ]		C	GET	C	FGGT	C	YTPG	C	S	–	X	–	GDALK
cliotide T53[βγ]		C	GET	C	LRGR	C	YTPG	C	–	–	X	–	GSSIVT
Cter 1*	cliotide T35	C	GET	C	FGGT	C	NTPN	C	V	C	DPWPI	C	TNNGLPI
Cter 10*	cliotide T34	C	GES	C	VYIP	C	TVTALLG	C	S	C	SNKV	C	YKNSYIP
Cter 11*	cliotide T24	C	GER	C	LLGR	C	HRPG	C	T	C	IRRI	C	RRNGSIR
Cter 12*		C	GET	C	VLGT	C	YTPD	C	S	C	KAVV	C	IKNNTAF

Cter 13*		C	GER	C	LLGR	C	HRPG	C	T	C	IRRI	C	RRNGSAIR
Cter 14*	cliotide T40	C	GES	C	VFIP	C	TITALLG	C	S	C	KSKV	C	YKNGIP
Cter 15*		C	GES	C	VFIP	C	TVTALLG	C	S	C	KSKV	C	YKNGIP
Cter 16*	cliotide T28	C	GES	C	VFLP	C	FLPG	C	S	C	KSSV	C	YLNGGSIP
Cter 18*	cliotide T43	C	SST	C	LHTP	C	KASV	C	Y	C	KNAV	C	YKNDLI
Cter 19*		C	GES	C	VYIP	C	LTTIVG	C	S	C	KSNV	C	YSNSIP
Cter 2*	cliotide T29	C	GES	C	FAGK	C	YTPG	C	T	C	EYPI	C	MNNGDPLK
Cter 20*		C	GES	C	VYLP	C	LTTIVG	C	S	C	KNNV	C	YTNGVIP
Cter 22*		C	GET	C	VLGT	C	YTPD	C	S	C	TAIV	C	IKNTAF
Cter 25*	cliotide T41	C	GET	C	FFQK	C	YTPG	C	S	C	DAVI	C	TNNGNPIV
Cter 28*	cliotide T36	C	GES	C	VWIP	C	ISAAIG	C	S	C	KKNV	C	YRNGVIP
Cter 29*	cliotide T44	C	DER	C	TYVP	C	ISAARG	C	S	C	NIHRV	C	SMNGAL
Cter 3*		C	GET	C	VLGT	C	YTPD	C	S	C	KAVV	C	IKNGAF
Cter 30*	cliotide T45	C	GET	C	FKTK	C	YTPG	C	S	C	SYPV	C	KKNGFPI
Cter 31*	cliotide T46	C	AET	C	VHSP	C	IGP	C	Y	C	KHGVI	C	YKNDLQ
Cter 32*		C	GES	C	VWIP	C	ISSILG	C	S	C	KDKV	C	YHNKIP
Cter 33*		C	GET	C	FGGT	C	YTPG	C	S	C	DYPI	C	KKNGDLFK
Cter 34*	cliotide T33	C	SEA	C	VYLP	C	FSKG	C	S	C	FKRQ	C	YKNGFNS
Cter 35*		C	GET	C	VLGT	C	YTPG	C	S	C	APVI	C	LNNGAF
Cter 36*		C	GET	C	FGGT	C	YTPN	C	V	C	DPWPI	C	TKNGSPT
Cter 37*		C	GET	C	FGGT	C	YTPG	C	V	C	DPWPI	C	TKNGSPT
Cter 4*	cliotide T39	C	GET	C	FGGT	C	YTPG	C	V	C	DPWPI	C	TKNGDPLA
Cter 5*		C	GES	C	VQGE	C	YTPG	C	S	C	DYPI	C	KNNGGEFLK
Cter 7*		C	GES	C	FAGK	C	YTPG	C	T	C	EYPI	C	MNNGDPFK
Cter 8*		C	GET	C	VLGT	C	YTPD	C	S	C	KAVV	C	IKNGSAF
Cter 9*		C	GES	C	VYIP	C	TVTALLG	C	S	C	RDKV	C	YKNGIP
Cter A		C	GES	C	VFIP	C	ISTVIG	C	S	C	KNKV	C	YRNGVIP
Cter B		C	AES	C	VWIP	C	TVTALLG	C	S	C	KDKV	C	YLNGVP
Cter C		C	AES	C	VWIP	C	TVTALLG	C	S	C	KDKV	C	YLDGVP
Cter D		C	AES	C	VWIP	C	TVTALLG	C	S	C	KDKV	C	YLNGIP
Cter E		C	AES	C	VWIP	C	TVTALLG	C	S	C	KDKV	C	YLDGIP
Cter F		C	GES	C	VFIP	C	ISSVVG	C	S	C	KSKV	C	YLDGIP
Cter G		C	GES	C	VFIP	C	ITTVVG	C	S	C	KNKV	C	YNNGLP
Cter H		C	GES	C	VFIP	C	ITTVVG	C	S	C	KNKV	C	YNDGLP
Cter I		C	GES	C	VFIP	C	ITGIAG	C	S	C	KNKV	C	YINGTVP
Cter J		C	GES	C	VFIP	C	ITGIAG	C	S	C	KNKV	C	YIDGTVP
Cter K		C	GES	C	VFIP	C	ITTVVG	C	S	C	KNKV	C	YNHEP
Cter L		C	GES	C	VFIP	C	ITTVVG	C	S	C	KNKV	C	YDHEP
Cter M	cliotide T3	C	GET	C	TLGT	C	YVPD	C	S	C	SWPI	C	MKNGLPT
Cter N		C	GET	C	VLGT	C	YTPD	C	S	C	TALV	C	LKNGSAF
Cter O		C	GES	C	VFIP	C	ITGIAG	C	S	C	KSKV	C	YRNGIP
Cter Q	cliotide T5	C	GES	C	VFIP	C	ISTVIG	C	S	C	KNKV	C	YRNGIP
Cter R	cliotide T7	C	GES	C	VFIP	C	TVTALLG	C	S	C	KDKV	C	YKNGIP
Cterneg C1$^{\alpha\gamma}$		–	GET	C	VLQT	C	YTPG	C	S	C	TIAI	C	LNNGSPLLR
ctr pep 3		C	GET	C	FKTK	C	YTPG	C	S	C	SYPI	C	KKNGRPT
ctr pep 4		C	GET	C	FGGT	C	YTPG	C	S	C	DYPI	C	KKNGDALK
ctr pep 5$^{\gamma}$		–	–	–	FKTK	C	YTPG	C	S	C	SYPV	C	KRN
ctr pep 6		C	GET	C	FKTK	C	YTKG	C	S	C	SYPI	C	KKNGCLPI
ctr pep 8		C	GET	C	FKTK	C	YTKG	C	S	C	SYPV	C	KRNGRPT
ctr pep 11		C	GP	C	LTDE	C	WTPG	C	EYH	C	KY	C	KNSANIPMT
ctr pep 12		C	GET	C	TLGT	C	YVPD	C	S	C	SWPI	C	MKNXLPT
ctr pep 14$^{\gamma}$		–	S	C	VWIP	C	ITGAIG	C	S	C	KNRV	C	YRN

ctr pep 18		C	GES	C	VFIP	C	ITGIAG	C	S	C	KNKV	C	YLNGIS
ctr pep 19[γ]		–	ES	C	VWIP	C	LTGYFG	C	Y	C	QSKV	C	YHN
ctr pep 20		C	GES	C	VFIP	C	ITGAIG	C	S	C	KSKV	C	YRNCVIP
ctr pep 21[γ]		C	GES	C	VWIP	C	TTTALVG	C	A	C	HEK	–	–
ctr pep 22		C	GGG	C	LLGR	C	YRPG	C	T	C	VRRI	C	RRNGSITG
ctr pep 24		C	GER	C	LLGR	C	HRPG	C	T	C	VRRI	C	RRNGSAIR
ctr pep 26		C	GET	C	LRGK	C	YTPG	C	T	C	VRPI	C	KKNGSSIVT
ctr pep 29		C	GES	C	FAGK	C	YTPG	C	T	C	SRPI	C	KKNGDPFK
ctr pep 30		C	GES	C	VFIP	C	ISTVIG	C	S	C	KNKV	C	YRNGSP
ctr pep 31		C	GES	C	VYIP	C	TVTALLG	C	S	C	KNKV	C	FRNGIP
ctr pep 33		C	GES	C	VFIS	C	TVTALLG	C	S	C	KDKV	C	YKNGIP
ctr pep 35		C	GES	C	VYIP	C	ITSIVG	C	S	C	QNKV	C	YNNNIFP
ctr pep 37		C	GES	C	VFIP	C	TSSVLG	C	S	C	KDKF	C	YNNDTIP
ctr pep 38		C	GES	C	IYLP	C	ISGVFEG	C	S	C	QNKA	C	YKNGGTA
ctr pep 43		C	GES	C	VFLP	C	FIIPG	C	S	C	KDKV	C	YLNGIP
ctr pep 47		C	VES	C	VFIP	C	TVTALLG	C	S	C	KDKV	C	YKNGIP
ctr pep 48		C	GES	C	VYIP	C	TVTALLG	C	S	C	KNKV	C	YRNGVP
ctr pep 50[γ]		–	–	–	ALP	–	ISTIVG	C	S	C	KSNV	C	YSN

* no mass-spectrometry derived data available; [α] predicted to be linear; [β] ambiguous sequences; [γ] partial sequences

1.5 ACTIVITIES AND MODE OF ACTION

1.5.1 Natural activities

1.5.1.1 Insecticidal activity

Cyclotides are produced in plants for defense against pests and pathogens based on their potent insecticidal and antimicrobial activities (Jennings et al., 2001; Tam et al., 1999). They also display wide-ranging biological activities, including uterotonic, molluscicidal, anti-HIV and cytotoxic to cancer cell lines. Due to the wide range of biological activities and unique properties, they are rendered advantageous for use as ultrastable frameworks for drug design and in agriculture. The insecticidal activity of cyclotides were first demonstrated in 2001 where artificial diets fed to *Helicoverpa punctigera*, a common pest moth seen on economically significant crops like cotton and legumes, were supplemented with kalata B1 (Jennings et al., 2001). It was observed that growth of larvae was arrested at first instar when fed with kB1 at a comparable concentration present in *O. affinis* leaves. Subsequently, cyclotide from *C. ternatea* (Cter M), had similar insecticidal activity against *Helicoverpa armigera* larvae (Poth et al., 2011). In 2016, Gilding et al., reported that cyclotide containing extracts prepared from the aerial tissues of *C. ternatea* interact more strongly with insect like lipid membranes than those prepared from roots, for instance, the shoot extracts exhibiting the greatest potency (0.31 mg/mL LC50) (Gilding et al., 2016). On the other hand, the root extracts show potent nematocidal activity against soil-living nematodes compared to the extracts from aerial parts. More recent reports have demonstrated highest cytotoxic

activity against Sf9 (*Spodoptera frugiperda*) insect cells by extracts from butterfly pea accessions with the highest cyclotide expression (Oguis et al., 2020). As the primary role of cyclotide production in plants is for defence purposes, cyclotides have gained significant interest in agriculture, especially from the recent release of an eco-friendly pesticide - SeroX®, made from *C. ternatea* extracts and marketed in Australia for use on cotton and macadamia crops. In 2017, an Australian company, InnovateAg successfully commercialised SeroX®, which is a mixture of cyclotides that showed activity against larvae and nymph of several insects such as native budworms (*Helicoverpa spp.*), apple dimpling bug (*Campylomma liebknechti*), southern green stink bug (*Nezara viridula*), green mirid (*Creontiades dilutus*) and silverleaf whiteflies (*Bemisia tabaci*).

1.5.1.2 Antimicrobial activity

Cyclotides are reported to exhibit greater antimicrobial activity against Gram-negative bacteria than Gram-positive bacteria (Nguyen et al., 2016; Pränting et al., 2010; Tam et al., 1999). Antimicrobial activity of *C. ternatea* cyclotides against Gram-negative has been established, but not against Gram-positive bacteria (Nguyen et al., 2011, 2016). The main reason for this is believed to be because of the structural difference and composition of the cell walls between Gram-negative and Gram-positive bacteria. Cycloviolacin O2 has been proven to be the most potent antimicrobial cyclotide, as it showed antimicrobial activity against *Staphylococcus aureus* in a mouse infection model (Fensterseifer et al., 2015). Cyclotides act on bacteria by membrane permeabilization through interaction with phosphatidylethanolamine (PE) phospholipids, while electrostatic and hydrophobic parameters bring about broad-spectrum activity (Henriques et al., 2019; Strömstedt et al., 2017). The binding of cyclotides to the PE phospholipids on the membranes leads to cyclotide internalization, disruption of the physical integrity of the membrane, formation of pores, and finally leakage of cellular contents. However, the observed antimicrobial activity was established under non-physiological conditions i.e. in low ionic strength buffers, which significantly limits its potential as potent antimicrobial agents.

1.5.2 Cyclotides in drug discovery

1.5.2.1 Anti-HIV activity

In one of the first of its kind studies, circulins A and B (Bracelet cyclotides isolated from *Chassalia parvifolia*) were reported to have antiviral cytoprotective effects on various human immunodeficiency virus (HIV) strains at concentrations ranging from 40-260 nM (Gustafson et al., 1994). Anti-HIV activity were subsequently verified with several

cyclotides such as circulins C−F isolated from *C. parvifolia* (Gustafson et al., 2000), cycloviolins A−D isolated from *Leonia cymosa* (Hallock et al., 2000), and palicourein from *Palicourea condensate* (Bokesch et al., 2001). Cycloviolins A-D and palicourein inhibited the cytopathic effects of HIV-1 infection in human T-lymphoblastoid cell lines, with EC50 value of 130 nM and 100 nM. However, the exact molecular mode of action is not fully understood.

1.5.2.2 Anticancer activity

Another activity of cyclotides that has been identified is the cytotoxic activity against several human cancer cell lines. Cycloviolacins and vitri A from Violaceae plants have shown activity against lymphoma and myeloma cell lines at 0.96 and 5.0 mM concentrations with potency similar to the chemotherapy drugs used in cancer treatments (Lindholm et al., 2002; Svangård et al., 2004). Another cyclotide vigno 5 significantly reduced Hela cell survival and induced cell apoptosis mainly via the mitochondrial pathway, including the release of cytochrome C to the cytosol and caspases activation in cervical cancer cells (Esmaeili et al., 2016). Additional cyclotides that are also cytotoxic to cancer cells include psyle E from *Psychotria leptothyrsa* (Gerlach et al., 2010), hedyotide B7 from *Hedyotis biflora* (Ding et al., 2014), and cliotides from *C. ternatea* are also cytotoxic to cancer cells (Zhang et al., 2013). Cyclotides have a higher binding affinity towards PE phospholipids on the membranes and thus can preferentially target membranes of cancer cells over the normal cells whose membranes mainly composed of phosphatidylcholine (PC) lipids (Troeira Henriques et al., 2014).

1.5.2.3 Neurological diseases

Few years ago, synthetically created cyclotides, analogous to the *O. affinis* plant-derived peptide, could be orally administered in animal models to suppress multiple sclerosis (MS), an autoimmune disease affecting the central nervous system. The immunosuppressive effects of kalata B1 (kB1) from *O. affinis* on activated human T-lymphocytes was examined through in vitro studies (Gründemann et al., 2012, 2013). Subsequently, [T20K] kB1 cyclotide in particular was described to inhibit T-cell proliferation by down-regulation of interleukin-2 (IL-2) release as well as IL-2R/CD25 surface expression without inducing major adverse effects in mouse model of autoimmune encephalomyelitis or multiple sclerosis (Thell et al., 2016). Cyclotides displayed immense potential as peptide therapeutics for drug development for the treatment of diseases affecting the central nervous system and are now on their way from preclinical studies to clinical trials for blocking the

progression of multiple sclerosis. More recently, cyclotides' anti-neurodegenerative properties were studied in *Psychotria solitudinum*, where they acted as inhibitors of the human prolyl oligopeptidase (POP), a promising target for the treatment of cognitive deficits in several psychiatric and neurodegenerative diseases. Cyclotides were tested as an inhibitor of serine-type protease, namely POP, through pharmacology-guided fractionation of several *Psychotria* species (Hellinger, 2018; Hellinger et al., 2015). The isolated peptide psysol 2 showed inhibitory potency (IC_{50}) of 25 μM. In addition, the kalata B1 inhibited POP activity with an IC_{50} of 5.6 μM. Prolyl oligopeptidase, an 81 kDa protein, is considered as a promising therapeutic target. Inhibition of POP protease leads to neuroprotection and enhancement of cognition. These findings have promoted the development of POP inhibitors for the treatment of several neurodegenerative diseases, such as Parkinson's disease, where cognition-related deficits are associated. In both these studies the identified peptides were studied through target-based drug discovery and pharmacologically characterized. Highly constrained cyclic peptides have gained special attention for the modulation of protein-protein interactions. Cyclic peptides, SFTI-1 (sunflower trypsin inhibitor-1) and kB1, showed inhibitory activity against hexapeptide AcVQIVYK-NH2 (AcPHF6) involved in tau fibril formation, highlighting their appeal as candidates for inhibition of fibril formation. This study highlights the promise of cyclic peptide scaffolds as leads for fibril inhibitor design. Collectively, the reports mentioned above demonstrate the therapeutic potential of cyclotides for drug design and development in several age-related diseases, such as Alzheimer's or Parkinson's disease.

1.6 AIM OF THE THESIS

The broad aims of the work reported in this thesis are to:

(1) Examine the underlying mechanisms of cyclotide diversity and biosynthesis in *Clitoria ternatea* (Fabaceae) by transcriptomic approaches.
(2) Survey the proteomic evidences and variation of cyclotides across multiple tissues of *C. ternatea* using high-resolution mass-spectrometric approaches.
(3) Develop a diagnostic method for mass-spectral fingerprinting to facilitate rapid identification of novel cyclotides in a mixture.
(4) Develop and update DSDBASE2.0, a database of native and modelled disulphides in proteins and include Protein Data Bank entries from with the latest release.
(5) Investigate cyclotides' therapeutic potential as an inhibitor of β-amyloid (Aβ) toxicity and demonstrate the activities using transgenic *Caenorhabditis elegans* model organism.

1.7 SIGNIFICANCE AND SCOPE OF THE THESIS

Cyclotides present themselves as fascinating biomolecules with a unique set of structural characteristics and a wide array of pharmaceutical applications. Cyclotides posses the unique 'CCK motif' structural feature formed by six cysteine residues and cyclic backbone. Furthermore, naturally isolated cyclotides have displayed several therapeutically relevant biological activities making them as attractive scaffolds for peptide-based drug design. *Clitoria ternatea* or butterfly pea is a perennial climber of Fabaceae plant family. Also known as Shankhapushpi, Aparajita, Sangu poo, etc in India, it is widely used in traditional Ayurvedic medicine as an enhancer of cognitive functions, nootropic, tranquilizing and sedative agent, blood platelet aggregation-inhibitor and to treat pain, fever and inflammation (Mukherjee et al., 2008). Apart from considered as a pharmaceutical treasure troves, cyclotides' insecticidal property makes them as desirable candidates in agriculture too. In 2017, a biopesticide, SeroX® (https://innovate-ag.com.au/our-product/) based on plant extract was commercialised. Collectively, understanding structural diversity and biosynthesis of cyclotides is crucial for fundamental research as well as developing potential future applications.

The introductory chapter (**Chapter 1**) in this thesis provides a general overview of cyclotides, their sequence variation, structural characteristics and biological activities, underscoring on those specifically produced in *C. ternatea*. The diversity, the biosynthesis machinery and their unique structural features are outlined, along with the potential pharmaceutical applications. In **Chapter 2**, key results of our work on the usage of RNA-Seq data for *de novo* assembly of the transcriptome of *C. ternatea* and identification of ~70 precursor genes of cyclotides arising from multiple tissues are presented. Further, I have proposed a model of cyclotide biosynthesis, wherein I followed the presence and expression of oxidative folding enzymes such as PDIs, ERO-1 and PPIases that are probably responsible for proper processing of cyclotides. The third chapter (**Chapter 3**) outlines the usage of high-resolution mass-spectrometry and proteomics methods to get a complete suite of cyclotides present in *C. ternatea* and their differential expression across the tissues. This chapter also describes our attempts to develop a protocol for rapid identification of novel cyclotides based on distinct MS/MS fragmentations patterns observed in prototypical cyclotides used as templates. **Chapter 4** discusses our efforts on improving DSDBASE database (http://caps.ncbs.res.in/dsdbase2). The database has been updated to include over 150,000 entries from PDB January 2021 release and the web interface provides several user-friendly tools for disulphide bonds of proteins. DSDBASE2.0 presently remains as the largest open-access repository that organises all

disulphide bonds of proteins in single platform. A curated database like this can be a very useful resource for future research related to disulfide-rich proteins such as cyclotides. In the penultimate chapter (**Chapter 5**), I have discussed the potential of such cysteine-rich peptides as therapeutic leads, explicitly as potential inhibitors of β-amyloid fibrils that are implicated in Alzheimer's disease (AD). The study was performed using transgenic *Caenorhabditis elegans*, a popular model organism in recent times to understand β-amyloid induced toxicity and progression of AD. Cyclotides delayed Aβ-induced paralysis in transgenic *C. elegans* strain expressing muscle-specific $A\beta_{1-42}$ gene (CL4176). Cyclotides also ameliorated Aβ-induced defects in chemotaxis behaviour in a different transgenic *C. elegans* strain (CL2355) expressing the $A\beta_{1-42}$ gene pan-neuronally. Additionally, the deposition of Aβ were significantly reduced in the transgenic in *C. elegans* (CL2006). As overwhelming amount of data support the hypothesis that oxidative stress is an important operator in the pathogenesis of AD, we performed the reactive oxygen species (ROS) assay. The results showed that cyclotides are potent antioxidants and the protection against Aβ toxicity is most likely mediated by an antioxidant mechanism. This concluding work provides insights for the first time that cyclotides are novel pharmacophore leads against neurodegenerative diseases, particularly against β-amyloid fibril formation. To sum up, **Chapter 6** provides an overall discussion of the chapters and key results mentioned above. A broad conclusion and future research directions are also given in this concluding chapter.

1.8 REFERENCES OF CHAPTER 1

Bokesch, H. R., Pannell, L. K., Cochran, P. K., Sowder, R. C., McKee, T. C., and Boyd, M. R. (2001). A novel anti-HIV macrocyclic peptide from Palicourea condensata. *J. Nat. Prod.* 64. doi:10.1021/np000372l.

Burman, R., Herrmann, A., Tran, R., Kivelä, J. E., Lomize, A., Gullbo, J., et al. (2011). Cytotoxic potency of small macrocyclic knot proteins: Structure-activity and mechanistic studies of native and chemically modified cyclotides. *Org. Biomol. Chem.* 9. doi:10.1039/c0ob00966k.

Burman, R., Yeshak, M. Y., Larsson, S., Craik, D. J., Rosengren, K. J., and Göransson, U. (2015). Distribution of circular proteins in plants: Large-scale mapping of cyclotides in the Violaceae. *Front. Plant Sci.* 6. doi:10.3389/fpls.2015.00855.

Camarero, J. A. (2011). Legume cyclotides shed light on the genetic origin of knotted circular proteins. *Proc. Natl. Acad. Sci. U. S. A.* 108. doi:10.1073/pnas.1107849108.

Cascales, L., Henriques, S. T., Kerr, M. C., Huang, Y. H., Sweet, M. J., Daly, N. L., et al. (2011). Identification and characterization of a new family of cell-penetrating peptides: Cyclic cell-penetrating peptides. *J. Biol. Chem.* 286. doi:10.1074/jbc.M111.264424.

Chen, B., Colgrave, M. L., Daly, N. L., Rosengren, K. J., Gustafson, K. R., and Craik, D. J. (2005). Isolation and characterization of novel cyclotides from Viola hederaceae: Solution structure and anti-HIV activity of vhl-1, a leaf-specific expressed cyclotide. *J. Biol. Chem.* 280. doi:10.1074/jbc.M501737200.

Colgrave, M. L., and Craik, D. J. (2004). Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: The importance of the cyclic cystine knot. *Biochemistry* 43, 5965–5975. doi:10.1021/bi049711q.

Colgrave, M. L., Kotze, A. C., Ireland, D. C., Wang, C. K., and Craik, D. J. (2008). The anthelmintic activity of the cyclotides: Natural variants with enhanced activity. *ChemBioChem* 9. doi:10.1002/cbic.200800174.

Conlan, B. F., Colgrave, M. L., Gillon, A. D., Guarino, R., Craik, D. J., and Anderson, M. A. (2012). Insights into processing and cyclization events associated with biosynthesis of the cyclic peptide kalata B1. *J. Biol. Chem.* 287. doi:10.1074/jbc.M112.347823.

Daly, N. L., Clark, R. J., Plan, M. R., and Craik, D. J. (2006). Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif. *Biochem. J.* 393, 619–626. doi:10.1042/BJ20051371.

De Veer, S. J., Kan, M. W., and Craik, D. J. (2019). Cyclotides: From Structure to Function. *Chem. Rev.* 119. doi:10.1021/acs.chemrev.9b00402.

Ding, X., Bai, D., and Qian, J. (2014). Novel cyclotides from Hedyotis biflora inhibit proliferation and migration of pancreatic cancer cell in vitro and in vivo. *Med. Chem. Res.* 23. doi:10.1007/s00044-013-0746-6.

Esmaeili, M. A., Abagheri-Mahabadi, N., Hashempour, H., Farhadpour, M., Gruber, C. W., and Ghassempour, A. (2016). Viola plant cyclotide vigno 5 induces mitochondria-mediated apoptosis via cytochrome C release and caspases activation in cervical cancer cells. *Fitoterapia* 109. doi:10.1016/j.fitote.2015.12.021.

Fensterseifer, I. C. M., Silva, O. N., Malik, U., Ravipati, A. S., Novaes, N. R. F., Miranda, P. R. R., et al. (2015). Effects of cyclotides against cutaneous infections caused by Staphylococcus aureus. *Peptides* 63. doi:10.1016/j.peptides.2014.10.019.

Gerlach, S. L., Burman, R., Bohlin, L., Mondal, D., and Göransson, U. (2010). Isolation, characterization, and bioactivity of cyclotides from the micronesian plant psychotria leptothyrsa. *J. Nat. Prod.* 73. doi:10.1021/np9007365.

Gerlach, S., and Mondal, D. (2012). The bountiful biological activities of cyclotides. *Chronicles Young Sci.* 3. doi:10.4103/2229-5186.99559.

Gilding, E. K., Jackson, M. A., Poth, A. G., Henriques, S. T., Prentis, P. J., Mahatmanto, T., et al. (2016). Gene coevolution and regulation lock cyclic plant defence peptides to their targets. *New Phytol.* 210, 717–730. doi:10.1111/nph.13789.

Gillon, A. D., Saska, I., Jennings, C. V., Guarino, R. F., Craik, D. J., and Anderson, M. A. (2008). Biosynthesis of circular proteins in plants. *Plant J.* 53. doi:10.1111/j.1365-313X.2007.03357.x.

Göransson, U., Herrmann, A., Burman, R., Haugaard-Jönsson, L. M., and Rosengren, K. J. (2009). The conserved glu in the cyclotide cycloviolacin O2 has a key structural role. *ChemBioChem* 10. doi:10.1002/cbic.200900342.

Gran, L. (1973a). On the Effect of a Polypeptide Isolated from "Kalata-Kalata" (Oldenlandia affinis DC) on the Oestrogen Dominated Uterus. *Acta Pharmacol. Toxicol. (Copenh).* 33, 400–408. doi:10.1111/j.1600-0773.1973.tb01541.x.

Gran, L. (1973b). Oxytocic principles of Oldenlandia affinis. *Lloydia* 36, 174–178.

Gruber, C. W., Čemažar, M., Clark, R. J., Horibe, T., Renda, R. F., Anderson, M. A., et al. (2007). A novel plant protein-disulfide isomerase involved in the oxidative folding of cystine knot defense proteins. *J. Biol. Chem.* 282, 20435–20446. doi:10.1074/jbc.M700018200.

Gruber, C. W., Elliott, A. G., Ireland, D. C., Delprete, P. G., Dessein, S., Göransson, U., et al. (2008). Distribution and evolution of circular miniproteins in flowering plants. *Plant Cell* 20, 2471–2483. doi:10.1105/tpc.108.062331.

Gründemann, C., Koehbach, J., Huber, R., and Gruber, C. W. (2012). Do plant cyclotides have potential as immunosuppressant peptides? *J. Nat. Prod.* 75. doi:10.1021/np200722w.

Gründemann, C., Thell, K., Lengen, K., Garcia-Käufer, M., Huang, Y. H., Huber, R., et al. (2013). Cyclotides Suppress Human T-Lymphocyte Proliferation by an Interleukin 2-Dependent Mechanism. *PLoS One* 8. doi:10.1371/journal.pone.0068016.

Gustafson, K. R., Parsons, I. C., Kashman, Y., Cardellina, J. H., McMahon, J. B., Sowder, R. C., et al. (1994). Circulins A and B: Novel HlV-Inhibitory Macrocyclic Peptides from the Tropical Tree Chassalia parnifolia. *J. Am. Chem. Soc.* 116. doi:10.1021/ja00099a064.

Gustafson, K. R., Walton, L. K., Sowder, R. C., Johnson, D. G., Pannell, L. K., Cardellina, J. H., et al. (2000). New circulin macrocyclic polypeptides from Chassalia parvifolia. *J. Nat. Prod.* 63. doi:10.1021/np990432r.

Hallock, Y. F., Sowder, R. C., Pannell, L. K., Hughes, C. B., Johnson, D. G., Gulakowski, R., et al. (2000). Cycloviolins A-D, anti-HIV macrocyclic peptides from Leonia cymosa. *J. Org. Chem.* 65. doi:10.1021/jo990952r.

Harris, K. S., Durek, T., Kaas, Q., Poth, A. G., Gilding, E. K., Conlan, B. F., et al. (2015). Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. *Nat. Commun.* 6. doi:10.1038/ncomms10199.

Haywood, J., Schmidberger, J. W., James, A. M., Nonis, S. G., Sukhoverkov, K. V., Elias,

M., et al. (2018). Structural basis of ribosomal peptide macrocyclization in plants. *Elife* 7. doi:10.7554/eLife.32955.

Hellinger, R. (2018). Cyclotides as novel inhibitors of human prolyl oligopeptidase. *Intrinsic Act.* doi:10.25006/ia.6.s1-a6.1.

Hellinger, R., Koehbach, J., Puigpinós, A., Clark, R. J., Tarragó, T., Giralt, E., et al. (2015). Inhibition of human prolyl oligopeptidase activity by the cyclotide psysol 2 isolated from psychotria solitudinum. *J. Nat. Prod.* doi:10.1021/np501061t.

Hemu, X., Sahili, A. El, Hu, S., Wong, K., Chen, Y., Wong, Y. H., et al. (2019). Structural determinants for peptide-bond formation by asparaginyl ligases. *Proc. Natl. Acad. Sci. U. S. A.* 116, 11737–11746. doi:10.1073/pnas.1818568116.

Henriques, S. T., Peacock, H., Benfield, A. H., Wang, C. K., and Craik, D. J. (2019). Is the Mirror Image a True Reflection? Intrinsic Membrane Chirality Modulates Peptide Binding. *J. Am. Chem. Soc.* 141. doi:10.1021/jacs.9b11194.

Hernandez, J. F., Gagnon, J., Chiche, L., Nguyen, T. M., Andrieu, J. P., Heitz, A., et al. (2000). Squash trypsin inhibitors from Momordica cochinchinensis exhibit an atypical macrocyclic structure. *Biochemistry* 39. doi:10.1021/bi9929756.

Jackson, M. A., Gilding, E. K., Shafee, T., Harris, K. S., Kaas, Q., Poon, S., et al. (2018). Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases. *Nat. Commun.* 9. doi:10.1038/s41467-018-04669-9.

Jennings, C., West, J., Waine, C., Craik, D., and Anderson, M. (2001). Biosynthesis and insecticidal properties of plant cyclotides: The cyclic knotted proteins from Oldenlandia affinis. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10614–10619. doi:10.1073/pnas.191366898.

Lindholm, P., Goränsson, U., Johansson, S., Claeson, P., Gullbo, J., Larsson, R., et al. (2002). Cyclotides: A novel type of cytotoxic agents. *Mol. Cancer Ther.* 1.

Mukherjee, P. K., Kumar, V., Kumar, N. S., and Heinrich, M. (2008). The Ayurvedic medicine Clitoria ternatea-From traditional use to scientific assessment. *J. Ethnopharmacol.* 120, 291–301. doi:10.1016/j.jep.2008.09.009.

Mulvenna, J. P., Mylne, J. S., Bharathi, R., Burton, R. A., Shirley, N. J., Fincher, G. B., et al. (2006). Discovery of cyclotide-like protein sequences in graminaceous crop plants: Ancestral precursors of circular proteins? *Plant Cell* 18. doi:10.1105/tpc.106.042812.

Mulvenna, J. P., Sando, L., and Craik, D. J. (2005). Processing of a 22 kDa precursor protein to produce the circular protein tricyclon A. *Structure* 13. doi:10.1016/j.str.2005.02.013.

Nguyen, G. K. T., Wang, S., Qiu, Y., Hemu, X., Lian, Y., and Tam, J. P. (2014). Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nat. Chem. Biol.* 10, 732–738. doi:10.1038/nchembio.1586.

Nguyen, G. K. T., Zhang, S., Nguyen, N. T. K., Nguyen, P. Q. T., Chiu, M. S., Hardjojo, A., et al. (2011). Discovery and characterization of novel cyclotides originated from chimeric precursors consisting of albumin-1 chain a and cyclotide domains in the fabaceae family. *J. Biol. Chem.* 286, 24275–24287. doi:10.1074/jbc.M111.229922.

Nguyen, K. N. T., Nguyen, G. K. T., Nguyen, P. Q. T., Ang, K. H., Dedon, P. C., and Tam, J. P. (2016). Immunostimulating and Gram-negative-specific antibacterial cyclotides from the butterfly pea (Clitoria ternatea). *FEBS J.* 283, 2067–2090. doi:10.1111/febs.13720.

Oguis, G. K., Gilding, E. K., Huang, Y. H., Poth, A. G., Jackson, M. A., and Craik, D. J.

(2020). Insecticidal diversity of butterfly pea (Clitoria ternatea) accessions. *Ind. Crops Prod.* 147. doi:10.1016/j.indcrop.2020.112214.

Pelegrini, P. B., Quirino, B. F., and Franco, O. L. (2007). Plant cyclotides: An unusual class of defense compounds. *Peptides* 28. doi:10.1016/j.peptides.2007.04.025.

Poon, S., Harris, K. S., Jackson, M. A., McCorkelle, O. C., Gilding, E. K., Durek, T., et al. (2018). Co-expression of a cyclizing asparaginyl endopeptidase enables efficient production of cyclic peptides in planta. *J. Exp. Bot.* 69. doi:10.1093/jxb/erx422.

Poth, A. G., Colgrave, M. L., Lyons, R. E., Dalya, N. L., and Craik, D. J. (2011). Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10127–10132. doi:10.1073/pnas.1103660108.

Pränting, M., Lööv, C., Burman, R., Göransson, U., and Andersson, D. I. (2010). The cyclotide cycloviolacin O2 from Viola odorata has potent bactericidal activity against Gram-negative bacteria. *J. Antimicrob. Chemother.* 65. doi:10.1093/jac/dkq220.

Rosengren, K. J., Daly, N. L., Plan, M. R., Waine, C., and Craik, D. J. (2003). Twists, knots, and rings in proteins: Structural definition of the cyclotide framework. *J. Biol. Chem.* 278, 8606–8616. doi:10.1074/jbc.M211147200.

Saether, O., Craik, D. J., Campbell, I. D., Sletten, K., Juul, J., and Norman, D. G. (1995). Elucidation of the Primary and Three-Dimensional Structure of the Uterotonic Polypeptide Kalata B1. *Biochemistry* 34, 4147–4158. doi:10.1021/bi00013a002.

Saska, I., Gillon, A. D., Hatsugai, N., Dietzgen, R. G., Hara-Nishimura, I., Anderson, M. A., et al. (2007). An asparaginyl endopeptidase mediates in vivo protein backbone cyclization. *J. Biol. Chem.* 282, 29721–29728. doi:10.1074/jbc.M705185200.

Schopke, T., Agha, M. I. H., Kraft, R., Otto, A., and Hiller, K. (1993). HAMOLYTISCH AKTIVE KOMPONENTEN AUS VIOLA TRICOLOR L. UND VIOLA ARVENSIS MURRAY. *Sci. Pharm.* 61.

Strömstedt, A. A., Park, S., Burman, R., and Göransson, U. (2017). Bactericidal activity of cyclotides where phosphatidylethanolamine-lipid selectivity determines antimicrobial spectra. *Biochim. Biophys. Acta - Biomembr.* 1859. doi:10.1016/j.bbamem.2017.06.018.

Svangård, E., Göransson, U., Hocaoglu, Z., Gullbo, J., Larsson, R., Claeson, P., et al. (2004). Cytotoxic Cyclotides from Viola tricolor. *J. Nat. Prod.* 67. doi:10.1021/np0301011.

Tam, J. P., Lu, Y. A., Yang, J. L., and Chiu, K. W. (1999). An unusual structural motif of antimicrobial peptides containing end-to-end macrocycle and cystine-knot disulfides. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8913–8918. doi:10.1073/pnas.96.16.8913.

Thell, K., Hellinger, R., Sahin, E., Michenthaler, P., Gold-Binder, M., Haider, T., et al. (2016). Oral activity of a nature-derived cyclic peptide for the treatment of multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* 113, 3960–3965. doi:10.1073/pnas.1519960113.

Troeira Henriques, S., Huang, Y. H., Chaousis, S., Wang, C. K., and Craik, D. J. (2014). Anticancer and toxic properties of cyclotides are dependent on phosphatidylethanolamine phospholipid targeting. *Chembiochem* 15. doi:10.1002/cbic.201402144.

Wang, C. K. L., Clark, R. J., Harvey, P. J., Johan Rosengren, K., Cemazar, M., and Craik, D. J. (2011). The role of conserved glu residue on cyclotide stability and activity: A structural and functional study of kalata B12, a naturally occurring glu to asp mutant.

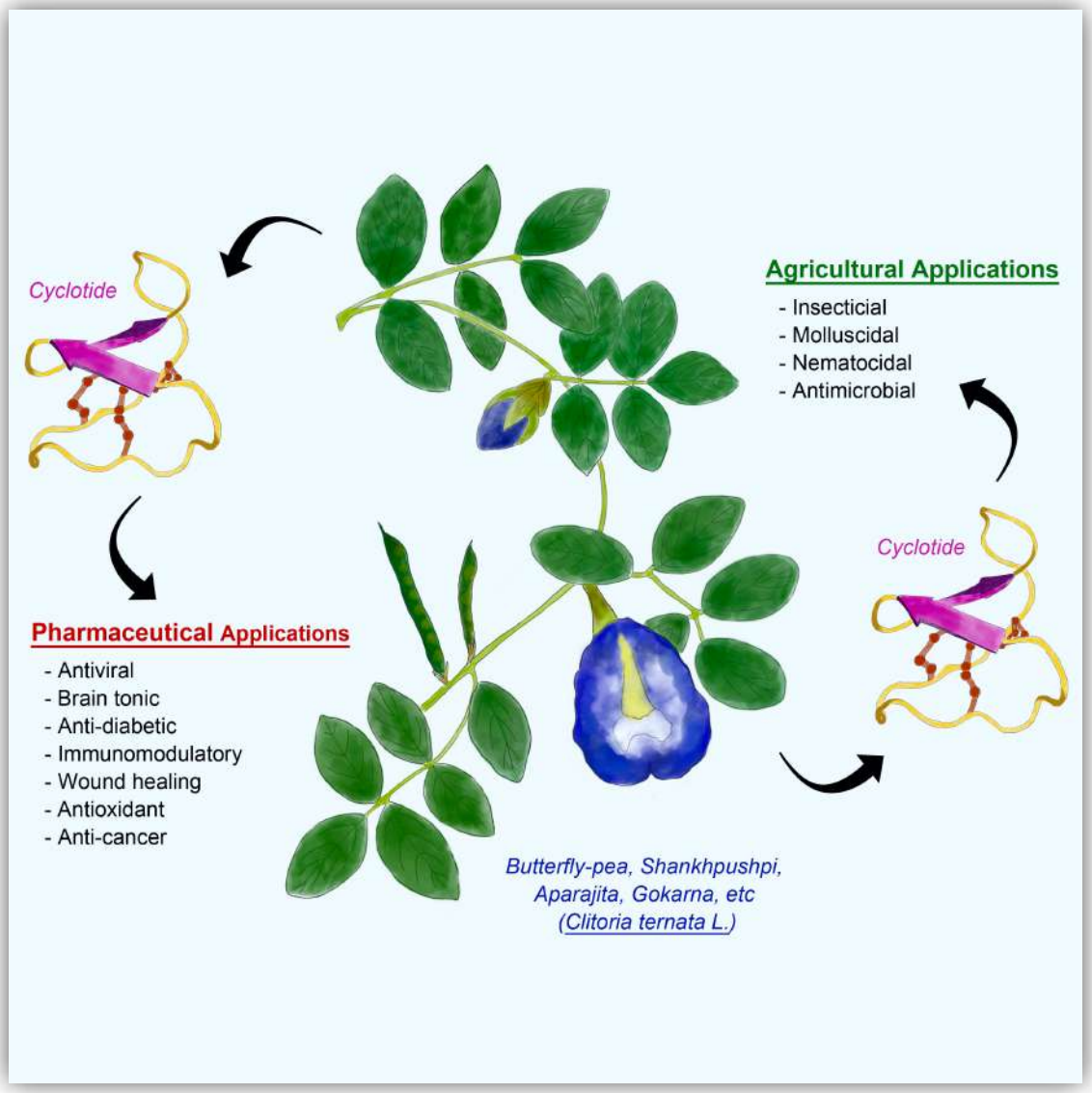
Biochemistry 50, 4077–4086. doi:10.1021/bi2004153.

Wang, C. K. L., Colgrave, M. L., Gustafson, K. R., Ireland, D. C., Goransson, U., and Craik, D. J. (2008a). Anti-HIV cyclotides from the Chinese medicinal herb Viola yedoensis. *J. Nat. Prod.* 71. doi:10.1021/np070393g.

Wang, C. K. L., Kaas, Q., Chiche, L., and Craik, D. J. (2008b). CyBase: A database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res.* 36. doi:10.1093/nar/gkm953.

Zhang, J., Li, J., Huang, Z., Yang, B., Zhang, X., Li, D., et al. (2015). Transcriptomic screening for cyclotides and other cysteine-rich proteins in the metallophyte Viola baoshanensis. *J. Plant Physiol.* 178, 17–26. doi:10.1016/j.jplph.2015.01.017.

Zhang, S., Xiao, K. Z., Jin, J., Zhang, Y., and Zhou, W. (2013). Chemosensitizing activities of cyclotides from clitoria ternatea in paclitaxel-resistant lung cancer cells. *Oncol. Lett.* 5. doi:10.3892/ol.2012.1042.


Cyclotide
Agricultural Applications
- Insecticial
- Molluscidal
- Nematocidal
- Antimicrobial
Cyclotide
Pharmaceutical Applications
- Antiviral
- Brain tonic
- Anti-diabetic
- Immunomodulatory
- Wound healing
- Antioxidant
- Anti-cancer
Butterfly-pea, Shankhpushpi,
Aparajita, Gokarna, etc
(Clitoria ternata L.)

2

Transcriptomic Profiling of Cyclotide Genes and Biosynthetic Enzymes

Neha V. Kalmankar[1,2], Radhika Venkatesan[1], Padmanabhan Balaram[1,3] and Ramanathan Sowdhamini[1*]

Manuscript 1

Scientific Reports, 10, 12658 (2020).

https://doi.org/10.1038/s41598-020-69452-7

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore 560065, Karnataka, India.

[2]The University of Trans-Disciplinary Health Sciences and Technology (TDU), #74/2, Jarakabande Kaval, Post Attur, Via Yelahanka, Bangalore 560064, Karnataka, India

[3]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, Karnataka, India

*Correspondence:

Prof. Ramanathan Sowdhamini

mini@ncbs.res.in

Manuscript 1

Transcriptomic profiling of the medicinal plant *Clitoria ternatea*: Identification of potential genes in cyclotide biosynthesis

Neha V. Kalmankar, Radhika Venkatesan, Padmanabhan Balaram and Ramanathan Sowdhamini*

2.1 SUMMARY

Cyclotides are gene-encoded, ribosomally synthesized, cyclic cystine knotted peptides produced in several plants as their defense strategy. In this study, we profiled *Clitoria ternatea* (butterfly pea) plant using transcriptomic techniques to identify the cyclotide transcripts. Using RNA-Seq data and *de novo* transcriptome assembly, we identified 71 precursor genes of cyclotides. 51 of these display unique cyclotide domains and 26 of them represent novel cyclotides. RNA was purified from four different tissues of the plant – pod, stem, leaf and flower, compiled and compared across tissues. Additionally, we modelled the cyclotide biosynthesis in the cell by identifying the enzymes implicated in cyclotide biosynthesis such as the asparaginyl endopeptidase (AEP), protein-disulphide isomerases (PDI), ER-oxidoreductin (Ero1) and peptidylprolyl *cis-trans* isomerases (PPIase). Collectively, this work provides a comprehensive transcriptomic profiling data on cyclotides and their biosynthetic enzymes in *C. ternatea*.

2.2 RELATED INFORMATION

The published manuscript attached here refers to - Kalmankar NV, Venkatesan R, Balaram P, Sowdhamini R. Transcriptomic profiling of the medicinal plant *Clitoria ternatea*: identification of potential genes in cyclotide biosynthesis. Sci Rep. 2020 Jul 29;10(1):12658. DOI: 10.1038/s41598-020-69452-7. PMID: 32728092; PMCID: PMC7391643. Please note: The page numbers indicated within the PDF is independent of rest of the thesis

SCIENTIFIC REPORTS
nature research

OPEN

Transcriptomic profiling of the medicinal plant *Clitoria ternatea*: identification of potential genes in cyclotide biosynthesis

Neha V. Kalmankar[1,2], Radhika Venkatesan[1,3], Padmanabhan Balaram[1,4] & Ramanathan Sowdhamini[1]✉

Clitoria ternatea **a perennial climber of the Fabaceae family, is well known for its agricultural and medical applications. It is also currently the only known member of the Fabaceae family that produces abundant amounts of the ultra-stable macrocyclic peptides, cyclotides, across all tissues. Cyclotides are a class of gene-encoded, disulphide-rich, macrocyclic peptides (26–37 residues) acting as defensive metabolites in several plant species. Previous transcriptomic studies have demonstrated the genetic origin of cyclotides from the Fabaceae plant family to be embedded in the albumin-1 genes, unlike its counterparts in other plant families. However, the complete mechanism of its biosynthesis and the repertoire of enzymes involved in cyclotide folding and processing remains to be understood. In this study, using RNA-Seq data and de novo transcriptome assembly of *Clitoria ternatea*, we have identified 71 precursor genes of cyclotides. Out of 71 unique cyclotide precursor genes obtained, 51 sequences display unique cyclotide domains, of which 26 are novel cyclotide sequences, arising from four individual tissues. MALDI-TOF mass spectrometry analysis of fractions from different tissue extracts, coupled with precursor protein sequences obtained from transcriptomic data, established the cyclotide diversity in this plant species. Special focus in this study has also been on identifying possible enzymes responsible for proper folding and processing of cyclotides in the cell. Transcriptomic mining for oxidative folding enzymes such as protein-disulphide isomerases (PDI), ER oxidoreductin-1 (ERO1) and peptidylprolyl *cis-trans* isomerases (PPIases)/cyclophilins, and their levels of expression are also reported. In particular, it was observed that the CtPDI genes formed plant-specific clusters among PDI genes as compared to those from other plant species. Collectively, this work provides insights into the biogenesis of the medicinally important cyclotides and establishes the expression of certain key enzymes participating in peptide biosynthesis. Also, several novel cyclotide sequences are reported and precursor sequences are analysed in detail. In the absence of a published reference genome, a comprehensive transcriptomics approach was adopted to provide an overview of diverse properties and constituents of *C. ternatea*.**

Macrocyclics such as the cyclotides are a class of cyclic peptides (26–37 residues) containing three disulphide bonds. They are formed by cyclization of a gene encoded, linear precursor in specific plant species. In addition to the circular backbone, they form a cyclic cystine knot (CCK) arrangement formed by a conserved six cysteine framework (Cys I–Cys IV, Cys II–Cys V, Cys III–Cys VI)[1–3]. This knotted arrangement along with the circular backbone, renders these cyclotides exceptionally stable against enzymatic, chemical and thermal degradation, with their principal function thought to be in plant defense[4]. Thus, cyclotides are potential candidates for peptide-based drugs; either as scaffolds to stabilize susceptible peptide sequences or as drugs themselves[5]. To date, ~ 400 cyclotides have been sequenced and are largely from Rubiaceae, Violaceae, Fabaceae and others. A major challenge, however, is deciphering how their unique structural topology arises from linear precursors.

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore, Karnataka 560065, India. [2]The University of Trans-Disciplinary Health Sciences and Technology (TDU), #74/2, Jarakabande Kaval, Post Attur, Via Yelahanka, Bangalore, Karnataka 560064, India. [3]Present address: Department of Biological Sciences, Indian Institute of Science, Education and Research, Kolkata, Mohanpur, West Bengal 741246, India. [4]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka 560012, India. ✉email: mini@ncbs.res.in

Clitoria ternatea, a perennial climber of Fabaceae family, commonly known as 'Butterfly pea' or Shankhapushpi (in India), has been used in traditional Ayurvedic medicine as a memory enhancer, nootropic, antidepressant, anticonvulsant, blood platelet aggregation-inhibitor, tranquilizing and sedative agent. Its extracts possess a wide range of pharmacological activities including antimicrobial, antipyretic, anti-inflammatory, analgesic, antidiabetic, etc[6–10]. Initial phytochemical screenings of *C. ternatea* have shown that the biologically active ingredients are rich in peptides and depleted in alkaloids, flavonoids, saponins, lignans, etc[7,11]. As the primary mode of extraction was boiling, it is plausible to infer that the active agents are in fact heat-stable proteins[12]. Poth et al., had previously characterized 12 cyclotides (Cter A–L) from the seed extracts[13]. They also showed that Asn/Asp residues could occur at the cyclization site more commonly than previously understood. Later, they isolated 6 novel cyclotide sequences (Cter M–R) from leaf and flower tissues, and also determined the NMR structure of the chemically synthesized cyclotide Cter M[14]. Independently, Nguyen et al., screened for cyclotides in *C. ternatea* whole plant and reported 21 novel sequences and designated as cT1–21, of both cyclic and acyclic forms, and named them as 'cliotides'[12,15]. Presently, there are 85 cyclotide sequences reported from *C. ternatea*, out of which 74 have come from transcriptomic approaches[16], but given the abundance and potential variation of these macrocyclics, a larger number and diversity can be reasonably expected.

All the transcriptomic evidences from *C. ternatea* by various scientific groups have pointed out the occurrence of a single cyclotide domain in Fabaceae family. Unlike its counterparts Violaceae, Rubiaceae, Cucurbitaceae and Solanaceae, Faboideae possess pea-like albumin-1 genes[13,17–20]. These albumin-1 gene architecture contain a signal peptide, followed by a shorter albumin-1b chain, a short ~10 aa intervening linker region and ~50 aa long albumin-1a chain. Gilding et al., have demonstrated the evolution and co-option of the albumin-1b chain by the cyclotide domain. More importantly, they showed that the distribution of cyclotides in the plant tissues depends on the kind of herbivore attack, and suggested a combinatorial cyclotide defense[21].

Asparaginyl endopeptidase (AEP, EC 3.4.22.34) or vacuolar-processing enzyme (VPE) are a class of cysteine proteases that recognize and target Asn/Asp residues for peptide bond hydrolysis. It was established a while ago that AEPs not only perform cleavage, but also have the enzymatic capability of ligating linear peptides[3]. Plants that lacked AEP produced acyclic versions of cyclotides. However, Nguyen et al., in 2014 reported the first experimental evidence on the ligation property of AEP in *C. ternatea*. It was termed butelase-1 and shown to be highly efficient in intermolecular peptide bond formation, even on non-native-targets[22,23]. Prior and subsequent studies showed that mutations of Asn/Asp residue at the C-terminus of target proteins resulted in reduced cyclizing ability of the enzyme, implying the functional importance of an Asn/Asp residue in loop 6 of the cyclotide domain[24–27]. Since then several AEPs have been discovered in plants that make cyclotides, such as OaAEPs from *Oldenlandia affinis*, PxAEPs from *Petunia x hybrida* and HeAEP from *Hybanthus enneaspermus*[28,29]. Bioinformatics analysis on sequence and structural differences between the conventional AEPs (proteases) and dual functioning AEPs (proteases and ligases) needs to be explored.

The disulphide-bond formation is primarily catalysed by an ER-resident protein disulphide isomerase (PDI, EC 5.3.4.1), thereby acting as ER-chaperones[30–32]. PDIs (~50 kDa size) are oxidoreductase enzymes belonging to the thioredoxin superfamily. The classical PDI (hPDI) is composed of four thioredoxin domains (**a**, **b**, **b′**, **a′**) and an additional short α-helical **c** domain at the C-terminus. **a** and **a′** are the catalytically active domains characterized by the presence of a 'CX_1X_2C' motif, where X_1 and X_2 are mostly glycine and histidine residues[33]. The **b** and **b′** domains also bear the thioredoxin-like fold but are redox-inactive domains and do not bear the catalytic 'CX_1X_2C' motif. Also, C-terminal end of PDI sequences, bear an ER-retention signal composed of a highly conserved '[K/H/N][D/E]EL' motif[30,34,35]. While, PDIs have been identified in several plant systems, there is no conclusive evidence for the involvement of PDIs in cyclotide biogenesis, except for one study that shows increased yields of folded kalata B1 in the presence of a disulphide isomerase isolated from *O. affinis* (OaPDI)[36]. To the best of our knowledge, despite all the transcriptomic evidence reported on cyclotide-producing plants, no reports on the sequences of PDIs have been reported in such *planta*. Here, we obtained several *C. ternatea* PDI sequences by transcriptomic mining and the functional annotation has been established by mapping all the functionally important residues of plant PDIs.

PDIs oxidize cysteine residues in the substrate polypeptides and become reduced. One of the enzymes involved in PDI reoxidation and activation is the endoplasmic reticulum oxidoreductin-1 (ERO1). This in conjunction with flavin adenine dinucleotide (FAD) cofactor generates and transfers disulphide bonds to PDI and its family members[37–40]. A recent study reports the identification of a *Conus* ERO1 and its role in the oxidative folding of conotoxins, a class of disulphide-rich peptides from venom glands, similar to the cyclotides[41]. The oxidative folding capacity of both ERO1 and its homolog ERO2 was investigated in plants using transgenic *A. thaliana* only recently[42]. Exactly how the PDIs are oxidized by an oxidoreductase upon cyclotide folding remains to be investigated. To address this, we have identified an ERO1 for the first time from *C. ternatea* transcriptome and compared it with yeast, human and plant ERO1 sequences.

Cyclotides are structurally grouped into three subfamilies, Möbius, Bracelet and trypsin inhibitor[2]. The core CCK topology is conserved in all subfamilies, but the amino acid composition in the six loops varies substantially. Möbius subfamily differs from the Bracelet subfamily by the presence of a *cis*-Pro peptide bond in loop 5 (intervening residues between Cys V and Cys VI). The majority of the cyclotide sequences identified to date belong to the Bracelet subfamily[43], but the reason for this preference is unclear. Given the presence of a key *cis*-Pro peptide bond in loop 5, an enzyme mediated *cis-trans* isomerization reaction is likely to facilitate the protein folding. Peptidylprolyl *cis-trans* isomerases (PPIases, EC 5.2.1.8), also known as cyclophilins (CYP) and immunophilins, catalyze the *cis-trans* isomerization of proline residues and are ubiquitously distributed[44]. The role of PPIases in folding small disulphide-rich peptides is not well understood except on maurotoxin, a four disulphide-bridged scorpion toxin[45] and on conotoxins in venom glands of cone snails[46,47]. In the present study, we have performed a comprehensive mining of PPIases genes and identified CYP gene members and their expression patterns in the *C. ternatea* transcriptome.

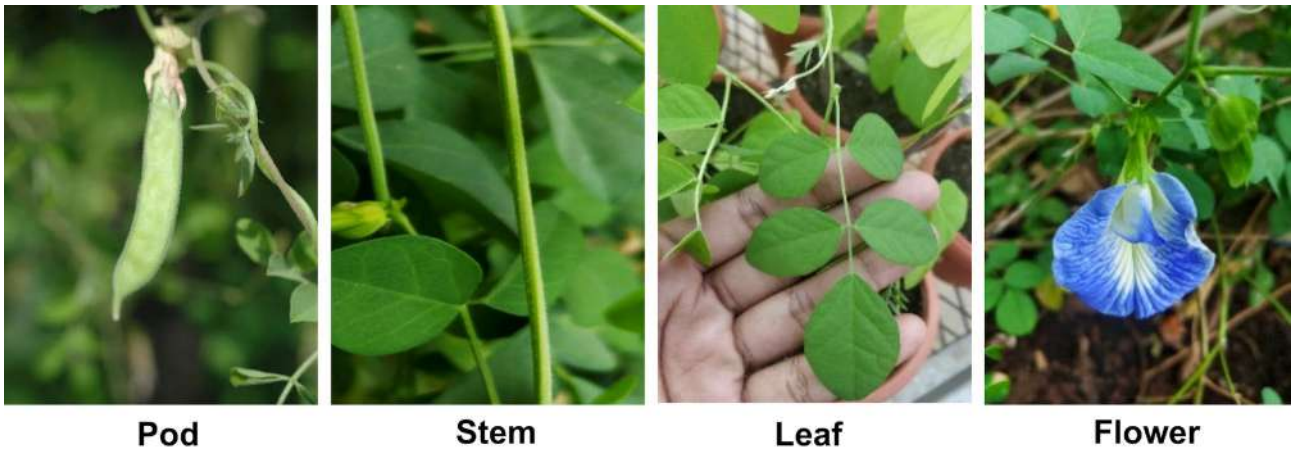


Figure 1. Morphology of *Clitoria ternatea* pods, stems, leaves and flowers.

Despite all the RNA-seq experiments on *C. ternatea*, a systematic functional annotation of a transcriptome assembly is yet to be performed. Here we present the de novo transcriptome assembly and annotation of *C. ternatea* plant (pod, stem, leaf and flower organs; Fig. 1), and report the presence of 26 new cyclotides that do not show high sequence similarity to those observed in previous studies. We have specifically investigated the expression of enzyme-coding genes such as asparaginyl endopeptidase (AEP), protein disulphide isomerase (PDI), ER oxidoreductin-1 (ERO1) and peptidylprolyl *cis-trans* isomerase (PPIase), which are key enzymes directly or indirectly involved, providing a concerted effect in the proper oxidative folding of disulphide-rich polypeptides, such as the cyclotides, in the cell.

Materials and methods

RNA preparation and sequencing. For transcriptome sequencing, RNA was extracted from four tissues of *C. ternatea* grown the on campus of National Centre for Biological Sciences, Bangalore (13°04′13.8″N 77°34′49.9″E). Samples were obtained from leaf, stem, pod and flower tissues (Fig. 1). Root tissue could be included in the current study, as quality-RNA from the root could not be purified. RNA was extracted using an RNeasy Plant Mini Kit (Qiagen, Inc.) as per manufacturer's guidelines. A total of 16 libraries were obtained from two biological replicates per tissue with two technical replicates per biological sample. RNA integrity and concentration were evaluated on a Bionalyzer (Agilent, Inc.). Samples with RNA integrity number (RIN) ≥ 6.5 were submitted to AgriGenome Labs Pvt. Ltd. (Kochi, India) for library preparation and sequencing using Illumina HiSeq 2500 platform V4 chemistry, to yield 100-bp paired-end reads (Raw data summary is detailed in Supplementary Table S1).

Transcriptome assembly and validation. Before transcriptome assembly, the quality of reads was assessed using FASTQC (v0.11.8) (www.bioinformatics.babraham.ac.uk/projects/fastqc/). A total of 293 million reads were obtained from 16 libraries. The raw reads were trimmed, to remove any low quality sequences or adapter contamination, using Trimmomatic v0.39[48] (parameters: *LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36)* and the trimmed reads was further used for assembly. The trimmed reads were de novo assembled using TRINITY assembler v2.6.6[49] (parameters: –seqType fq—normalize_by_read_set—left reads_1.fastq—right reads_2.fastq). A total of 278,991 transcripts were obtained with a mean contig of 1,441.89 bp and an N50 of 2,437 bp. We identified the candidate coding regions within the de novo assembly using TransDecoder v5.5.0 (https://transdecoder.github.io). First step in TransDecoder is TransDecoder.LongOrfs, which selects long open reading frame (ORF) per transcript (parameters: ORF in all 6 reading frames, minimum ORF amino acid length-m 50). The second step, TransDecoder.Predict predicts the likely coding regions. A total of 387,057 candidate genes were obtained. The transcriptome assembly was assessed for its completeness by using BUSCO v3 (Benchmarking Universal Single-Copy Orthologs)[50] on three datasets: Embryophyta_odb10, Eudicotyledons_odb10 and Viridiplantae_odb10.

Functional annotation and orthology analysis. Gene Ontology (GO) annotation was performed using OrthoVenn with an E-value cut-off of 10^{-10}.[51] Orthologous groups of *C. ternatea* encoded proteins with four other Fabaceae plant species *i.e. Glycine max*, *Medicago truncatula*, *Phaseolus vulgaris*, *Vigna unguiculata* and model plant *Arabidopsis thaliana* were obtained using OrthoVenn with an E-value cut-off of 10^{-10}.

Cyclotide precursor search. Protein and nucleotide sequences of published cyclotides were extracted from Cybase and UniProt[52,53]. These were used as query sequences to search our *C. ternatea* transcriptome through various BLAST methods (v.2.7.1)[54,55]. In parallel, functional motifs of cyclotide, corresponding to two familes (PS51052, PS60008 (Bracelet) and PS60009 (Möbius)) were queried using PROSITE[56,57] against our transcriptome database for pattern matching. The combined BLAST and PROSITE hits were manually inspected for

authenticity based on sequence conservation with known cyclotide sequences and redundancy was removed. The protein sequences were aligned through Clustal Omega (Supplementary Data 1)[58] and visualized using Unipro UGENE software[59,60]. The motif discovery algorithm (MEME Suite v5.1.0) was utilized for analysis of conserved motifs in cyclotide precursor sequences[61].

Identification of biosynthetic enzymes. The *C. ternatea* transcriptome was screened for several processing proteins AEP, PDI, ERO1 and PPIases that are known or likely to participate in the biosynthesis of cyclotides. Sequences of interest for each classical member of these enzyme/protein families, extracted from Uniprot, were used as queries to mine homologous sequences in the transcriptome using BLASTp (seq. cov. cut-off = ≥70%; seq. ID cut-off = ≥30%)[62]. The resulting hits were manually inspected for redundancy removal and identification. Identification involved deletion of signal peptides predicted by SignalP-5.0[63] and mapping of functionally important residues (FIR) onto a multiple sequence alignment of the enzymes. The protein sequences were aligned through Clustal Omega (Supplementary Data 1)[58] and visualized using BoxShade (https://sourceforge.net/projects/boxshade/).

Phylogenetic analysis. Neighbor-joining method was used to construct and understand phylogenetic relationships between the cyclotide transcript obtained from our assembly and all the known peptides reported from *C. ternatea*. The Maximum Likelihood method was used to construct phylogeny trees for the implicated biosynthetic enzyme transcripts from *C. ternatea* with orthologs from various species under Viridiplantae. Sequences of transcripts of interest were aligned with all the previously canonical examples using Clustal Omega and manually curated[58]. Phylogenetic construction of the final alignments for various queries discussed in this study were performed using MEGA version X[64] with the Neighbor-Joining/Maximum Likelihood method and nodal supports were calculated with 10,000 bootstrap iterations. The phylogenetic trees (bootstrap consensus) were visualized and imaged using FigTree v1.4.4[65] or TreeDyn 198.3[66]. All the protein sequences obtained in the current transcriptome and from other species are recorded in Supplementary Data 1.

Transcript quantification & differential expression analysis. RSEM was used to analyse and quantify the variation in gene expression across the four tissues[67]. The transcript abundance and Transcripts Per Million (TPM) was calculated by mapping raw reads to the assembled transcriptome. Clustering of TPMs of cyclotide genes was performed using Pearson correlation coefficient method with complete linkage and visualized with ClustVis webtool (https://biit.cs.ut.ee/clustvis)[68]. The TPM values for all the genes of interest and across the four tissues are provided in Supplementary Table S2.

Peptide extraction. Fresh plant parts (leaves, stems, flowers & pods) from *C. ternatea* were collected and oven dried at 70 °C. The peptide extraction was performed from literature protocols[12–15,69]. Briefly, plant material was extracted with 1:1 DCM: MeOH (v/v) with overnight stirring. The extract was partitioned with 1:1 mixture of DCM: (MeOH: ddH_2O, 2:3) using liquid–liquid phase extraction. MeOH/ddH_2O phase was further fractionated on RP-C18 column (0%, 30%, 50%, 70% and 100% gradient elution using ACN/ddH_2O). Initial screening of cyclotide-like masses in each elute was performed using MALDI-TOF mass-spectrometry following the protocol mentioned in Section 2.10 of Materials and Methods. Elutes 50%, 70% and 100% showed cyclotide-like masses (~3,000 Da). These elutes were pooled and freeze-dried for further analysis and will be referred to as "crude extract".

HPLC fractionation. The crude extract was analysed via LC–ESI–MS on a Maxis Impact Q-TOF mass spectrometer (Bruker Daltonics, Germany) interfaced to an Agilent 1,260 HPLC system. Samples were run on a reverse phase Agilent Poroshell C18 column (2.7 mm, 4.6 × 150 mm) using a linear gradient of 5%–95% ACN/ddH_2O, with 0.1% formic acid, and flow-rate of 0.2 mL/min. Data acquisition and analysis were performed using Bruker Data Analysis v4.1 software. The ESI–MS data obtained for *C. ternatea* crude extracts is not included in this study. HPLC fractionation of the crude extracts was performed using a Shimadzu Prominence series equipped with a binary pump, autosampler, PDA detector, and fraction collector. Chromatography was performed using a semi-preparative Phenomenex Proteo C18 column (250 × 10 mm, 10 µm, 110 Å) at flow-rate of 3 mL/min. A linear gradient of 1% min^{-1} of 0–95% buffer B (100% ACN, 0.1% trifluoroacetic acid) was applied, and the eluents were monitored at 254, and 280 nm. Late-eluting peaks were separated into five fractions (A–E) for each plant tissue, lyophilized for further analysis and will be referred to as "purified fraction".

Cyclotide analysis by MALDI-TOF mass spectrometry. Each of the five purified fractions (A–E), per tissue, was subjected to MADLI-TOF mass-spectrometry (UltraFlex Bruker Daltonics) analysis with 50 Hz pulsed nitrogen laser (λ = 337 nm), in positive ion reflectron mode. The samples were prepared by mixing an equal amount of purified fractions (0.5 µL) with matrices α-cyano-4-hydroxycinnamic acid (CCA) or 2,5-dihydroxy benzoic acid (DHB) prepared in 1:1 mixture of H_2O-ACN with 0.1% triflouroacetic acid, spotted on a stainless-steel target plate and air dried. The data were analysed using Flex Analysis v3.3.74 software. Identification of cyclotides was performed by matching the observed monoisotopic masses and calculated molecular weights of cyclotide transcript sequences.

Results and discussion

Clitoria ternatea transcriptome assembly, validation and annotation. A total of 281.8 million quality reads were assembled into a complete de novo assembly (details of assembly is provided in Supplementary Table S3). Quality of the assembly was assessed by examining the RNA-Seq read representation in the transcriptome by mapping reads back to the assembly using Bowtie2[70]. ~95% of the read fragments mapped as proper pairs and concordantly aligned 1 or more times to the transcriptome. Benchmarking Universal Single-Copy Orthologs (BUSCO) revealed the presence of 95.5% of complete BUSCOs out of the 1,375 orthologues searched from Embryophyta dataset, 94.5% of the 2,121 orthologues searched from Eudicotyledons dataset and 97.7% of the 430 orthologues searched from Viridiplantae dataset.

Gene orthology and functional annotation of _Clitoria ternatea_ transcriptome. Gene orthology analysis was performed using the predicted proteome of *C. ternatea* against proteomes of closely related species such as *Glycine max*, *Medicago truncatula*, *Phaseolus vulgaris*, *Vigna unguiculata* and *A. thaliana* using OrthoVenn tool. 10,778 orthogroups are common to all these species (Supplementary Fig. S1A). The plants form 57,821 orthologous clusters (at least contain two species) and 572 single-copy gene clusters. *C. ternatea* shares 753 and 494 orthogroups with closely related species *G. max* and *M. truncatula*, respectively. Gene Ontology (GO) annotation categories were assigned for *C. ternatea* and shown in Supplementary Fig. S1B. 24,950 orthogroups are specific to *C. ternatea* as seen in Supplementary Fig. S1A, and GO term enrichment analysis of these species-specific orthogroups show top GO-term hits with respect to stress-related genes, defense-response genes and serine-type endopeptidase activity, apart from the usual cell-processes genes (Supplementary Table S4).

Cyclotide identification in the transcriptome assembly. To mine cyclotide-coding genes in the transcriptome, mature cyclotides reported from *C. ternatea,* deposited in Cybase, were used as queries and a total of 3,753 redundant hits was obtained. The hits were filtered to remove redundancy and resulted in a final set of 71 unique cyclotide precursor genes. Out of 71 unique cyclotide precursor genes obtained, 51 displays unique cyclotide domain (redundant hits encode the same cyclotide but different albumin and/or ER-signal region), of which 26 are novel cyclotide domains that have not been reported earlier (Supplementary Table S5). These data expands our understanding of the cyclotide repertoire found in *C. ternatea* plant. RNA-seq and proteomic experiments from various groups have facilitated cataloguing of at least 85 cyclotide sequences produced in *C. ternatea* and independently we have identified additional new sequences. A comparison of the numbers reported by other groups and our group is depicted in Supplementary Fig. S2[15,16,21]. Figure 2 shows the multiple sequence alignment of all the 71 cyclotide precursor genes mined from the *C. ternatea* transcriptome. The figure highlights the ER signal peptide, immediately followed by the cyclotide mature domain, linked by a 10-residue linker and co-occurring with the albumin-1a chain, and a short ~10 residue C-terminal propeptide tail. Characterization of 'cliotide' genes from various transcriptomic approaches suggest a unique architecture in Fabaceae, that differs greatly from Violaceae and Rubiaceae cyclotide genes[15,20,21,69,71,72]. The cyclotide gene arrangement in *C. ternatea* is very similar to that of albumin gene of many legume species, wherein a signal peptide is followed by an albumin-1b domain, a linker region, and an albumin-1a domain. Similarities in arrangement and conservation of residues between albumin-1b and cliotide sequences suggests the albumin-1b domain could have been adapted into a cyclotide domain[21]. Out of 71 cyclotide precursor transcripts, 42 of them displayed nearly complete ER signal peptide region. 53 out of 71 hits displayed nearly complete albumin-1a domain. Figure 2 shows that the cyclotide mature peptide is cleaved off from the ER signal peptide at the highly conserved Thr-Glu-Ala (TEA) motif, most probably by a signal peptidase-1 enzyme[12]. Nine of the 71 precursor transcripts were partial sequences, lacking either 5′ end or 3′ end, or both.

Interestingly, several sequences display novel residue stretches at the N-terminal side of the cyclotide domain (Fig. 3). Furthermore, the starting site was not necessarily a glycine residue and included residues such as valine, serine, alanine and aspartic acid, contrary to some of the early reports on cyclotides. However, an asparagine residue at the C-terminal processing site was highly conserved (Fig. 3), suggesting that a non-Gly residue can possibly undergo a transpeptidation reaction with the conserved Asn residue. Three sequences lacked the C-terminal asparagine, implying that these could form acyclic products. Cyclotides with Möbius topology were seen in 11 cases, whereas Bracelet topology was seen in 27 cases, and Hybrid topology[73,74], i.e. bearing loop 2 and 3 similar to the Möbius and but lacking the *cis*-Pro in loop 5, were seen in 8 cases. One sequence contained an additional cysteine residue in the intervening loop 1 of the cyclotide domain (ctr28192_c2_g6_i2) and two sequences displayed atypical cysteine frameworks (ctr28495_c0_g5_i3 and ctr28841_c2_g3_i1) and these were termed as "unusual" cyclotide transcripts (Fig. 3).Interestingly, one the most abundant cyclotides known to be produced in *C. ternatea*, Cter M, and the only representative with a known structure[13], was obtained as a truncated sequence in our transcriptome assembly. The transcripts, ctr28495_c0_g7_i1 and ctr28495_c0_g7_i3, lacks the ER-signal peptide region and the starting glycine residue of the cyclotide domain.

Out of the 51 unique cyclotide domain containing transcripts, five were truncated sequences. Some of the challenges of a de novo assembly are misassembly, sequencing errors and redundancy problems that couldlead to truncation of transcripts. We have explored whether the partial cyclotide transcripts could be further extended by mapping them to the raw reads. We were able to extend the incomplete transcripts to at least contain the complete cyclotide domain by manually searching through the raw reads. The results for these partial cyclotide transcripts (ctr28192_c2_g6_i1, ctr28841_c1_g1_i4, ctr28841_c1_g4_i1, ctr28841_c2_g4_i2 and ctr29746_g1_g3_i3) are provided in the Supplementary Data 2. However, as these extension of boundaries at the 5′ or 3′ ends are merely predicted information, the sequences have been considered originally as partial sequences in this study.

Figure 4 represents the phylogenetic relationship between cyclotide precursor proteins from our transcriptome and those reported earlier. Cluster A represents 16 transcripts, 13 of which are classical *i.e.* bearing the

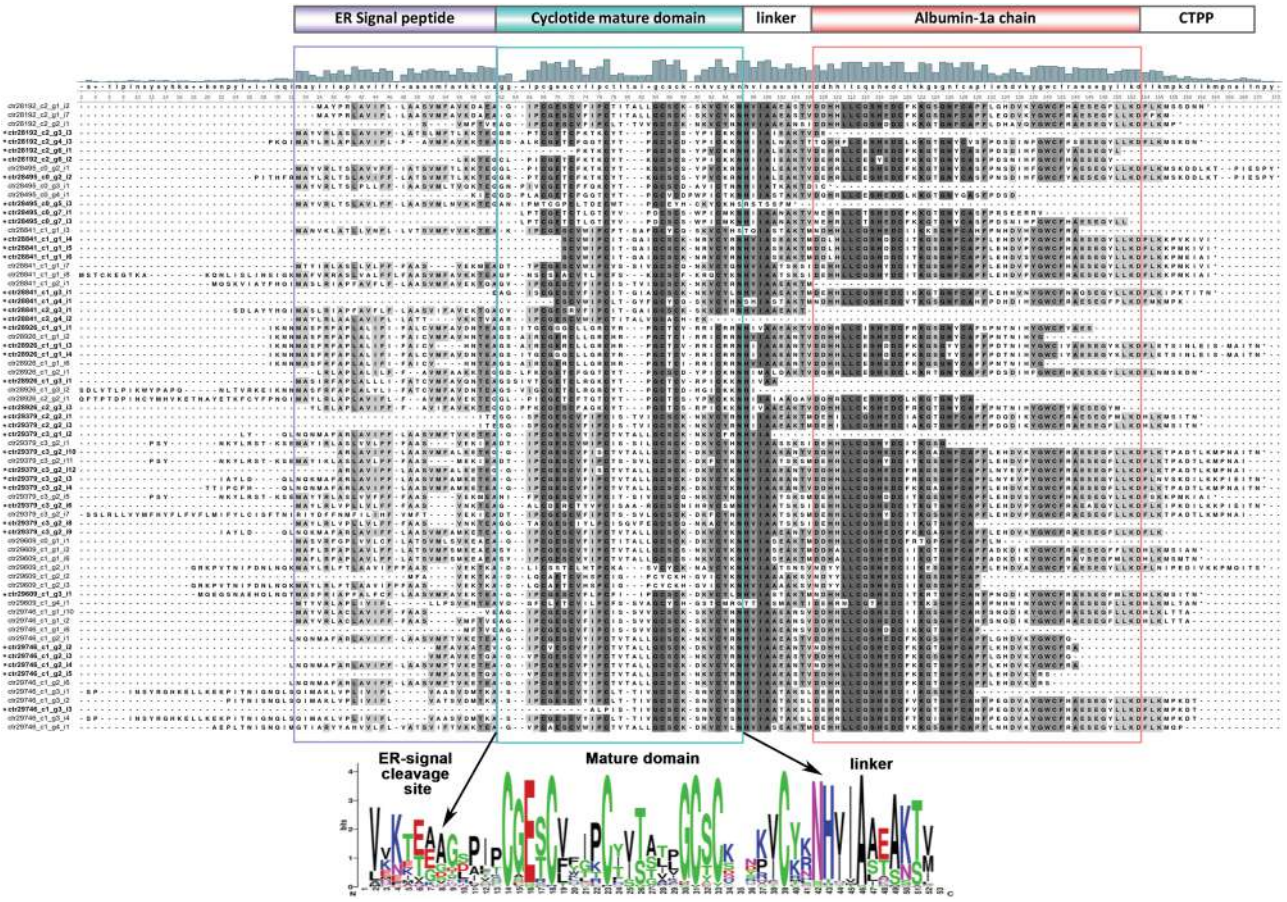


Figure 2. Multiple sequence alignment of 71 cyclotide precursor proteins obtained from the de novo transcriptome assembly of *C. ternatea*. Novel cyclotide sequences are highlighted in bold and an asterisk '*', alongside the transcript ID. ER-signal peptide highlighted in purple, cyclotide mature domain highlighted in cyan and albumin-1a chain highlighted in red. Sequence logo of the cyclotide mature domain, ER-signal cleavage site and linker region, generated using MEME Suite[61], is presented at the bottom of the figure. Key residues such as the cyclization site containing conserved Asn/Asp and membrane-binding conserved Glu residue is highlighted in red with arrows.

typical ER-signal peptide, immediately followed by the mature cyclotide domain, a linker region and albumin-1a chain. The remaining three precursors are truncated and possess atypical N-terminal motif in the cyclotide mature domain (but end with an asparagine at the C-terminus). Transcript ctr28495_c0_g5_i3 from cluster A bears least similarity with the conventional 'cliotide' sequences, displaying unusual cysteine framework and lacks the required asparagine at the C-terminus. Notably, cluster A represents primarily a Bracelet family of 'cliotides', having longer loop 3 and lacking the proline residue in loop 5. Transcripts ctr29609_c1_g2_i1, ctr29609_c1_g2_i2, ctr29609_c1_g2_i3 and ctr28841_c1_g1_i8 are an exception to this rule, displaying shorter loop 3. Cluster B highlighted in Fig. 4 represents 31 members, of which 18 carry the ER-signal peptide and the remaining 13 transcripts are partial. Cluster B also displays high sequence similarity with known Bracelet family members of cyclotides, having a glycine at the N-terminus and a conserved asparagine at the C-terminus and classical sequence architecture. Cluster C contains 20 members, 13 of which have the ER-signal peptide followed by cyclotide domain. The remaining 7 transcripts are partial sequences. Most sequences in this cluster contain the conserved glycine at the N-terminus and all members bear the crucial asparagine at the C-terminus. However, transcript ctr28192_c2_g6_i2 contains an additional cysteine residue in the cyclotide domain at the N-terminus apart from the conventional six conserved cysteines. Interestingly, cluster C represents the Möbius family of cyclotides, displaying shorter sequence lengths, with 65% of the hits bearing the *cis*-Pro residue in loop 5 of the cyclotide mature domain. Out of the 71 precursor protein sequences, 4 transcripts do not cluster with any of the above and remain as outliers. Transcript ctr29609_c1_g4_i1 sequence shows high dissimilarity with most cyclotide precursor sequences, showing unusual stretches in the ER-signal region, cyclotide domain and linker region. Despite the presence of six conserved cysteine framework, it lacks the C-terminal asparagine residue required for cyclization step. The sequence is reported previously and named as 'Cter acyc1' precursor[21]. Occurrence of uncommon precursor sequences hint at the divergent evolution of cyclotides from the ancestral albumin gene. Transcript ctr29609_c1_g3_i1 clusters and shares 67% sequence identity with cliotide T28 (also known as Cter 16) precursor sequences[15,21]. Transcript ctr29746_c1_g4_i1 clusters with Cter B precursor sequence[14],

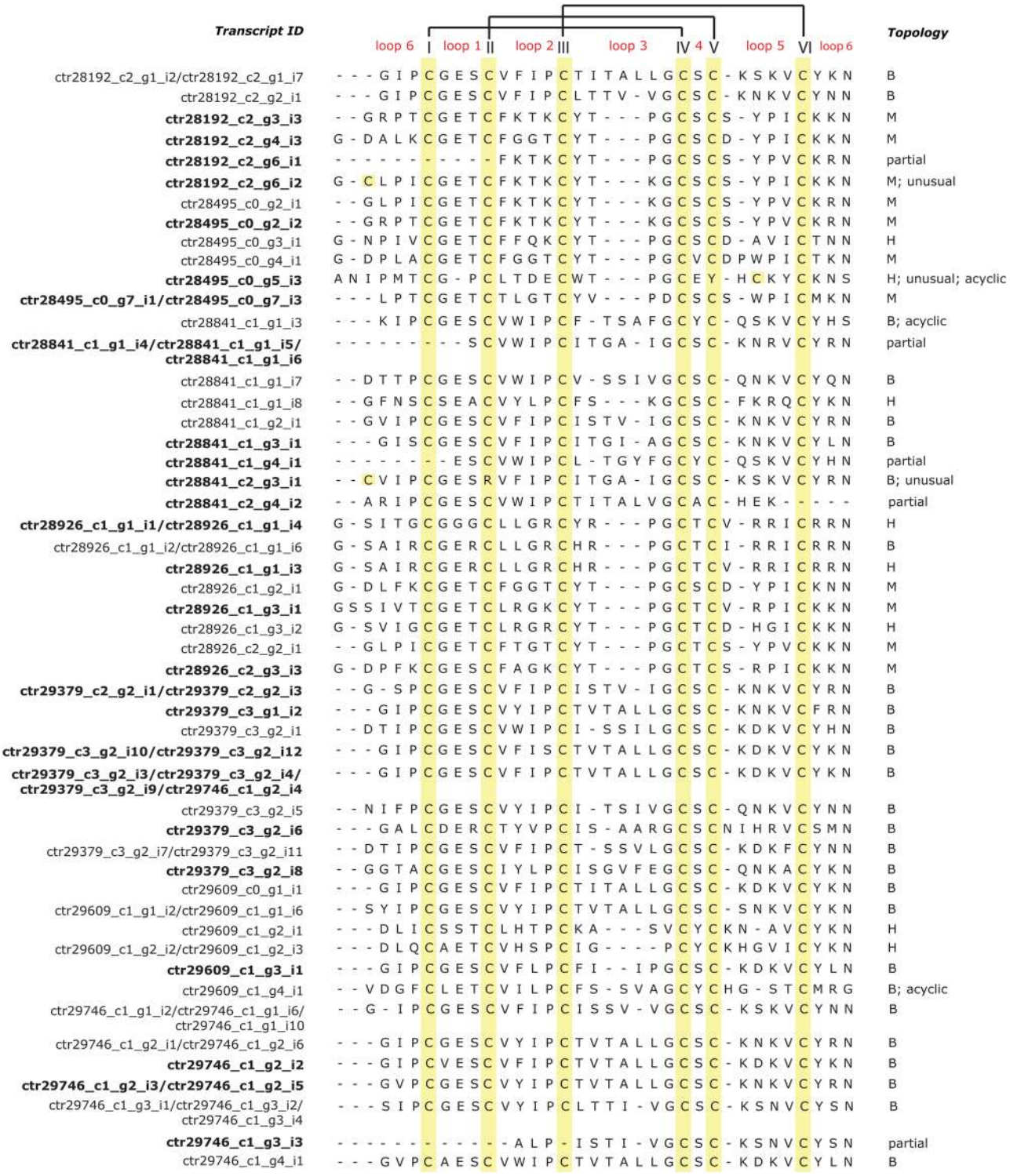


Figure 3. Sequence alignment of 51 unique cyclotide domains observed in *C. ternatea* transcriptome assembly. The transcript IDs of precursor genes corresponding to a unique cyclotide domain, are mentioned. Novel cyclotide domains are highlighted in bold. Cyclotide loop numbers are indicated above the alignment and the conserved cysteine residues are highlighted in yellow boxes. Based on sequence signatures of known cyclotides, the domains are classified as M-Möbius, B-Bracelet or H-Hybrid topology.

sharing 98% sequence identity. The MEME logos of motifs predicted for the above three phylogenetic clusters are presented in Supplementary Fig. S3.

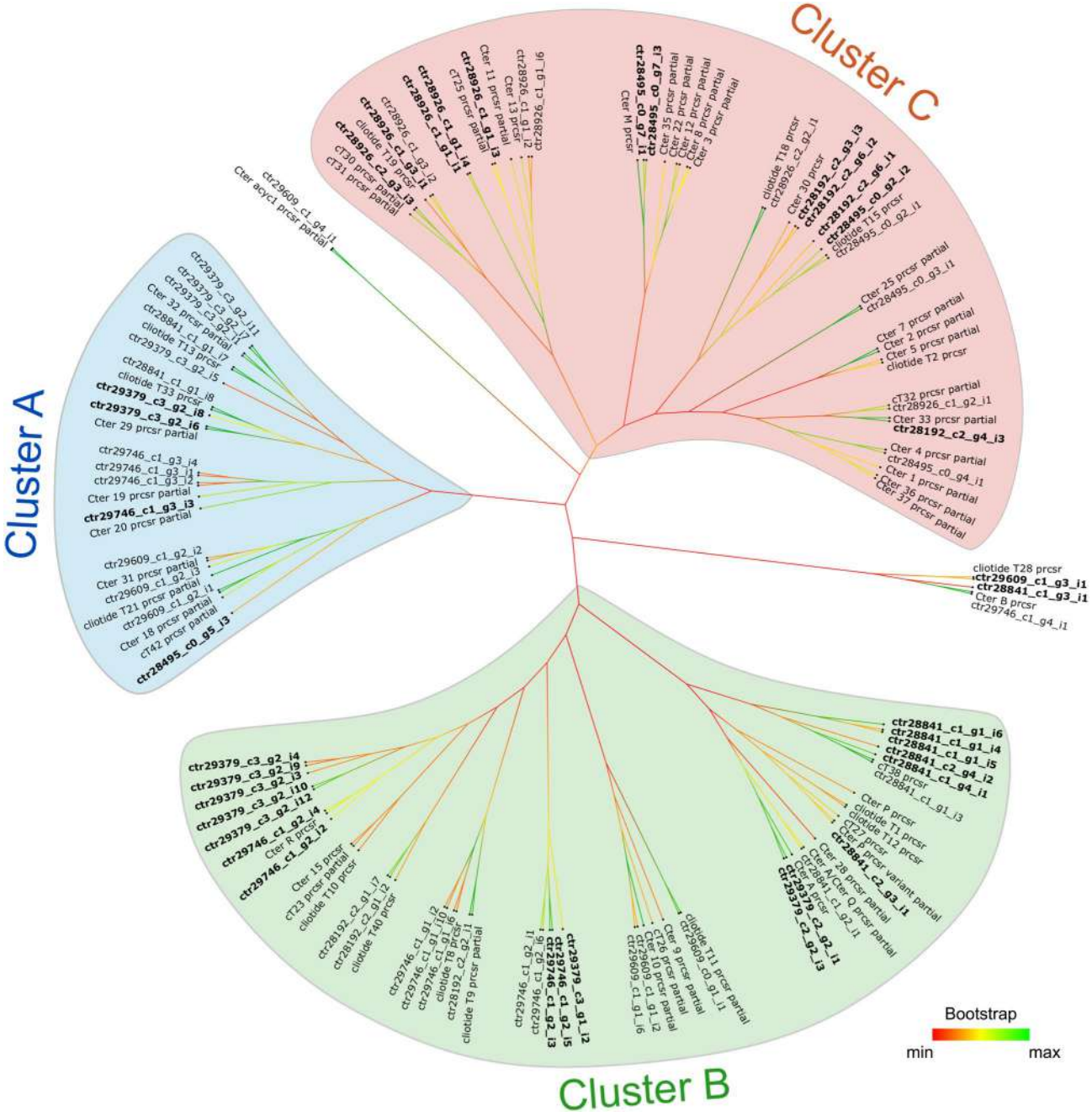


Figure 4. Phylogenetic tree of cyclotide precursor proteins from *Clitoria ternatea* (our transcriptome hits with all previously reported sequences from *Clitoria ternatea*). The tree contains 130 precursor sequences in total and was constructed using Neighbor-joining (NJ) method using MEGA version X software[64]. Three distinct cluster are highlighted as cluster A (blue), B (green) and C (red). Novel cyclotides obtained in the current study is highlighted in bold font. Bootstrap values for the unrooted tree is presented as a colour code ranging from minimum (red), mid-point (yellow) and maximum (green). "prcsr" = precursor.

Differential expression of cyclotides. Transcriptome mining undertaken in this study resulted in a diverse set of cyclotide genes with reasonably high sequence diversity, all ranging from typical topologies of Möbius and Bracelet, to Hybrids and acyclic gene products. To find relationships between expression of different cyclotide sequences and tissue localization, we performed a hierarchical clustering of TPM (Transcript per million) values of cyclotide genes using Pearson correlation coefficient method with complete linkage. Figure 5 illustrates three major transcript clusters across the tissues, in terms of expression profiles. Pods and stem displayed high expression of most cyclotide genes, while leaves and flowers showed relatively lesser expression values. Most of the members of the Bracelet topology of cyclotides show significantly high expression in pods and stems, whereas members of the Möbius and Hybrid topology cluster mostly in leaf tissue. Furthermore, genes classified as 'unusual' were mostly found in the leaf cluster, as seen in Fig. 5.

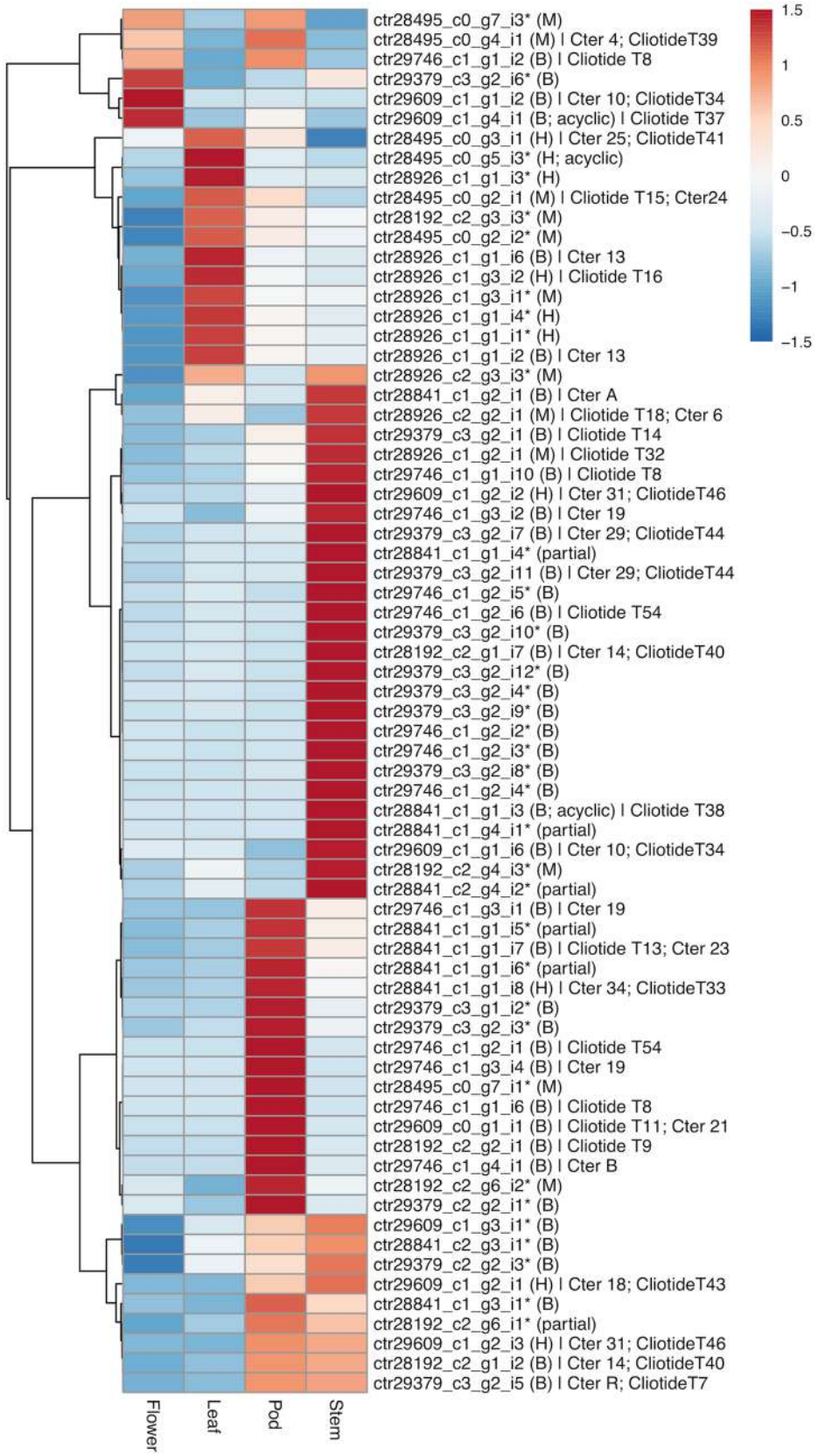


Figure 5. Heatmap of differential expression of cyclotide genes across four tissues of *C. ternatea* de novo transcriptome. The heatmap was generated using average TPM values and Pearson correlation coefficient clustering method with complete linkage on genes, using ClustVis tool[68]. Möbius (M) and Bracelet (B) topologies are highlighted in brackets and novel cyclotide sequences are highlighted with an asterisk '*', alongside the transcript ID. Positive (red) and negative (blue) values correspond to up-and down-regulated clusters, respectively.

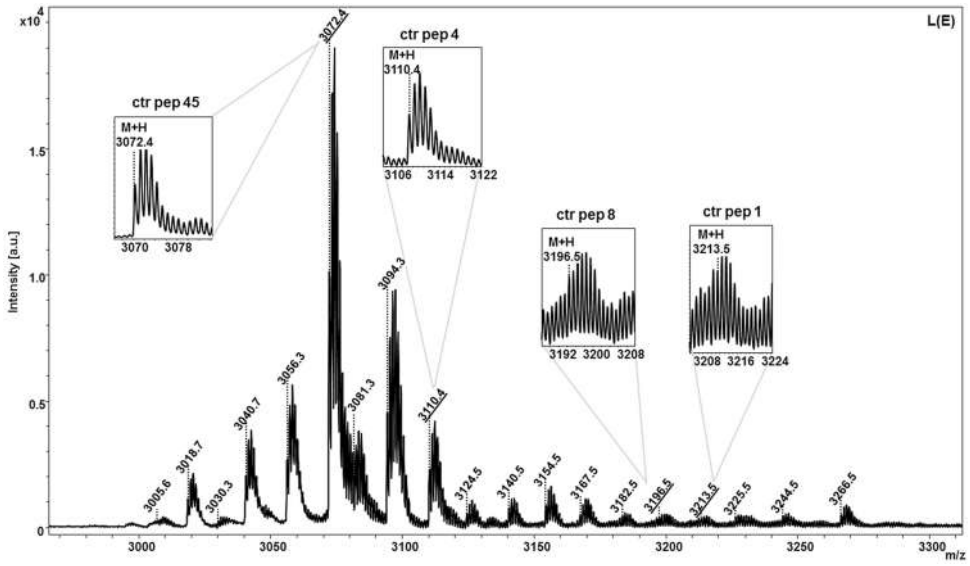


Figure 6. Representative MALDI-TOF mass-spectra of HPLC-purified fraction E from crude extract of leaves of *C. ternatea* (Refer to Supplementary Fig. S5). (Insets) The expanded isotopic multiplets for M + H masses that correspond to predicted cyclotide transcripts (peptide identifiers highlighted) are shown.

Novel transcripts ctr28926_c2_g3_i3 and ctr28495_c0_g7_i3, with Möbius topology, showed highest expression in the leaf, stem and pod, flower tissues respectively. However, novel transcripts ctr29379_c2_g2_i3 and ctr29609_c1_g3_i1, with Bracelet topology, displayed highest expression in pod and stem. Flower exhibited low cyclotide expression profile, with Cter 10 (ctr29609_c1_g1_i2) being the highest expressing gene in flower, clustering with the novel cyclotide ctr29379_c3_g2_i6. Unusual cyclotide gene ctr28495_c0_g5_i3 showed highest expression in leaf and acyclic 'Cter acyc1′ precursor (ctr29609_c1_g4_i1) showed high expression in flower. The predicted cyclotide gene expression was indeed tissue-specific, and they showed correspondence with untransformed TPM values (Supplementary Table S2). Additionally, Supplementary Table S2 shows that there is little variation between the biological replicates.

Such differential expression of cyclotide genes across different tissues, with varying levels of expression values indicates that cyclotides could be participating in varied plant responses towards biotic and/or abiotic stresses. Further, given the metabolic cost involved in production and maintenance of these peptides, it could be hypothesized that their distribution with the plant is controlled optimally. Whether differential distribution of cyclotides is useful to combat herbivore attack or environmental stresses remains to be studied.

Proteomic identification of cyclotides. We aimed to discover the presence of cyclotides and identify their molecular diversity in *C. ternatea* by using a combined approach of proteomics and transcriptomics analysis. We were interested to determine the presence of cyclotides at the protein level by using primarily MALDI-TOF mass spectrometry method. Mass detected analytical LC ESI–MS profile (ion chromatogram) of the crude extracts are shown in Supplementary Fig. S4. Masses corresponding to cyclotides (> 3,000 Da) were detected in the retention time range 25–55 min. Presence of cyclotides in all the four tissue crude extracts of *C. ternatea* were confirmed by ESI–MS (data not included in this study). Further fractionation of these crude extracts using semi-preparative HPLC was performed and illustrated in Supplementary Fig. S5. The purified fractions were subjected MALDI-TOF mass-spectrometry analysis for detection of > 3,000 Da masses. Cyclotides were distributed in fractions B-E (Supplementary Fig. S5). A representative MALDI-TOF spectra of fraction E from crude extract of leaves of *C. ternatea* is shown in Fig. 6. The MS analysis showed two significant peaks of about 3,071.4 m/z (M + H = 3,072.4) and 3,109.4 m/z (M + H = 3,110.4) and two low intensity peaks of about 3,195.5 m/z (M + H = 3,196.5 m/z) and 3,212.5 m/z (M + H = 3,213.5) corresponding to predicted peptides 'ctr pep 45', 'ctr pep 4', 'ctr pep 8' and 'ctr pep 1', respectively. Based on sequence analysis, ctr pep 45 and ctr pep corresponds to cliotide T8 and cliotideT40 respectively, and ctr pep 4 and ctr pep 8 are novel sequences. The expanded isotopic multiplets for M + H masses in all fractions from four tissues that correspond to predicted cyclotide transcripts, are shown in Supplementary Fig. S6. Identification of cyclotides at the proteome level was done by matching the observed monoisotopic masses and calculated molecular weights (Supplementary Table S6). Out of the 51 unique cyclotide sequences obtained at the transcriptomic level, 30 of them were detected at the protein level. Proteomic detection combined with transcriptomic analysis provided us useful information on the diversity of cyclotides produced in *C. ternatea.* The combined -omics approach resulted in more than half of the transcript sequences being predicted at protein level. Nevertheless, to uncover complete cyclotide diversity,novel structural templates and possible PTMs in a single plant specimen, a combinatorial procedure of transcriptome analysis, LC–MS deconvolution and tandem MS sequencing and transcriptome analysis is essential[75–77].

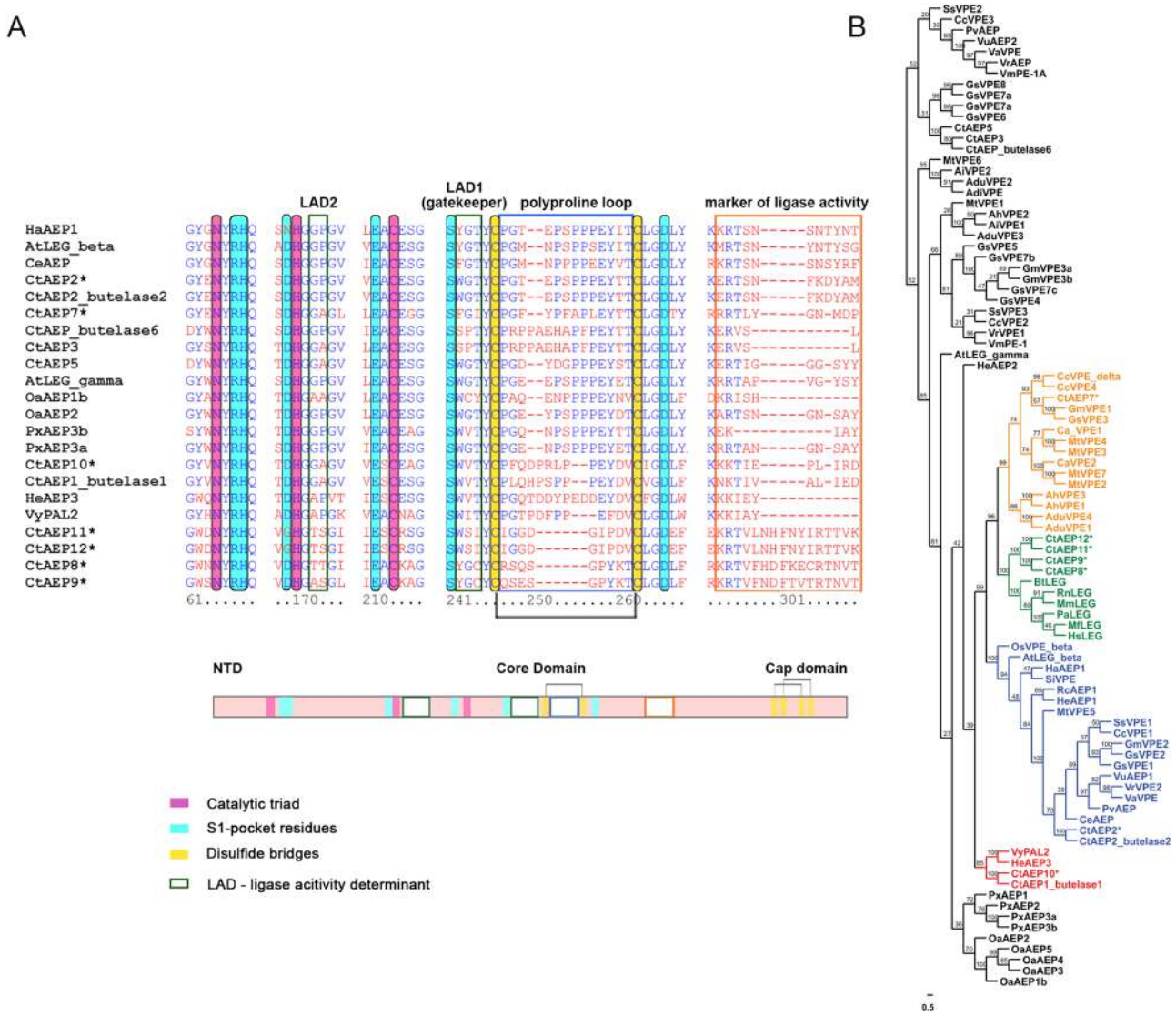


Figure 7. Asparaginyl endopeptidase sequences (AEPs) from *C. ternatea* transcriptome (**A**) Comparative sequence alignment of 7 nearly full-length AEPs from the current de novo transcriptome assembly and AEPs/legumains from *Oldenlandia affinis*, *Canavalia ensiformis*, *Petunia x hybrida*, *Arabidopsis thaliana*, *Helianthus annuus*, *Viola yedoensis* and *Clitoria ternatea* (butelases). The catalytic triad "Asn-His-Cys"is highlighted in pink. The conserved cysteines forming disulphide bridges are highlighted in yellow. S1-pocket residues is highlighted in blue. Ligase-activity determinant residues (LAD1 and LAD2), poly-proline loop (PPL) and marker of ligase activity (MAL) are boxed in green. (**B**) Phylogenetic tree of *Clitoria ternatea* AEP proteins with known mammalian legumains and plant AEP/VPEs. The tree contains 91 sequences in total and was constructed using Maximum likelihood (ML) method using MEGA version X software[64]. Bootstrap values at roots are indicated as percentages. Distinct clades are highlighted in different colour fonts, and outliers are coloured in black font. Novel cyclotide sequences are highlighted with an asterisk '*'.

Asparaginyl endopeptidase (AEP). We identified 14 AEP sequences in the transcriptome assembly. Seven out of the 14 hits were complete transcripts composed of the ER-signal and the mature active domain (Supplementary Table S7). Till date, five AEP sequences from *C. ternatea* (CtAEP1/butelase-1, CtAEP2/butelase-2, CtAEP3, CtAEP5 and butelase-6) have been well characterized[21,22]. AEP multiple sequence alignment and phylogeny was performed using AEP transcripts obtained in our transcriptome, Fabaceae and other plant AEP/VPE sequences, and mammalian legumains (Fig. 7). All the 7 hits contain the catalytic His and Cys residues, as highlighted in Fig. 7A. CtAEP10 is a partial sequence, lacking the N-terminal ER signal peptide and an incomplete 3′ end. Only one transcript (ctr29014_c2_g1_i2), bares 100% sequence identity with the known CtAEP2 sequence.

Recent studies indicate that two sites play an important role in differentiating a protease-type AEP and a ligase-type AEP. One is a "gatekeeper" position (GK) and the other is a "marker for ligase activity" (MLA) motif[29]. GK residue position (Cys247 in OaAEP1b), was found to be responsible for the ligation efficiency of OaAEP1b. This cysteine residue in some instances is replaced by a valine residue in the case of ligase-type AEPs (PxAEP3b)

and glycine residue in case of protease-type AEP (PxAEP3a). The ligase-type PxAEP3b was also shown to present a truncated MLA region and the protease version, PxAEP3a, had a longer and hydrophilic MLA region[29]. On the contrary, a recent study determined that GK and MLA sites alone are not responsible for ascribing ligase activity to AEPs. Using X-ray crystallography and site-directed mutagenesis, it was suggested that as the MLA region is distantly located from the catalytic core and it might not play a crucial role in activity[78]. They identified putative ligase-activity determinants, LAD1 (contains GK site) and LAD2 as the key sites for activity determination, apart from the remotely located MLA region. Specifically, they propose that the first position in LAD1 should be a bulky aromatic amino acid and the second position should be hydrophobic such as Val/Ile/Cys/Ala but not Gly. Figure 7A shows that all the seven transcripts have aromatic residues such as phenylalanine, tyrosine and tryptophan at the first position in LAD1. However, the second position is a glycine or a serine residue at the GK site (position 243 in the alignment) in CtAEP2, CtAEP7, CtAEP8, CtAEP9, CtAEP11 and CtAEP12. This suggests that these could be the protease-type AEPs. A single sequence, CtAEP10, displays a valine residue at the GK site, implying this sequence to be a ligase-type AEP. The investigators also proposed that a small residue such as Gly or Ala be present at the first position of LAD2, which is the case in our CtAEP10 sequence (Fig. 7A).

Figure 7B reveals the phylogenetic clustering of CtAEP8 and CtAEP9, CtAEP11 and CtAEP12 transcripts with mammalian AEP sequences (green clade). Interestingly, CtAEP7 clusters with pigeon pea's VPE4 and VPEδ (yellow clade). As these sequences also lack the GK Val/Ile/Cys/Ala at position 243 in the alignment (Fig. 7A), it is likely that these represent the protease-type AEP (Fig. 7B). Only two transcripts in the transcriptome are nested with known ligase-type AEP sequences clade. –CtAEP10 sequence co-clusters with CtAEP1/butelase1 and OaAEP1b (red clade), and CtAEP2 sequence co-clusters with CtAEP2/butelase2 (blue clade), as seen in Fig. 7B. TPM values of the AEP transcripts expressed across different tissues were calculated and compared with cyclotide transcripts (Supplementary Fig. S7). Most of the transcripts showed high expression in the flower based on the mean TPM value (Supplementary Table S2). However, CtAEP2 (butelase 2), CtAEP7 and CtAEP10, showed high expression in terms of TPM values in leaf, pod and stem tissues, respectively. Clustering of CtAEP10 (predicted ligase-type AEP) with several cyclotide transcripts, which get expressed highly in stem, provides evidence on the involvement of AEP in cyclotide cyclization.

Protein disulphide isomerase (PDI). PDI is a soluble protein in the ER and is involved in catalysing disulphide bond formation of substrate proteins. Furthermore, it is known to be involved in disulphide-bond formation of several seed storage proteins[79–81]. In butterfly pea, cyclotide gene is embedded within the albumin 1b gene, hence it is reasonable to assume the involvement of PDI for cyclotide folding. A variety of plant PDIs and PDI-like proteins have been identified through independent genomic approaches, resulting in a total of about 22 members in *A. thaliana*, 12 members from *Oryza sativa*, 32 from *Brassica rapa ssp. Pekinensis*, and 22 in *Zea mays* and 9 from *Triticum aestivum L*[82–85]. A novel PDI from the Rubiaceae plant *O. affinis* (OaPDI) was reported to be involved in the in vitro folding of knotted circular proteins[36]. In-depth analysis of PDI sequences in terrestrial plants have led to several distinct types of nomenclature[82,83,86,87]. The former study named *Arabidopsis* PDI isoforms as PDI1.1 to 1.6, PDI2.1 to 2.3 and PDI5.1 to 5.4[82]. More recent studies, specifically on soybean PDIs, a letter for various PDI classes and consecutive numbers for each isoform was used[88,89]. For example, classical PDIs were named PDI-L, and each subclass was distinctly numbered. Due to the ambiguity in nomenclature, we have not followed any of the above methods and instead named our *C. ternatea* PDI sequences numerically. We have found 34 PDI and PDI-like hits, of which 15 are nearly full-length sequences (exhibiting two TRX motifs) and 19 are partial/incomplete sequences. The list of full-length PDI sequences from *C. ternatea* are presented in Supplementary Table S8. Structural characteristics such as the active site residues, C-terminal signal sequence and the domain composition is also presented in Supplementary Table S8. The domain architecture adopted was found to be predominantly the classical type i.e. **a**–**b**–**b′**–**a′** arrangement of successive thioredoxin domains.

Figure 8A highlights the phylogenetic construction of known plant PDIs with CtPDI sequences from the current transcriptome. Four sequences, CtPDI3, CtPDI7, CtPDI11 and CtPDI12, are nested within PDIL1 subfamily, highlighted as clade 3 (pink). Sequences CtPDI14, CtPDI15, CtPDI13 and CtPDI8 are nested within PDIL5 subfamily, highlighted as clade 4 (green). CtPDI4 sequence is nested within PDIL4 subfamily, highlighted as clade 5 (yellow). CtPDI10 is a single sequence that is nested within PDIL3 and PDIL2 subfamilies, (clade 2, blue). Four sequences, CtPDI1, CtPDI5, CtPDI6 and CtPDI9, form a separate cluster (clade 1, red). CtPDI2 sequence is an outlier and does not cluster with any of the above PDIL subfamilies, suggesting it could be a species-specific foldase. Figure 8B shows the domain structure of 15 PDI sequences extracted from the current *C. ternatea* transcriptome. Domain boundaries were identified by sequence alignment to *Arabidopsis* PDI sequences and human PDI sequence. It also highlights the sequence conservation of the redox active sites (**a** or **a′** or **a°**) and the ER-retention signal. The two thioredoxin active site motifs (CXXC) in **a** and **a′** (or **a°**) domains are fully conserved. **a** domain in fourteen of the CtPDI sequences has the motif CGHC, while CtPDI10 sequence has motif CPRS. **a′** (or **a°**) domain in 13 CtPDI sequences has the motif CGHC, while two CtPDI sequences have CHFC and CINC motif (CtPDI2 and CtPDI10 respectively). Thirteen out of the 15 full length sequences show the C-terminal ER-retention signal 'K[D/E]EL' motif. CtPDI13 has a unique 'KDQI' motif at the C-terminal end, suggesting that it may get secreted from the endoplasmic reticulum, instead of being retained in the ER (Fig. 8B). CtPDI2 bears reasonable sequence similarity with AtPDIL2-1, AtPDIL2-2 and OaPDI sequence[36,82]; however, it lacks the N-terminal acidic domain that is characteristic of PDIL2 subfamily (Fig. 8B) and possesses 'HDEL' motif at the C-terminus. A distinguishing characteristic of OaPDI is that the two active sites contain Gln residue (instead of a Lys residue) as commonly seen in many PDI sequences in humans and in Arabidopsis[36]. Similarly, CtPDI2 sequence also seems to be different from several known PDIs. Firstly, it has a CHFCXA motif in the active site of a' domain and secondly, it bears a Gln residue in the first active site (**a** domain) and a Lys residue in the second active site (**a′** domain) at the X-position, which is an uncommon combination in canonical PDI

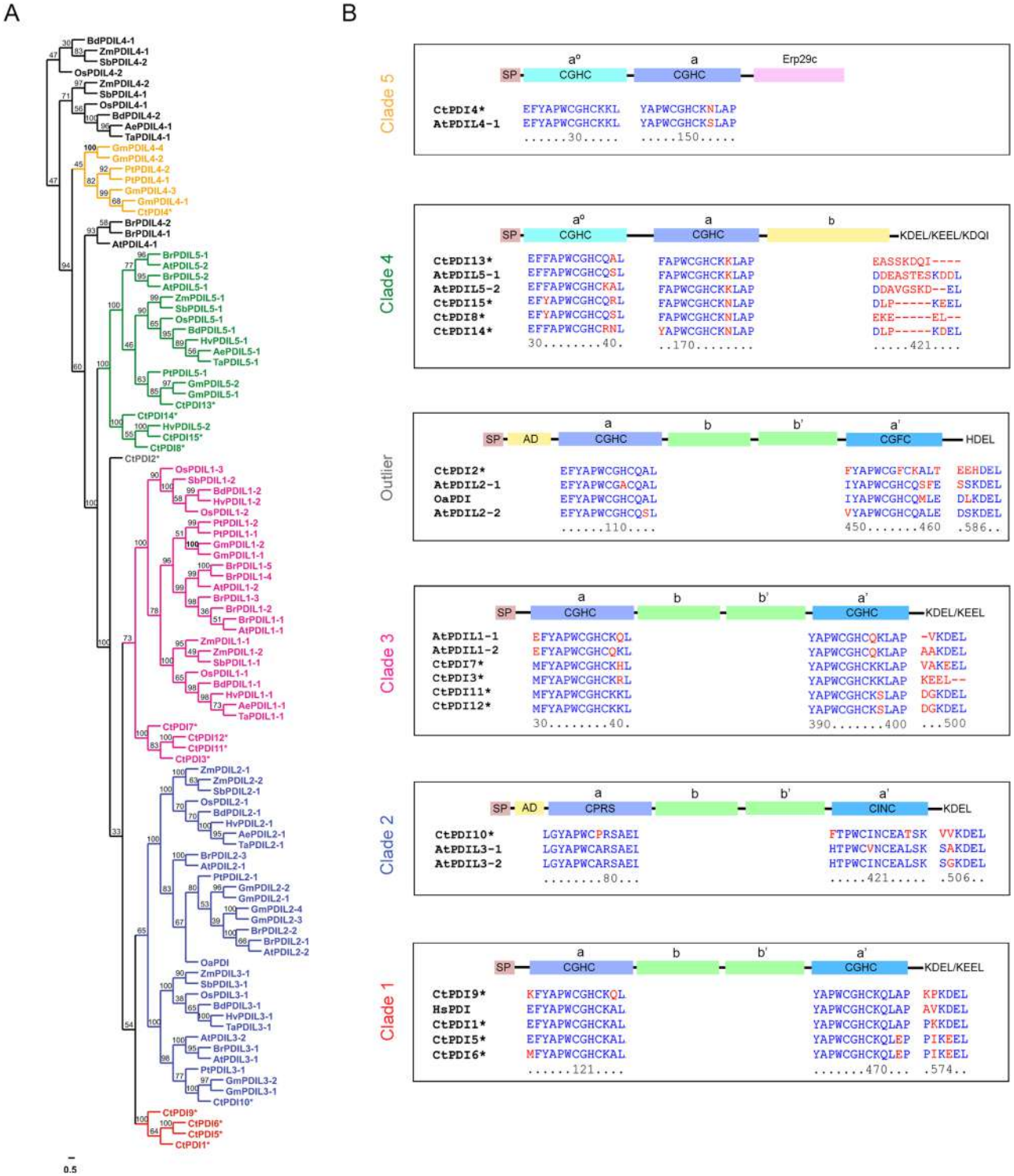
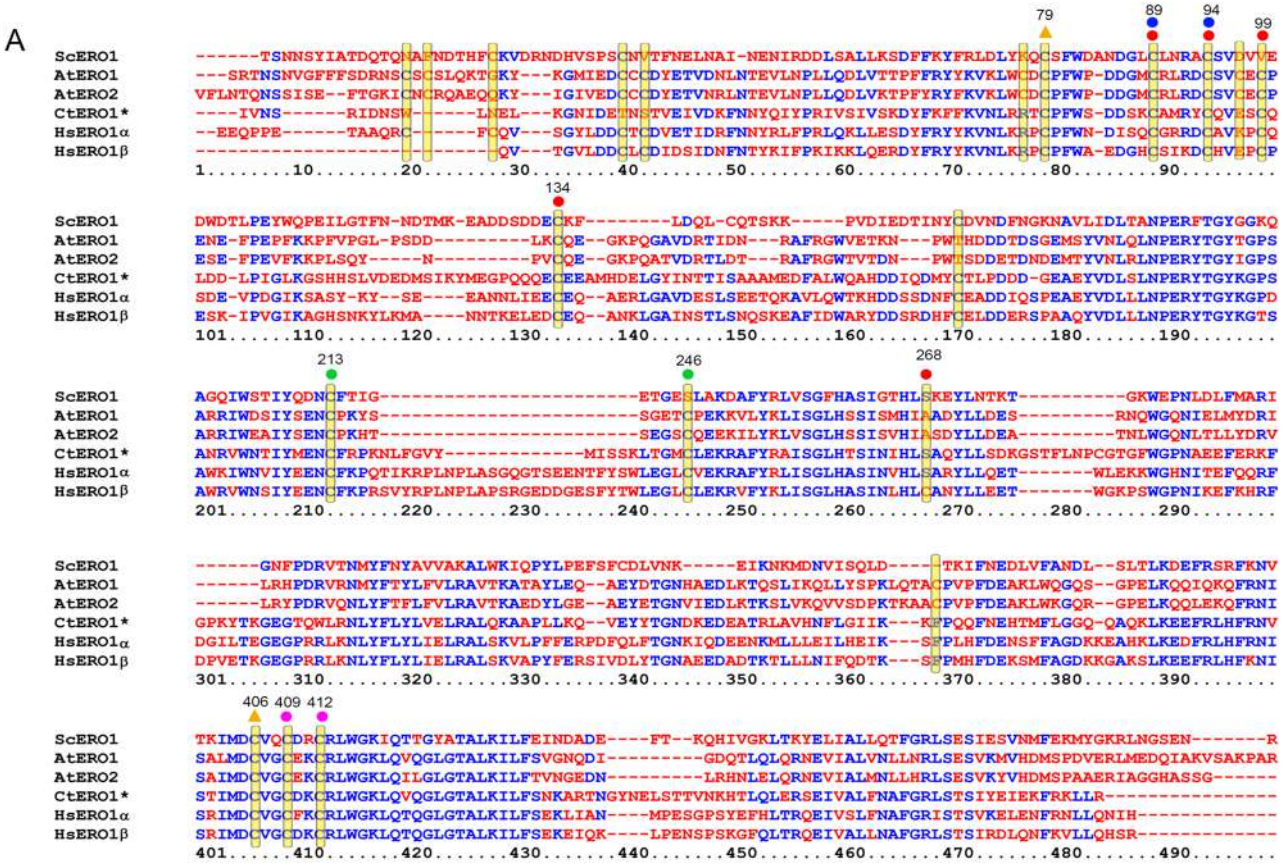
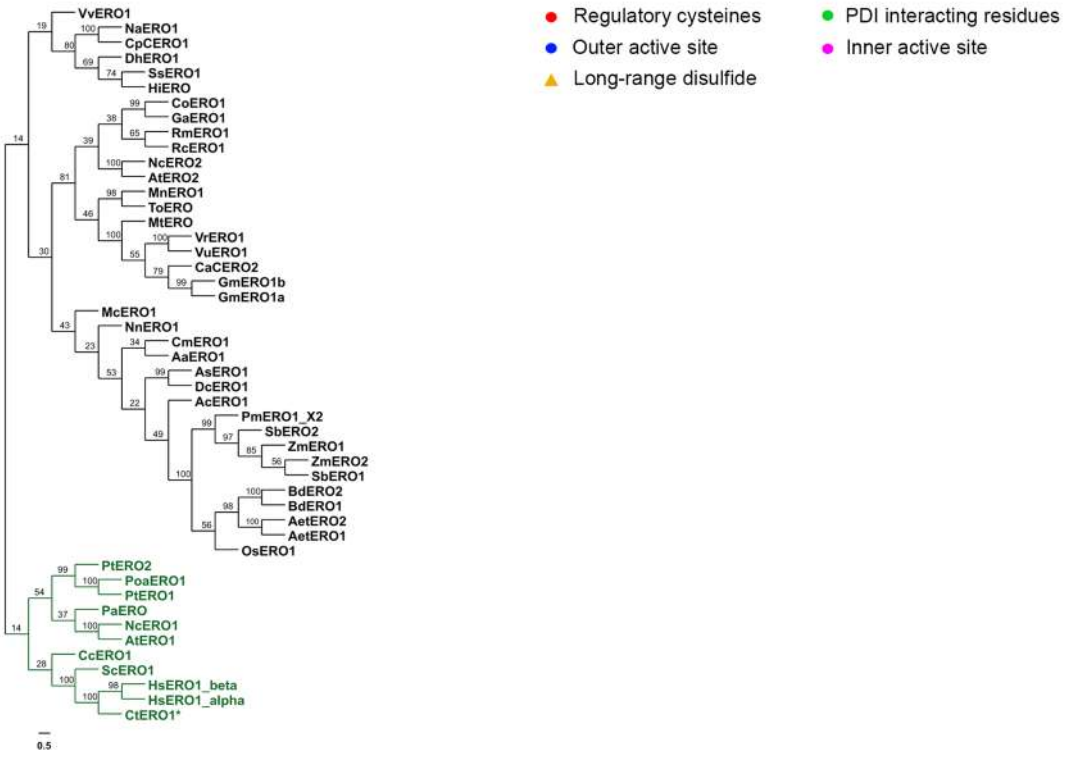


Figure 8. Protein disulphide isomerase sequences (PDIs) from *C. ternatea* transcriptome (**A**) Phylogenetic tree of *Clitoria ternatea* PDI hits with other plant PDIs. Distinct clades are highlighted in different colour fonts, and outliers are coloured in black font. The tree contains 104 sequences in total and was constructed using Neighbor-joining (NJ) method using MEGA version X software[64]. Bootstrap values at roots are indicated as percentages. (**B**) Domain architecture of CtPDI sequences aligned with canonical PDI sequences and categorized into different clades are displayed. The catalytic thioredoxin-like tetrapeptide "CXXC" and C-terminal ER-retention signal tetrapeptide of the 15 PDI sequences obtained from the de novo transcriptome assembly of *C. ternatea* along are reported. The putative signal peptides (SP), the a and b type domains, the N-terminal acidic domain (AD) and the Erp29c domain are highlighted.


A
ScERO1
AtERO1
AtERO2
CtERO1*
HsERO1α
HsERO1β
79
89
94
99
134
213
246
268
406 409 412


B
VvERO1
NaERO1
CpCERO1
DhERO1
SsERO1
HiERO
CoERO1
GaERO1
RmERO1
RcERO1
NcERO2
AtERO2
MnERO1
ToERO
MtERO
VrERO1
VuERO1
CaCERO2
GmERO1b
GmERO1a
McERO1
NnERO1
CmERO1
AaERO1
AsERO1
DcERO1
AcERO1
PmERO1_X2
SbERO2
ZmERO1
ZmERO2
SbERO1
BdERO2
BdERO1
AetERO2
AetERO1
OsERO1
PtERO2
PoaERO1
PtERO1
PaERO
NcERO1
AtERO1
CcERO1
ScERO1
HsERO1_beta
HsERO1_alpha
CtERO1*
0.5
Regulatory cysteines
Outer active site
Long-range disulfide
PDI interacting residues
Inner active site

◂**Figure 9.** ER oxidoreductin-1 sequence from *C. ternatea* transcriptome (**A**) Comparative sequence alignment of the sole ERO1 protein sequence (highlighted with an asterisk "*") obtained from the *C. ternatea* transcriptome with yeast (ScERO1), human (HsERO1α and HsERO1β) and *Arabidopsis* (AtERO1 and AtERO2) ER oxidoreductin-1 sequences. Cysteines are highlighted in yellow bars, regulatory cysteines are highlighted with a red circle, long-range disulphide is highlighted with an orange triangle, PDI-interacting partners are highlighted with a green circle, outer active site is highlighted with a blue circle, inner active site is highlighted with a pink circle and non-catalytic cysteines are not marked with a circle or triangle. Numbering of residues is based on the alignment. (**B**) Phylogenetic tree of *Clitoria ternatea* ERO1 hits with yeast (ScERO1), human (HsERO1α and HsERO1β) and plant ERO1 homologs. Distinct clades are highlighted in different colour fonts, and outliers are coloured in black font. The tree contains 48 sequences in total and was constructed using Neighbor-joining (NJ) method using MEGA version X software[64]. Bootstrap values at roots are indicated as percentages.

sequences. A recent study on cone snail PDI family suggest that two distinct isoforms exist, canonical PDIs and conotoxin-specific PDIs (CsPDI)[90]. Supplementary Fig. S8 shows a multiple sequence alignment of PDIs from *C. ternatea* transcriptome, *Conus geographus* cone-specific PDI and *Conus geographus* canonical PDI. There is a high degree of sequence conservation between the butterfly pea PDI sequences and the cone snail PDIs, especially in the two thioredoxin domains (**a** and **a′**). However, two CtPDIs show significant differences in the + 2 position of the C-terminus and in both the CGHC active sites. CtPDI9 has CGHCKQ in both the catalytic domains. Such apparent differences in the active site motifs were observed in conotoxin-specific PDI (*C. geographus* csPDI) sequences[90]. This suggests that protein disulphide isomerases could have diversified in *C. ternatea* and in a correlated manner with the diversification of cyclotide sequences. Overall, Fig. 8 highlights the phylogenetic and domain analyses of CtPDI gene family. The amino acid conservation and co-clustering of PDI genes demonstrates that the PDI gene family is divergent in plants. Specifically, the phylogenetic analyses highlight that the PDIs from *C. ternatea* cluster independently to form phylogenetically distinct groups.

Hierarchical clustering analysis of the PDI expression levels, shows that out of the 15 sequences, 11 are highly expressed in the flower tissue (Supplementary Table S8). The expression level (TPM) of the fifteen genes of PDI family from *C. ternatea* are plotted along with the cyclotide transcripts across the four tissues (Supplementary Fig. S9). Consistent with the hypothesis of the role of PDI in disulphide isomerase activity of cyclotides, we observed high expression levels of few PDI genes correlated with cyclotide expression in each tissue. Especially CtPDI2, a predicted species-specific outlier in phylogeny, is up-regulated in the stem and clusters well with several cyclotide sequences. CtPDI4 and CtPDI10 also clusters with five cyclotide sequences (Cter 19, Cter 31, cliotide T8, cliotide T32 and cliotide T13) in the stem tissues. CtPDI13 shows high expression levels in pod and stem tissues, along with novel cyclotide transcripts ctr28841_c1_g3_i1 and ctr28192_c2_g6_i1.

Endoplasmic reticulum oxidoreductin-1 (ERO1). ERO1 catalyzes the formation of disulphide bonds in its substrate PDI in conjunction with FAD cofactor[91]. ERO1 proteins are highly conserved across kingdoms of life. Disulphide bond formation and proper folding of cyclotides in plants takes place in the ER, where several cyclotides are expressed and properly folded. Proper folding of disulphide-rich peptides requires the concerted efforts of multiple enzymes, as this is critical to their biological activity in the cellular environment[46,47,92]. Here, we have identified a single ERO1 protein sequence (474 amino acid) in the *C. ternatea* transcriptome, which has not been attempted before. Sequence of CtERO1 transcript (ctr203_c0_g1_i1) is described in Fig. 9A, with all the active site residues highlighted on the multiple sequence alignment. Figure 9A highlights the high amino acid sequence homology of CtERO1 to human ERO1α 48% sequence identity) and ERO1β (~ 50% sequence identity), as opposed to *A. thaliana* ERO sequences (< 39% sequence identity). This is substantiated again in the phylogeny of homologs from plants, yeast and human, where CtERO1 clusters with human and yeast ERO1 homologs with high nodal support (green clade) (Fig. 9B). The putative outer active sites (Cys89-Cys94) and inner active sites (Cys409–Cys412) are fully conserved in ERO1s across yeast, plant and human ERO sequence. The putative regulatory cysteines Cys89, Cys94, Cys99 and Cys134 are fully conserved. However, position 268 in CtERO1 sequence is occupied by serine instead of a cysteine residue, characteristic of HsERO1β. The putative long-range disulphide bridges (Cys79–Cys406) and PDI interacting partners (Cys213 and Cys246) are also conserved across the yeast, plant and human ERO sequences (Fig. 9A). Similar to the expression levels of several CtPDI genes, CtERO1 is also upregulated in the flower tissue and substantiates the fact that ERO1 might be participating in PDI re-oxidation and activation (Supplementary Fig. S9).

Peptidylprolyl *cis-trans* isomerase (PPIase)/cyclophilin. To identify all members of the CYP gene family in *C. ternatea*, a BLASTp sequence search was performed against the transcriptome using *G. max* CYP1 sequence (UniProt entry Q8W171). This search identified a total of 70 unique CYP genes and named CtCYP1 to CtCYP70 (Supplementary Table S9). All the identified proteins were analysed for presence of CLD domain and additional domains using CDART, SMART and DoMosaics tool[93–95]. Of the 70 CYP genes, 53 encoded a protein with a single cyclophilin-like domain (CLD). 17 of the 70 genes contain cyclophilin-like domain (CLD) and additional domains, hence classified as multi-domain CYP genes (Supplementary Table S9). Figure 10 highlights a schematic representation of domain architectures of all CtCYPs sequences. CtCYP1, CtCYP34, CtCYP62, CtCYP63 and CtCYP64 comprises of an RNA recognition motif (RRM). CtCYP62 and CtCYP63 also contain an additional Zinc-finger (zf-CCHC) motif at the C-terminus end. CtCYP2, CtCYP23, CtCYP66 includes Tetratricopeptide-like repeats (TPR) at the C-terminus. CtCYP8 contains a Sialic acid binding adhesion domain (SabA adhesion) at the C-terminus end, which is not commonly known to exist in cyclophilin sequences. CtCYP9

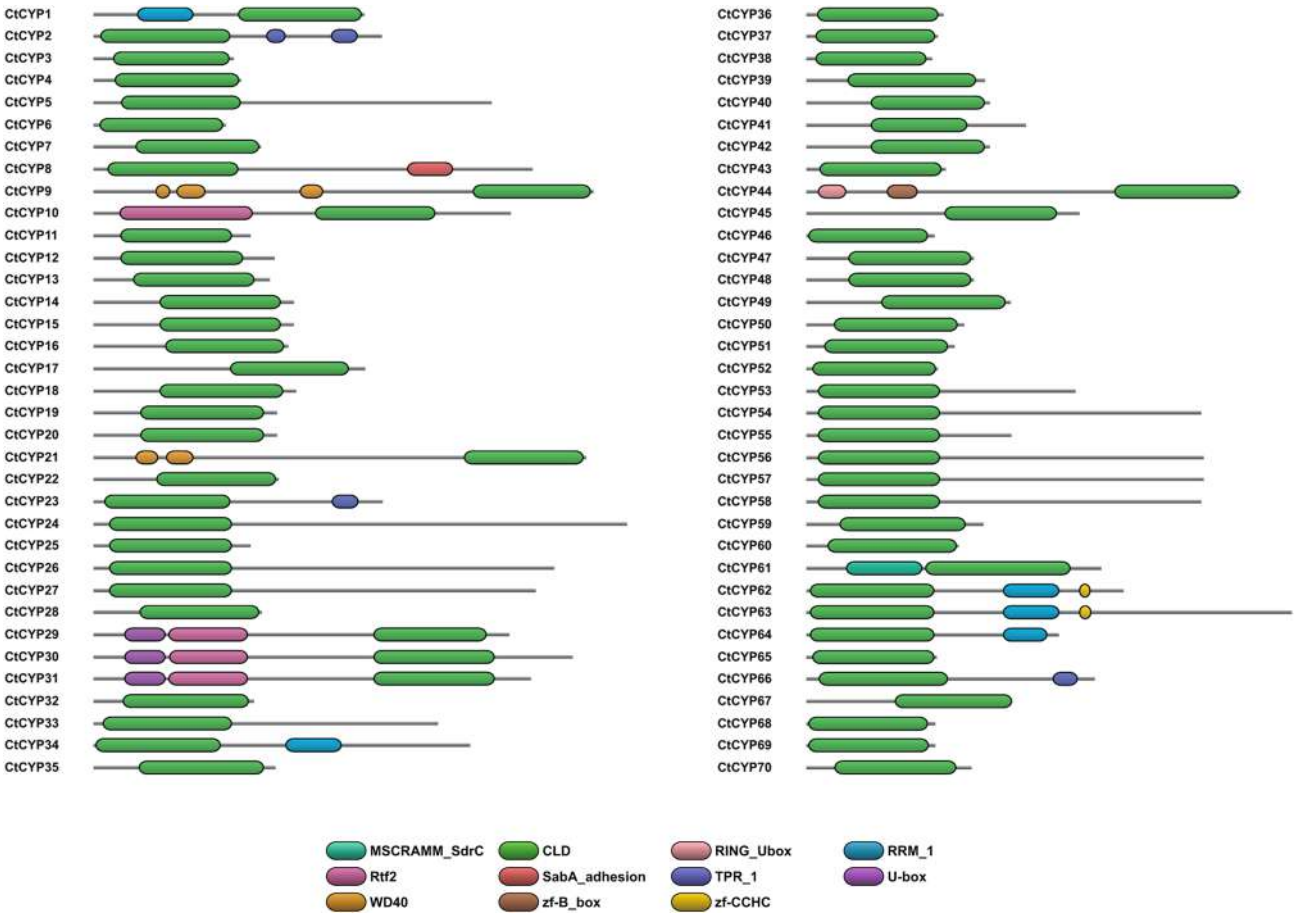


Figure 10. Schematic representation of single-domain and multi-domain peptidylprolyl *cis-trans* isomerase/cyclophilins in *C. ternatea* (70 sequences). The domain arrangement was created using DoMosaics tool[95]. The cyclophilin-like domain (CLD) is represented in green. *MSCRAMM_SdrC* microbial surface components recognizing adhesive matrix molecules-SdrC adhesion domain, *RING_Ubox* RING U-box domain, *RRM_1* RNA recognition motif 1, *Rtf2* Replication termination factor 2 RING-finger domain, *SabA_adhesion* Sialic acid binding adhesion domain, *TPR_1* Tetratricopeptide-like repeats, *U-box* U-box domain, *WD40* tryptophan-aspartate (WD) repeats, *zf-B_box* B-box zinc finger domain, *zf-CCHC* Zinc-finger motif.

and CtCYP21 comprise two/three tryptophan-aspartate (WD) repeats at the N-terminus. CtCYP10 includes a large Replication termination factor 2 (Rtf2) RING-finger domain at the N-terminus. CtCYP29, CtCYP30 and CtCYP31 contain Rtf2 and U-box domains at the N-terminus end of the protein. CtCYP44 includes a B-box zinc finger and RING U-box domains at the N-terminus. Lastly, CtCYP61 comprises of MSCRAMM (microbial surface components recognizing adhesive matrix molecules)-SdrC adhesion domains at the N-terminus, which is a novel arrangement seen in cyclophilin proteins. Subcellular localization was predicted using five different algorithms—DeepLoc1.0, YLoc, SherLoc2, CELLO v2.5 and MultiLoc2[96–100]. The results show that several CYPs are predicted to localise in the ER and there is a consensus among the different algorithms (Supplementary Table S9). Of the 70 CtCYPs, 19 have been predicted by at least one predictor to be targeted to the ER, therefore, likely to be involved in cyclotide biosynthesis. Hierarchical clustering analysis of the CYP expression levels show that these 19 CtCYPs are distributed among all the four tissues, clustering with differentially expressed cyclotide genes (Supplementary Fig. S10). TPM values for majority of PPIases in *C. ternatea* indicate high expression in the stem and flower tissue, and these predominantly are single domain CYP genes (Supplementary Table S9).

Conclusions

In the present study, we describe the de novo assembly and annotation of the *C. ternatea* transcriptome. We identified a total of 71 cyclotide precursor sequences and 26 of them describe new cyclotide sequences. In-depth analysis of the ER-signal regions and cyclotide mature domains, highlight the sequence diversity at the transcript and proteome level. We have also highlighted tissue-specific gene expression of cyclotides through hierarchical clustering method. The results suggest that cyclotide expression occurs in a tissue specific manner and the levels of expression vary greatly. Figure 11 highlights the cyclotide gene architecture and the possible biosynthetic pathway, displaying the concerted effect of various enzymes in the proper folding and processing of cyclotides. Specifically, the cyclization enzyme and the oxidative folding enzymes have been explored in this study. Macrocyclizing AEPs are essential for producing cyclic peptides that impart ultra-stability to already stable molecules. Such AEPs have a wide potential in peptide engineering and synthetic biology. However, recombinant

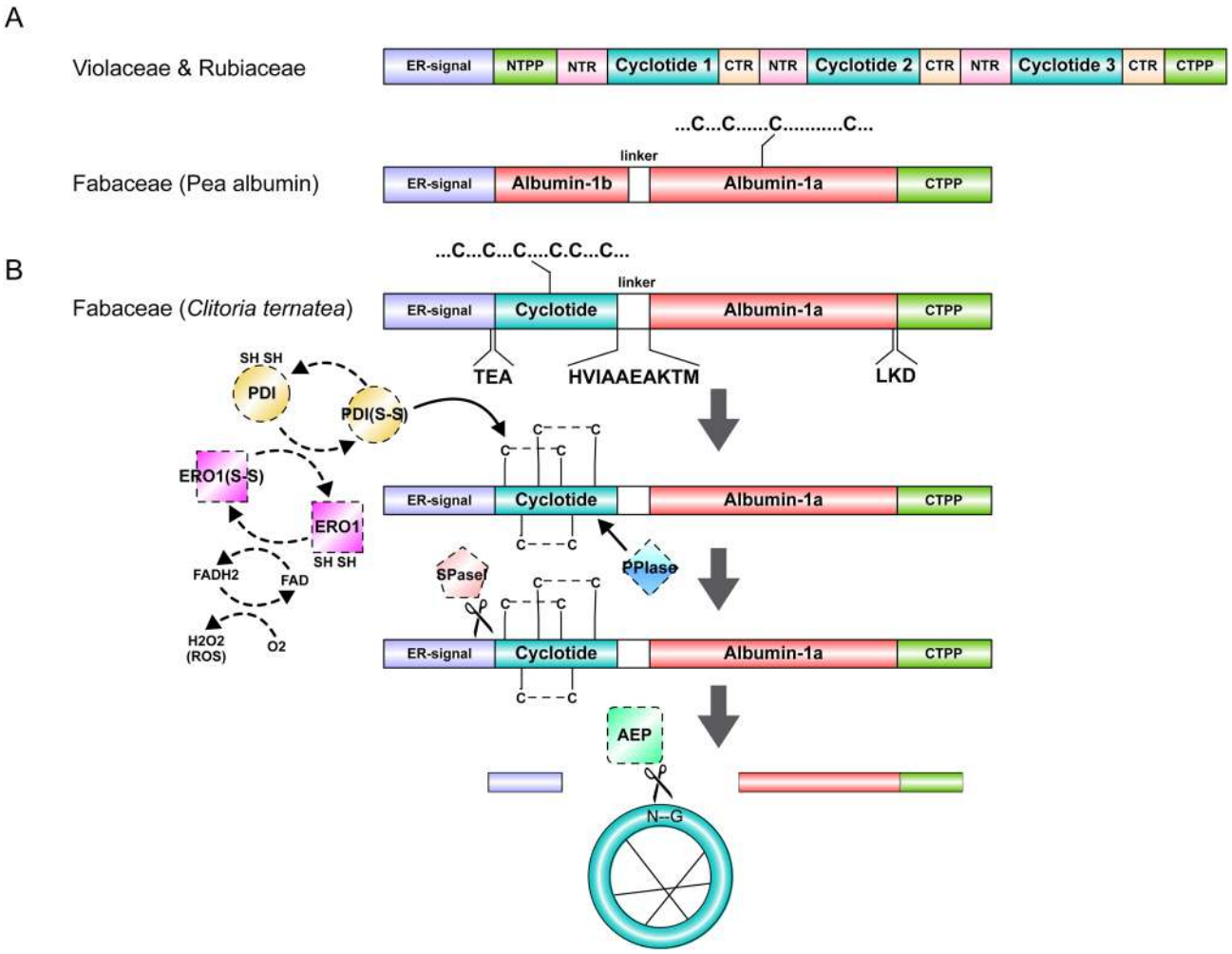


Figure 11. (**A**) Schematic view of the genetic arrangements of Violaceae and Rubiaceae cyclotide precursors and Fabaceae albumin precursor. (**B**) Overview of proposed biosynthetic pathway of cyclotides in *Clitoria ternatea* (Fabaceae). The concept figure was generated using Illustrator for Biological Sequences (IBS)[101]. *NTPP* N-terminal propeptide, *NTR* N-terminal repeat, *CTPP* C-terminal propeptide, *CTR* C-terminal repeat, *PDI* protein disulphide isomerase, *ERO1* ER oxidoreductin-1, *SPase1* signal peptidase-1, *PPIase* peptidylprolyl *cis-trans* isomerase, *AEP* asparaginyl endopeptidase.

expression of macrocyclizing AEPs is challenging. The current model of cyclotide production suggests formation of disulphide bonds before cyclization step, within the ER. For this PDIs are essential and this study is the first to provide transcriptomic evidence for diverse PDI isoforms in *C. ternatea*. Furthermore, re-oxidization of reduced PDIs has not been addressed in plants producing cyclotides. However, the current study permits identification of candidate enzymes such as the ER oxidoreductin. A single isoform of ERO1 could be retrieved from *C. ternatea* transcriptome. The main difference between the two structural families of cyclotides is a *cis*-proline in loop 5, observed only in the Möbius family. To catalyze this crucial reaction, PPIases are essential and no sequence search efforts in a cyclotide producing plants have been attempted earlier. This study is the first to report 70 isoforms of PPIases in *C. ternatea*. In the absence of a complete genome sequence of *C. ternatea*, comprehensive transcriptomics analyses reported in this study might be useful for further developing applications of this medicinally relevant plant.

Data availability

The raw reads generated for this study have been deposited in BioProject Accession: PRJNA573557 (https://www.ncbi.nlm.nih.gov). The nucleotide and protein precursors generated and/or analysed during the current study are available in GenBank under the accession numbers for *C. ternatea* (a) cyclotide precursor sequences: MT468661-MT468731, (b) asparaginyl endopeptidase sequences: MT468732- MT468738, (c) protein disulphide isomerase sequences: MT468739- MT468753, (d) endoplasmic reticulum oxidoreductin-1 sequence: MT468754 and (e) peptidyl prolyl *cis-trans* isomerase sequences: MT468755- MT468824, respectively.

Received: 11 February 2020; Accepted: 10 July 2020
Published online: 29 July 2020

References

1. Saether, O. *et al.* Elucidation of the primary and three-dimensional structure of the uterotonic polypeptide Kalata B1. *Biochemistry* **34**, 4147–4158 (1995).
2. Craik, D. J., Daly, N. L., Bond, T. & Waine, C. Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J. Mol. Biol.* **294**, 1327–1336 (1999).
3. Jennings, C., West, J., Waine, C., Craik, D. & Anderson, M. Biosynthesis and insecticidal properties of plant cyclotides: The cyclic knotted proteins from Oldenlandia affinis. *Proc. Natl. Acad. Sci. USA* **98**, 10614–10619 (2001).
4. Colgrave, M. L. & Craik, D. J. Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: The importance of the cyclic cystine knot. *Biochemistry* **43**, 5965–5975 (2004).
5. Craik, D. J. & Conibear, A. C. The chemistry of cyclotides. *J. Org. Chem.* **76**, 4805–4817 (2011).
6. Kulkarni, C., Pattanshetty, J. R. & Amruthraj, G. Effect of alcoholic extract of *Clitoria ternatea* Linn. on central nervous system in rodents. *Indian J. Exp. Biol.* **26**, 957–960 (1988).
7. Rai, K. S. *et al.* Clitoria ternatea root extract enhances acetylcholine content in rat hippocampus. *Fitoterapia* **73**, 685–689 (2002).
8. Jain, N. N. *et al.* Clitoria ternatea and the CNS. *Pharmacol. Biochem. Behav.* **75**, 529–536 (2003).
9. Mukherjee, P. K., Kumar, V., Kumar, N. S. & Heinrich, M. The Ayurvedic medicine *Clitoria ternatea*-from traditional use to scientific assessment. *J. Ethnopharmacol.* **120**, 291–301 (2008).
10. Daisy, P. & Rajathi, M. Hypoglycemic effects of *Clitoria ternatea* Linn. (Fabaceae) in alloxan-induced diabetes in rats. *Trop. J. Pharm. Res.* **8**, 393–398 (2009).
11. Rai, K. S., Murthy, K. D., Karanth, K. S. & Rao, M. S. *Clitoria ternatea* (Linn) root extract treatment during growth spurt period enhances learning and memory in rats. *Indian J. Physiol. Pharmacol.* **45**, 305–313 (2001).
12. Nguyen, G. K. T. *et al.* Discovery and characterization of novel cyclotides originated from chimeric precursors consisting of albumin-1 chain a and cyclotide domains in the fabaceae family. *J. Biol. Chem.* **286**, 24275–24287 (2011).
13. Poth, A. G., Colgrave, M. L., Lyons, R. E., Dalya, N. L. & Craik, D. J. Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. USA* **108**, 10127–10132 (2011).
14. Poth, A. G. *et al.* Discovery of cyclotides in the Fabaceae plant family provides new insights into the cyclization, evolution, and distribution of circular proteins. *ACS Chem. Biol.* **6**, 345–355 (2011).
15. Nguyen, K. N. T. *et al.* Immunostimulating and Gram-negative-specific antibacterial cyclotides from the butterfly pea (*Clitoria ternatea*). *FEBS J.* **283**, 2067–2090 (2016).
16. Oguis, G. K., Gilding, E. K., Jackson, M. A. & Craik, D. J. Butterfly pea (*Clitoria ternatea*), a cyclotide-bearing plant with applications in agriculture and medicine. *Front. Plant Sci.* **10** (2019).
17. Elena Felizmenio-Quimio, M., Daly, N. L. & Craik, D. J. Circular proteins in plants. Solution structure of a novel macrocyclic trypsin inhibitor from *Momordica cochinchinensis*. *J. Biol. Chem.* **276**, 22875–22882 (2001).
18. Mylne, J. S. *et al.* Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nat. Chem. Biol.* **7**, 257–259 (2011).
19. Koehbach, J. & Gruber, C. W. Cyclotides in the Rubiaceae. *Adv. Bot. Res.* **76**, 51–78 (2015).
20. Park, S. *et al.* Cyclotide evolution: Insights from the analyses of their precursor sequences, structures and distribution in violets (viola). *Front. Plant Sci.* **8**, 2058 (2017).
21. Gilding, E. K. *et al.* Gene coevolution and regulation lock cyclic plant defence peptides to their targets. *New Phytol.* **210**, 717–730 (2016).
22. Nguyen, G. K. T. *et al.* Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nat. Chem. Biol.* **10**, 732–738 (2014).
23. Nguyen, G. K. T. *et al.* Butelase 1: A versatile ligase for peptide and protein macrocyclization. *J. Am. Chem. Soc.* **137**, 15398–15401 (2015).
24. Nguyen, G. K. T. *et al.* Discovery of a linear cyclotide from the bracelet subfamily and its disulfide mapping by top-down mass spectrometry. *J. Biol. Chem.* **286**, 44833–44844 (2011).
25. Nguyen, G. K. T., Lim, W. H., Nguyen, P. Q. T. & Tam, J. P. Novel cyclotides and uncyclotides with highly shortened precursors from *Chassalia chartacea* and effects of methionine oxidation on bioactivities. *J. Biol. Chem.* **287**, 17598–17607 (2012).
26. Mylne, J. S. *et al.* Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase-mediated biosynthesis. *Plant Cell* **24**, 2765–2778 (2012).
27. Nguyen, G. K. T. *et al.* Discovery of linear cyclotides in monocot plant *Panicum laxum* of Poaceae family provides new insights into evolution and distribution of cyclotides in plants. *J. Biol. Chem.* **288**, 3370–3380 (2013).
28. Harris, K. S. *et al.* Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. *Nat. Commun.* **6**, 10199 (2015).
29. Jackson, M. A. *et al.* Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases. *Nat. Commun.* **9**, 2411 (2018).
30. Edman, J. C., Ellis, L., Blacher, R. W., Roth, R. A. & Rutter, W. J. Sequence of protein disulphide isomerase and implications of its relationship to thioredoxin. *Nature* **317**, 267–270 (1985).
31. Wang, C. C. Protein disulfide isomerase as an enzyme and a chaperone in protein folding. *Methods Enzymol.* **348**, 66–75 (2002).
32. Van Anken, E. & Braakman, I. Versatility of the endoplasmic reticulum protein folding factory. *Crit. Rev. Biochem. Mol. Biol.* **40**, 191–228 (2005).
33. Wilkinson, B. & Gilbert, H. F. Protein disulfide isomerase. *Biochim. Biophys. Acta Prot. Proteom.* **1699**, 35–44 (2004).
34. Freedman, R. B., Hirst, T. R. & Tuite, M. F. Protein disulphide isomerase: Building bridges in protein folding. *Trends Biochem. Sci.* **19**, 331–336 (1994).
35. Kemmink, J., Darby, N. J., Dijkstrat, K., Nilges, M. & Creighton, T. E. The folding catalyst protein disulfide isomerase is constructed of active and inactive thioredoxin modules. *Curr. Biol.* **7**, 239–245 (1997).
36. Gruber, C. W. *et al.* A novel plant protein-disulfide isomerase involved in the oxidative folding of cystine knot defense proteins. *J. Biol. Chem.* **282**, 20435–20446 (2007).
37. Frand, A. R. & Kaiser, C. A. Ero1p oxidizes protein disulfide isomerase in a pathway for disulfide bond formation in the endoplasmic reticulum. *Mol. Cell* **4**, 469–477 (1999).
38. Bulleid, N. J. & Ellgaard, L. Multiple ways to make disulfides. *Trends Biochem. Sci.* **36**, 485–492 (2011).
39. Sato, Y. & Inaba, K. Disulfide bond formation network in the three biological kingdoms, bacteria, fungi and mammals. *FEBS J.* **279**, 2262–2271 (2012).
40. Zito, E. ERO1: A protein disulfide oxidase and H2O2 producer. *Free Radic. Biol. Med.* **83**, 299–304 (2015).
41. O'brien, H. *et al.* Ero1-mediated reoxidation of protein disulfide isomerase accelerates the folding of cone snail toxins. *Int. J. Mol. Sci.* **19**, 3418 (2018).
42. Fan, F. *et al.* ATERO1 and ATERO2 exhibit differences in catalyzing oxidative protein folding in the endoplasmic reticulum. *Plant Physiol.* **180**, 2022–2033 (2019).
43. Burman, R. *et al.* Distribution of circular proteins in plants: Large-scale mapping of cyclotides in the Violaceae. *Front. Plant Sci.* **6**, 855 (2015).

44. Galat, A. Peptidylprolyl Cis/trans isomerases (immunophilins): Biological diversity—targets—functions. *Curr. Top. Med. Chem.* **3**, 1315–1347 (2005).
45. Di Luccio, E. *et al.* Parameters affecting in vitro oxidation/folding of maurotoxin, a four-disulphide-bridged scorpion toxin. *Biochem. J.* **358**, 681–692 (2001).
46. Safavi-Hemami, H., Bulaj, G., Olivera, B. M., Williamson, N. A. & Purcell, A. W. Identification of Conus peptidylprolyl cis-trans isomerases (PPIases) and assessment of their role in the oxidative folding of conotoxins. *J. Biol. Chem.* **285**, 12735–12746 (2010).
47. Safavi-Hemami, H. *et al.* Modulation of conotoxin structure and function is achieved through a multienzyme complex in the venom glands of cone snails. *J. Biol. Chem.* **287**, 34288–34303 (2012).
48. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
49. Grabherr, M. G. *et al.* Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2013).
50. Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).
51. Wang, Y., Coleman-Derr, D., Chen, G. & Gu, Y. Q. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res.* **43**, W78–W84 (2015).
52. Wang, C. K. L., Kaas, Q., Chiche, L. & Craik, D. J. CyBase: A database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res.* **36**, D206–D210 (2008).
53. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
54. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
55. Boratyn, G. M. *et al.* BLAST: A more efficient report with usability improvements. *Nucleic Acids Res.* **41**, W29–W33 (2013).
56. Sigrist, C. J. A. *et al.* PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–274 (2002).
57. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344–D347 (2013).
58. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
59. Okonechnikov, K. *et al.* Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28**, 1166–1167 (2012).
60. Rose, R., Golosova, O., Sukhomlinov, D., Tiunov, A. & Prosperi, M. Flexible design of multiple metagenomics classification pipelines with ugene. *Bioinformatics* **35**, 1963–1965 (2019).
61. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
62. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
63. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
64. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
65. Rambaut, A. FigTree v. 1.4.4. https://tree.bio.ed.ac.uk/software/figtree/ (2018).
66. Chevenet, F., Brun, C., Bañuls, A. L., Jacq, B. & Christen, R. TreeDyn: Towards dynamic graphics and annotations for analyses of trees. *BMC Bioinform.* **7**, 439 (2006).
67. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *Bioinformatics* https://doi.org/10.1201/b16589 (2014).
68. Metsalu, T. & Vilo, J. ClustVis: A web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res.* **43**, W566–W570 (2015).
69. Serra, A. *et al.* A high-throughput peptidomic strategy to decipher the molecular diversity of cyclic cysteine-rich peptides. *Sci. Rep.* **6**, 23005 (2016).
70. Langmead, B. & Salzberg, S. Bowtie2. *Nat. Methods* **9**, 357–359 (2013).
71. Koehbach, J. *et al.* Cyclotide discovery in Gentianales revisited-identification and characterization of cyclic cystine-knot peptides and their phylogenetic distribution in Rubiaceae plants. *Biopolymers* **100**, 438–452 (2013).
72. Sternberger, A. L. *et al.* Transcriptomics identifies modules of differentially expressed genes and novel cyclotides in *Viola pubescens*. *Front. Plant Sci.* **10**, 156 (2019).
73. Daly, N. L., Clark, R. J., Plan, M. R. & Craik, D. J. Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif. *Biochem. J.* **393**, 619–626 (2006).
74. Daly, N. L., Rosengren, K. J. & Craik, D. J. Discovery, structure and biological activities of cyclotides. *Adv. Drug Deliv. Rev.* **61**, 918–930 (2009).
75. Colgrave, M. L., Poth, A. G., Kaas, Q. & Craik, D. J. A new 'era' for cyclotide sequencing. *Biopolymers* **94**, 592–601 (2010).
76. Hellinger, R. *et al.* Peptidomics of circular cysteine-rich plant peptides: Analysis of the diversity of cyclotides from viola tricolor by transcriptome and proteome mining. *J. Proteome Res.* **14**, 4851–4862 (2015).
77. Koehbach, J. & Clark, R. J. Unveiling the diversity of cyclotides by combining peptidome and transcriptome analysis. *Biopolymers* **106**, 774–783 (2016).
78. Hemu, X. *et al.* Structural determinants for peptide-bond formation by asparaginyl ligases. *Proc. Natl. Acad. Sci. USA* **116**, 11737–11746 (2019).
79. Shewry, P. R., Napier, J. A. & Tatham, A. S. Seed storage proteins: Structures and biosynthesis. *Plant Cell* **7**, 945–956 (1995).
80. Kamauchi, S. *et al.* Molecular cloning and characterization of two soybean protein disulfide isomerases as molecular chaperones for seed storage proteins. *FEBS J.* **275**, 2644–2658 (2008).
81. Kimura, S., Higashino, Y., Kitao, Y., Masuda, T. & Urade, R. Expression and characterization of protein disulfide isomerase family proteins in bread wheat. *BMC Plant Biol.* **15**, 73 (2015).
82. Houston, N. L. *et al.* Phylogenetic analyses identify 10 classes of the protein bisulfide isomerase family in plants, including single-domain protein disulfide isomerase-related proteins. *Plant Physiol.* **137**, 762–778 (2005).
83. d'Aloisio, E. *et al.* The Protein Disulfide Isomerase gene family in bread wheat (*T. aestivum* L.). *BMC Plant Biol.* **10**, 101 (2010).
84. Onda, Y. & Kobori, Y. Differential activity of rice protein disulfide isomerase family members for disulfide bond formation and reduction. *FEBS Open Biol.* **4**, 730–734 (2014).
85. Kayum, M. A. *et al.* Genome-wide characterization and expression profiling of PDI family gene reveals function as abiotic and biotic stress tolerance in Chinese cabbage (*Brassica rapa* ssp. pekinensis). *BMC Genomics* **18**, 885 (2017).
86. Lu, D. P. & Christopher, D. A. Endoplasmic reticulum stress activates the expression of a sub-group of protein disulfide isomerase genes and AtbZIP60 modulates the response in *Arabidopsis thaliana*. *Mol. Genet. Genomics* **280**, 199–210 (2008).
87. Selles, B., Jacquot, J. P. & Rouhier, N. Comparative genomic study of protein disulfide isomerases from photosynthetic organisms. *Genomics* **97**, 37–50 (2011).
88. Wadahama, H., Kamauchi, S., Ishimoto, M., Kawada, T. & Urade, R. Protein disulfide isomerase family proteins involved in soybean protein biogenesis. *FEBS J.* **274**, 687–703 (2007).
89. Iwasaki, K. *et al.* Molecular cloning and characterization of soybean protein disulfide isomerase family proteins with nonclassic active center motifs. *FEBS J.* **276**, 4130–4141 (2009).

90. Safavi-Hemami, H. *et al.* Rapid expansion of the protein disulfide isomerase gene family facilitates the folding of venom peptides. *Proc. Natl. Acad. Sci. USA* **113**, 3227–3232 (2016).
91. Tavender, T. J. & Bulleid, N. J. Molecular mechanisms regulating oxidative activity of the Ero1 family in the endoplasmic reticulum. *Antioxid. Redox Signal.* **13**, 1177–1187 (2010).
92. Wang, Z. Q., Han, Y. H., Shao, X. X., Chi, C. W. & Guo, Z. Y. Molecular cloning, expression and characterization of protein disulfide isomerase from Conus marmoreus. *FEBS J.* **274**, 4778–4787 (2007).
93. Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. CDART: Protein homology by domain architecture. *Genome Res.* **12**, 1619–1623 (2002).
94. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
95. Moore, A. D., Heldy, A., Terrapon, N., Weiner, J. & Bornberg-Bauer, E. DoMosaics: Software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics* **30**, 282–283 (2014).
96. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).
97. Briesemeister, S., Rahnenführer, J. & Kohlbacher, O. YLoc-an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.* **38**, W497–W502 (2010).
98. Briesemeister, S. *et al.* SherLoc2: A high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.* **8**, 5363–5366 (2009).
99. Yu, C. S., Chen, Y. C., Lu, C. H. & Hwang, J. K. Prediction of protein subcellular localization. *Prot. Struct. Funct. Genet.* **64**, 643–651 (2006).
100. Blum, T., Briesemeister, S. & Kohlbacher, O. MultiLoc2: Integrating phylogeny and gene ontology terms improves subcellular protein localization prediction. *BMC Bioinform.* **10**, 274 (2009).
101. Liu, W. *et al.* IBS: An illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**, 3359–3361 (2015).

Acknowledgements

NK was supported by the Tata Education and Development Trust. PB acknowledges his support by the DST-YOS Chair Professorship awarded by the SERB, Department of Science and Technology, Government of India. Transcriptome sequencing was done by Agrigenome, India. Mass spectra was obtained using the Proteomics Facility supported by the Department of Biotechnology at the Indian Institute of Science. We thank NCBS (TIFR) and JC Bose fellowship of RS (SB/S2/JC-071/2015). This work was supported by Department of Science and Technology (Early Career Award, Ramanujan Fellowship (SB/S2/RJN-189/2014)), Max Planck Society (DST-Max Planck Partner group program) and Department of Biotechnology NER grant of RV.

Author contributions

N.K. carried out the entire work and wrote first draft of the manuscript. Plant purifications were performed by N.K. and R.V. The transcriptome sequencing analysis was analysed by N.K., and R.S. Mass spectrometry work was carried by N.K. and P.B. All authors participated in the meta-analysis and interpretation of data and improvement of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-69452-7.

Correspondence and requests for materials should be addressed to R.S.

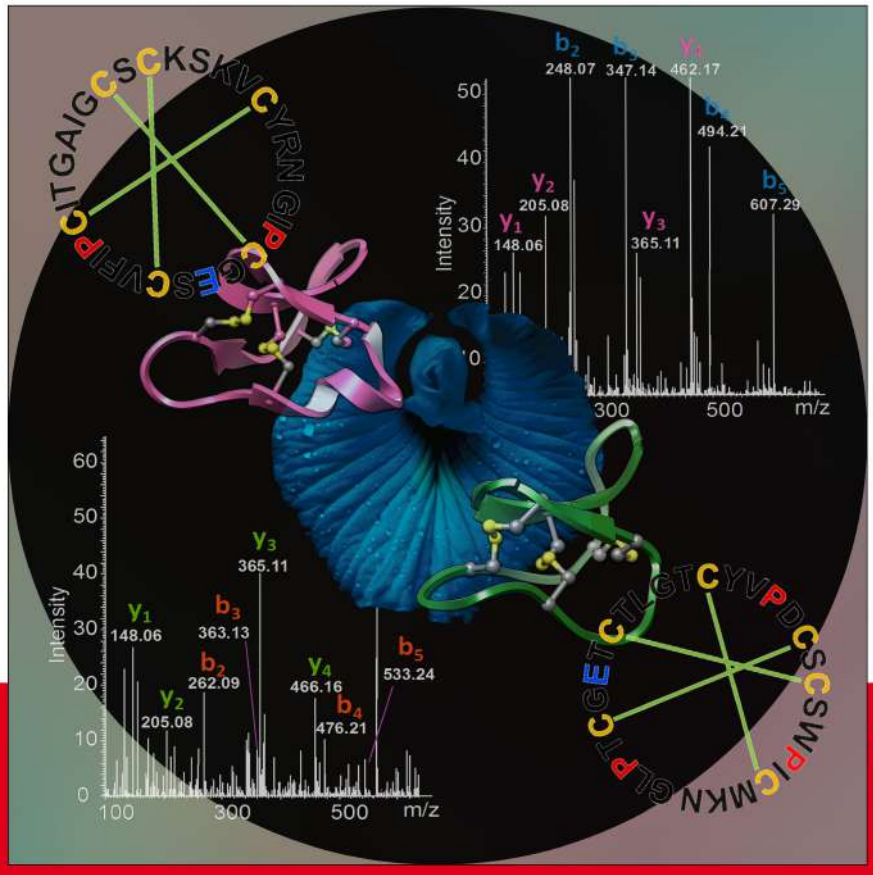
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHEMISTRY

AN ASIAN JOURNAL

www.chemasianj.org



16/19
2021

A Journal of

ACES
Asian Chemical Editorial Society

Front Cover:
Neha V. Kalmankar, Padmanabhan Balaram, Radhika Venkatesan
Mass Spectrometric Analysis of Cyclotides from *Clitoria ternatea*: Xxx-Pro Bond Fragmentation as Convenient Diagnostic of Pro Residue Positioning

WILEY-VCH

3

Proteome-Wide Discovery of Cyclotides from Butterfly Pea

Neha V. Kalmankar[1,2], Ramanathan Sowdhamini[1], Radhika Venkatesan[1,4] and Padmanabhan Balaram[1,3*]

Manuscript 3
Chemistry - an Asian Journal, 16: 2920 (2021)
https://doi.org/10.1002/asia.202100585
Manuscript 5 (In preparation)

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore 560065, Karnataka, India.
[2]The University of Trans-Disciplinary Health Sciences and Technology (TDU), #74/2, Jarakabande Kaval, Post Attur, Via Yelahanka, Bangalore 560064, Karnataka, India.
[3]Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, Karnataka, India.
[4]Department of Biological Sciences, Indian Institute of Science Education and Research, Mohanpur 741246 West Bengal, India.

Correspondence:
Prof. Padmanabhan Balaram
pb@iisc.ac.in

Manuscript 3

Mass spectrometric analysis of cyclotides from *Clitoria ternatea*: Xxx-Pro bond fragmentation as convenient diagnostic of Pro residue positioning

Neha V. Kalmankar*, Padmanabhan Balaram* and Radhika Venkatesan*

3.1 SUMMARY

To date, >600 cyclotides have been characterized from five major plant families such as - Rubiaceae, Violaceae, Fabaceae, Solanaceae and Cucurbitaceae. In the Fabaceae (legume) plant family, *Clitoria ternatea* represents the only member known to produce cyclotides. The study presented here describes the complete assortment of cyclotides obtained in the Indian variety of *C. ternatea* plant, across five tissues (pod, stem, leaf, flower and root), characterized by MALDI-TOF, high-resolution LC-MS and MS/MS techniques. Notable variations in MS/MS product ion distributions were observed in cyclotides belonging to the two structural subfamilies, Möbius and Bracelet, based on the number and positions of proline residues. Distinct fragmentation patterns was observed for prototypical cyclotides of the two structural subfamilies caused by the preferential fragmentation at Xxx-Pro bonds. These distinct mass spectral fingerprints were then used as a diagnostic method to rapidly identify and sequence novel cyclotides ctr pep 30 and ctr pep 43 from this plant.

3.2 RELATED INFORMATION

The published manuscript attached here refers to - Kalmankar NV*, Balaram P* and Venkatesan R* (2021) Mass spectrometric analysis of cyclotides from *Clitoria ternatea*: Xxx-Pro bond fragmentation as convenient diagnostic of Pro residue positioning. Chemistry - an Asian Journal. 16: 2920. DOI: 10.1002/asia.202100585. PMID: 34288513. This article is under the Agreement between Neha Kalmankar and John Wiley and Sons, licensed under Copyright Clearance Center's RightsLink® service (License number: 5207991024480), which permits personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce in the Dissertation/Thesis. Hence, this article has been incorporated here as it is without any changes and for non-commercial use as per the license conditions. Please note: The page numbers indicated within the PDF is independent of rest of the thesis.

ACES Asian Chemical Editorial Society
doi.org/10.1002/asia.202100585
CHEMISTRY AN ASIAN JOURNAL Full Paper

Special Collection

Mass Spectrometric Analysis of Cyclotides from *Clitoria ternatea*: Xxx-Pro Bond Fragmentation as Convenient Diagnostic of Pro Residue Positioning

Neha V. Kalmankar,*[a, b] Padmanabhan Balaram,*[a, c] and Radhika Venkatesan*[a, d]

Abstract: Cyclotides, a class of macrocyclic plant peptides, characterized by a cyclic backbone and three inter-locking disulfide bonds, may be divided into two major structural subfamilies, Möbius and Bracelet, based on the presence or absence of a specific proline residue. The present study describes the suite of cyclotides obtained from *Clitoria ternatea*, characterized by LC–MS and MS/MS techniques. Notable variations in product ion distributions were observed in cyclotides belonging to different structural subfamilies based on the number and positions of proline residues. For instance, Cter M which is an abundant Möbius cyclotide in this plant containing three proline residues, displayed distinct b- and y- ion characteristics in the MS/MS spectra compared to Cliotide T1, another commonly identified cyclotide but belonging to the Bracelet subfamily having two proline residues. The distinct fragmentation pattern of prototypical cyclotides of each structural subfamily, determined by Xxx-Pro bond fragmentation, was used to rapidly identify and sequence a novel cyclotide ctr pep 30 from this plant.

Introduction

Plants produce a wide variety of chemicals to defend themselves against microbes and herbivores. Among the various plant-defense molecules, cyclotides present as an interesting yet an untapped family of host defense agents. Cyclotides are plant defense mini-proteins comprising of 26–37 amino acids and bearing six conserved cysteine residues that form three disulfide bridges (Cys I-Cys IV, Cys II-Cys V, Cys III-Cys VI).[1,2] Additionally, they possess a cyclic backbone which contributes to the formation of a cyclic cystine knot (CCK) topology.[3] Specifically, this occurs at a highly conserved C-terminal asparagine residue, and the reaction is mediated by a distinct asparaginyl endopeptidase acting as a ligase.[4] Due to the conserved CCK topology, it renders these peptides with ultra-stability against thermal, chemical or enzymatic methods of degradation.[5] This unique property of high stability has made them excellent candidates for drug design and agriculture. They have attracted significant pharmaceutical attention due to activities such as cytotoxic, anti-cancer, anti-HIV, anti-microbial, antihelminthic, uterotonic, etc.[6–11] Cyclotides appear to be involved in plant defense mechanism, exhibiting insecticidal, antimicrobial, antifungal and nematocidal agents, hence, have also gained significant interest in agricultural applications.[2,12,13] This has led to the recent release of an ecofriendly pesticide - SeroX®, made from *Clitoria ternatea* extracts and marketed in Australia.[14] Growing interest in cyclotide research has led to their identification from a variety of plant families such as Rubiaceae, Violaceae, Cucurbitaceae, Fabaceae and Solanaceae.[11,15–18]

Cyclotide sequences and structures can be divided mainly into three subfamilies: Möbius, Bracelet and trypsin-inhibitor subfamilies[19] (Figure 1A). All three groups have the same CCK motif but differ from each other based on the composition of amino acid residues. The Möbius and Bracelet groups are similar to each other in terms of sequence, structure and functions, but differ in their size and content in the loops. However, the characteristic difference between the two is the presence of a *cis*-proline present in loop 5 in the Möbius group, which in turn creates a twist in the backbone, whereas this *cis*-proline is absent in loop 5 of the Bracelet group of sequences[3] (Figure 1A). Additionally, loops 2 and 3 of Möbius cyclotides contain mainly hydrophilic residues, whereas Bracelet cyclotides contain mainly hydrophobic residues in those loop regions.[20] The trypsin inhibitor group have distinct structural and sequence characteristics compared to the other subfamilies.[21]

C. ternatea is currently the only leguminous plant species from the Fabaceae family known to produce cyclotides in all the parts, despite examination of other diverse species of *Clitoria*.[22] Presently, there are ~110 cyclotide sequences

[a] *N. V. Kalmankar, P. Balaram, R. Venkatesan*
National Centre for Biological Sciences (NCBS), Tata Institute for Fundamental Research (TIFR)
GKVK Campus, Bangalore, Karnataka, 560065, India
[b] *N. V. Kalmankar*
The University of Trans-Disciplinary Health Sciences and Technology (TDU)
74/2, Jarakabande Kaval, Post Attur, Via Yelahanka,
Bangalore, Karnataka, 560064, India
E-mail: nehavk@ncbs.res.in
[c] *P. Balaram*
Molecular Biophysics Unit
Indian Institute of Science
Bangalore, Karnataka, 560012, India
E-mail: pb@iisc.ac.in
[d] *R. Venkatesan*
Department of Biological Sciences
Indian Institute of Science, Education and Research (IISER) Kolkata
Mohanpur, West Bengal, 741246, India
E-mail: rv@iiserkol.ac.in
radhika@ncbs.res.in

Supporting information for this article is available on the WWW under https://doi.org/10.1002/asia.202100585

This manuscript is part of a special collection celebrating the 15th Anniversary of IISER Inception.

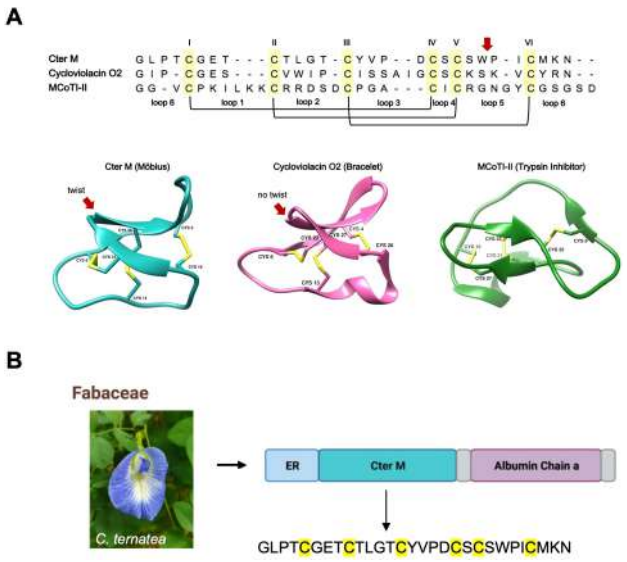


Figure 1. A) Top panel; Examples of cyclotide sequences. The six cysteines (I–VI) are highlighted in yellow. The intervening peptide stretches between the cysteine residues are termed as "loops" and numbered successively. Bottom panel; The three-dimensional structures (from left to right) of Cter M, cycloviolacin O2 and MCoTI-II belonging to the Möbius, Bracelet and trypsin inhibitor subfamilies, respectively. The disulfide bonds are indicated by yellow stick structures and the loop 5 region in Möbius and Bracelet structures is indicated by a red arrow. B) Schematic representation of the precursor protein arrangements of the butterfly pea cyclotide Cter M.

reported from *C. ternatea*, of which 90 have been confirmed at the transcript level from NGS technology.[12,22–24] Earlier studies on cyclotide characterization in *C. ternatea* have primarily made use of matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry, except for work by Serra et al. and Gilding et al., wherein electrospray ionization time-of-flight (ESI-qTOF) mass-spectrometry was used, respectively.[22,24] In the present study, in order to improve the quality and sensitivity of the data obtained, we made use of the Orbitrap Fusion Tribrid mass spectrometer, which performs electrospray ionization, using Orbitrap mass analyzer, coupled to liquid chromatography (LC) to identify and characterize cyclotides from five tissues of *C. ternatea*. Unlike in the counterpart plant families, the cyclotide gene in Fabaceae is unique as it appears to be a member of the albumin-1 gene family that have evolved to produce cyclotides from the region of albumin b-chain[17] (Figure 1B). Poth et al., had previously characterized 12 cyclotides (Cter A–L) from the seed extracts.[17] Soon after, they isolated 6 new sequences (Cter M–R) from leaf and flower tissues.[25] Separately, Nguyen et al., screened for cyclotides in the whole plant and reported 21 additional sequences (cT1–21), of both cyclic and acyclic forms, and named them as 'cliotides'.[12,26]

In the current study, transcriptomic data previously published by our group was used to accelerate the MS/MS sequencing process by creating a library of the transcriptome-derived sequences and CyBase entries producing a custom database in ERA format.[27–30] CyBase (Cyclic Proteins Database) is a publicly accessible curated databse of cyclic protein sequences and structures and documents ~700 cyclotide sequences discovered from multiple plant families.[28,29] A recent study by Parsley et al., describe a method to rapidly identify cysteine-rich-peptides in a mixture by identifying mass spectral (MS_2) fingerprint ions characteristic of homologus sequences before elucidating the full sequence.[31] Similarly, in order to facilitate rapid classification of cyclotides from the complex, microheterogenous, peptide mixtures occurring in plant extracts we have attempted to use mass spectral fragmentation pattern obtained during LC–ESI–MS analysis of *C. ternatea* extracts. Our goal was to distinguish the two broad structural subfamilies, Möbius and Bracelet, by inspection of MS/MS spectra based on the number of proline residues present in the sequence. The pronounced fragmentation of X-Pro bonds under CID conditions should in principle lead to distinct distributions of product ions.[32–34] The present study provides insights into the proteomic diversity of cyclotides across five different tissues of the wild-type *C. ternatea* plants. Several known cyclotide sequences were discovered and confirmed in all the tissues. Subsequently, we also performed a comparative analysis of the mass-spectral characteristics and signatures of the Möbius and Bracelet-type cyclotide sequences from this plant species. The common and abundant MS/MS ions characteristic of Möbius and Bracelet-type sequences were used to identify and sequence a novel cyclotide from the crude extract.

Results and Discussion

Utilizing the laborious purification and sequencing methods the maximum number of unique sequences reported from a single study is 54 cyclotides from *Viola tricolor*,[35] using MS/MS sequencing and mass deconvolution methods. Totally 702 sequences arising from various plant species have been deposited in CyBase to date.[28,29] As the ribosomal origin of cyclotides has been established, it paves an alternative way to discover novel sequences. The analysis at the nucleotide level not only determines the resulting peptide sequence but also provides the parent precursor sequence information. Therefore, we aimed to discover as many cyclotides as possible in butterfly pea plant using a combined approach of utilizing the nucleotide level information from our transcriptome study and peptide level information from the MS experiments taken from CyBase.[28] Our published transcriptome of *C. ternatea* revealed 51 distinct cyclotide domains arising from 71 precursor genes, and 26 of them were novel cyclotide sequences arising from four individual tissues (pod, stem, leaf and flower).[27] In the present study, apart from the four tissues that was used for RNA-Seq experiments, we have also included the root tissue. In this study, we report the MS-based discovery of cyclotides in five tissues of *C. ternatea*, the sole cyclotide producing member of the Fabaceae plant family.

Extraction of cyclotide-rich fractions

Initial screening for cyclotides in the herbal extracts of *C. ternatea* tissues revealed the presence of proteins with masses in the range 2500–4000 Da, consistent with those of known cyclotides.[30,35,36] Elutes of 50%, 70% and 100% ACN/ddH_2O showed cyclotide-like masses via MALDI-TOF MS. (Figure S1). These three elutes were pooled for each plant tissue and freeze-dried for further MS/MS analysis (crude extract).

Tissue-specific variation of cyclotides

To understand if there is tissue specific variation of cyclotides in *C. ternatea*, crude extracts of different tissues including that of pods, stems, leaves, flowers and roots were subjected to LC–MS analysis (Figure 2). An assortment of cyclotides were detected and tissue specific variation was observed. A few molecular masses detected in some tissue types were not detected in the others (Figure 2, Tables S1 and S2). The LC–MS ion chromatograms in Figure 2 suggest that cyclotides characteristic of a specific tissue can be identified. For example, a mass M = 3139.32 Da, corresponding to the cyclotide Cter 30, was observed at retention time (RT) of 22 min in the stem (S10), leaf (L3) and flower (F5) extracts, but were absent in other two tissues. Similarly, a mass of M = 3226.47 Da, corresponding to the cyclotide Cter R, was detected at RT of 33 min in pod, leaf, flower and root extracts (P14, L20, F18 and R12) and this mass was absent in the stem tissue (Tables S1 and S2). Figure 2 also shows the isotopic distribution of a charge state corresponding to a representative, unique cyclotide mass found in a specific tissue. For example, m/z = 1130.51 was found at RT of 33.1 min in the root extract, which corresponds to a cyclotide mass of 3550.59 Da of an unknown identity, and this mass signal was absent in the other tissues. It is also evident that the relative concentration of cyclotides varied with the type of tissue. However, as these are crude extracts, deconvoluting individual masses and comparing across the tissues is challenging due to the overlap of peaks.

LC–MS/MS characterization of cyclotides

Sequencing of cyclotides is not as straightforward as compared to linear peptides. The presence of the CCK motif makes cyclotides topologically complex, and the absence of a free N-terminus hinders fragmentation in tandem MS analyses. In the present study the knotted topology of cyclotides were disrupted by biochemical derivatization such as thiol reduction using dithiothreitol and the free sulfhydryl groups were alkylated using iodoacetamide. The cyclic backbone, which is biosynthetically mediated by asparaginyl endopeptidase at the highly conserved C-terminal Asn (N) residue, was further linearized using endoproteinase Glu-C (endo Glu-C), taking advantage of another highly conserved Glu (E) residue in loop 1[37] (Figure 3). The identification and sequencing of cyclotides from Indian variety of *C. ternatea* was performed using Orbitrap

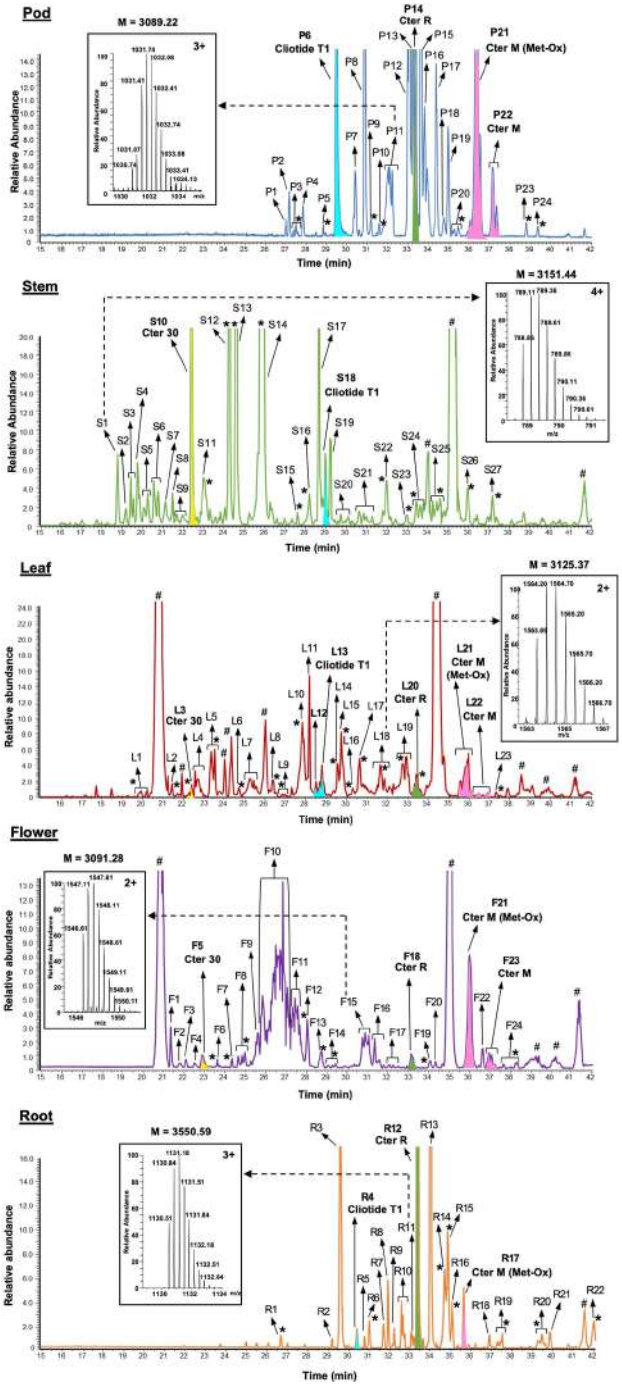


Figure 2. Crude extract from five tissues (pod, stem, leaf, flower and root) of *C. ternatea* was analyzed via LC–MS. Ion chromatograms of mass signals corresponding to cyclotide masses (~3 kDa) in each plant tissue are shown. Cyclotide masses were found to carry 2^+, 3^+, and 4^+ H-ions. Insets in each chromatogram shows the isotopic distribution of a charge state corresponding to a representative, unique, tissue specific cyclotide mass. The most abundant signals under each peak were assigned and the determined cyclotide mass signals were consecutively numbered as per retention times and tissue source. Peaks without any cyclotide-like masses and those containing mass signals corresponding to small molecules in high abundance are highlighted with a # and *, respectively. Peaks corresponding to Cliotide T1, Cter M (and Cter M (M20-oxidized)), Cter 30 and Cter R are highlighted in cyan, pink, yellow and green colours.

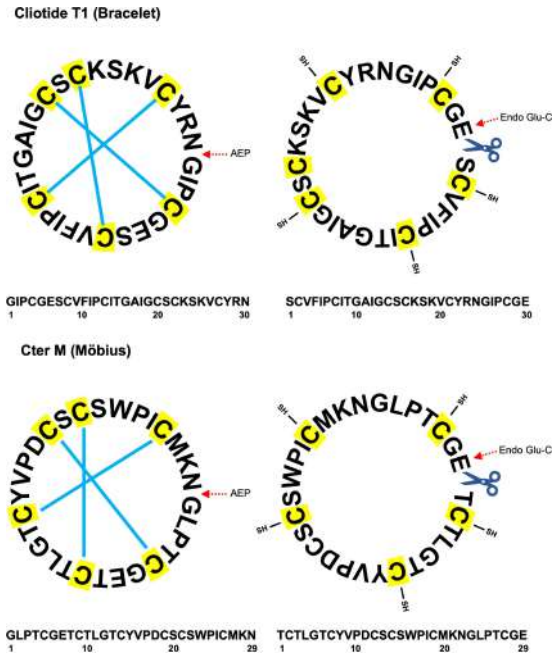


Figure 3. Circular representation of cyclotides. Cliotide T1 (top panel) and Cter M (bottom panel) sequences. One letter amino acid code showing the sequence, circular backbone and cystine knot arrangement. The positions where asparaginyl endopeptidase (AEP; red arrows, dashed line) ligates and endoproteinase GluC (endo Glu-C; red arrows, dashed line) digests the intact cyclotides are indicated.

Fusion Tribrid mass spectrometer coupled to liquid chromatography (LC). Figure 2 shows the ion chromatograms of the crude extracts from five tissues (pod, stem, leaf, flower and root) of *C. ternatea* analyzed via LC–MS. All the chromatograms were manually investigated for the presence of cyclotide-mass signals i.e. if their masses were in the range of 2,500–4000 Da, as deconvoluted from their double, triple and quadruple charged ions. The most abundant mass signals under each peak were assigned as a cyclotide mass and consecutively numbered by increasing retention times and tissue source (P1-Pn (pod); S1-Sn (stem); L1-Ln (leaf); F1-Fn (flower) and R1-Rn (root)). Then, the presence of three disulfide bonds was confirmed if the peaks in the reduced and alkylated extracts showed an increase in mass by 348.18 Da (Table S1).

MS fragmentation-based classification of cyclotides

The structural classification of cyclotides into Möbius and Bracelet subfamily has primarily been based on the presence of *cis*-Pro residue in the loop 5 region.[3] However, there are increasing reports of sequence intermediates which have characteristics of both Bracelet and Möbius cyclotides and have been classified as the hybrid subfamily.[38,39] Nonetheless, our analysis suggests that cyclotide sequences could be classified based on the number of proline residues present in regions other than loop 5. For mass spectrometry-based sequencing of cyclotides, reduction and alkylation of disulfides is followed by Glu-C digestion producing linear peptides with the C-terminus glutamic acid residue (Figure 3). In our subsequent discussion we will represent all linearized cyclotide sequences with the Glu (E) residue placed at the C-terminus. Conventionally, linearized sequences have been represented with Gly (G) at the N-terminus and Asn (N) at the C-terminus, which represents the linear precursor for enzymatically catalyzed ligation reaction. In broadly classifying mass spectra into cyclotide classes which differ in the number of proline residues we have used the tendency of X-Pro bonds undergo facile cleavage under collision-induced-dissociation (CID) conditions. Table 1 provides a representative list of cyclotides from *C. ternatea* characterized

Table 1. Representative list of mass spectrometrically characterized *C. ternatea* cyclotides classified by proline content[a]

Name	Alternative name[s]	Sequence	Average Mass [Da]	Monoisotopic Mass [Da]	Number of Basic residues [K, R, H][b]	Type [c]	Tissue-wise distribution[d]
1 Proline							
Cliotide T16	ctr pep 27	TCLRGRCYTPGCTCDHGICKKNGSVIGCGE	3108.57	3106.31	K = 2, R = 2, H = 1	H	S
2 Prolines							
Cliotide T1		SCVFIPCITGAIGCSCKSKVCYRNGIPCGE	3085.65	3083.36	K = 2, R = 1, H = 0	B	P, S, F, R
Cter P	Cliotide T4	SCVFIPCITAAIGCSCKSKVCYRNGIPCGE	3099.68	3097.38	K = 2, R = 1, H = 0	B	P, L, F, R
Cter Q	Cliotide T5	SCVFIPCISTVIGCSCKNKVCYRNGIPCGE	3170.76	3168.41	K = 2, R = 1, H = 0	B	P, R
Cliotide T12		SCVFIPCITGAIGCSCKSKVCYRDGIPCGE	3086.64	3084.35	K = 2, R = 1, H = 0	B	S, L
Cter R	Cliotide T7, ctr pep 34	SCVFIPCTVTALLGCSCKDKVCYKNGIPCGE	3228.84	3226.45	K = 3, R = 0, H = 0	B	P, L, R
3 or more Prolines							
Cter M	Cliotide T3	TCTLGTCYVPDCSCSWPICMKNGLPTCGE	3059.54	3057.20	K = 1, R = 0, H = 0	M	P, L, F, R
Cliotide T18	Cter 6, ctr pep 28	TCFTGTCYTPGCTCSYPVCKKNGLPICGE	3023.49	3021.23	K = 2, R = 0, H = 0	M	S, F
Cter 30	Cliotide T45	TCFKTKCYTPGCSCSYPVCKKNGFPICGE	3141.68	3139.32	K = 4, R = 0, H = 0	M	S, L
Cter 7		SCFAGKCYT**P**GCTCEY**P**ICMNNGD**P**FKCGE	3215.64	3213.19	K = 2, R = 0, H = 0	M	F

[a] A full list of mass spectrometrically characterized cyclotide sequences is provided in Table S3. [b] K: Lysine; R: Arginine; H: Histidine. [c] B: Bracelet; M: Möbius; H: Hybrid. [d] P: pod; S: stem; L: leaf; F: flower; R: root.

by mass spectral analysis and classified by number of proline residues. The presence of multiple basic residues permits the detection of 2^+, 3^+ and 4^+ charge states in cyclotide mass spectra providing a wealth of fragment information.

Cliotide T1: The linearized cyclotide sequence consists of 30 residues with Pro residues at positions 6 and 27. Cliotide T1 forms a Bracelet structure.[26] Figure 4 shows the ESI-MS/MS spectra of Cliotide T1 obtained for the 3^+ and 4^+ charge states from the pod extract. MS/MS spectrum of the 5^+ charge state is shown in Figure S2. Inspection of the 4^+ charge state fragmentation pattern reveals an extremely intense y_{25}^{3+} ion (m/z = 948.76), corresponding to cleavage of I5–P6 bond (Figure 4A). The intense y_4 ion (m/z = 462.17) and the internal fragment ion [P6-I26] arise from cleavage of I26-P27 bond. Several b- and y- ions confirming the sequence of Cliotide T1 are readily assigned (Figure 4). Cliotide T1 contains the positively charged residues – K17, K19 and R23. Together with the N-terminus amino group this provides four sites of protonation. Fragmentation along the backbone will be facilitated by proton migration. Interestingly, MS/MS spectrum of the 3^+ charge state reveals an intense y_{25}^{2+} (m/z = 1422.64) ion (Figure 4B). In this case, the distribution of triply charged ions might favor appreciable charge localization at the N-terminus due to the proximity of the lysine (K17, K19) residues. Fragmentation of the 5^+ charge state in which one proton must be localized along the backbone is a different distribution of intense product ions, with y_4 (m/z = 462.17) and the N-terminus ions (b_3, b_4 and b_5) being predominant (Figure S2). Preferential protonation at the tertiary amide bonds of the two Ile-Pro segments should facilitate these cleavages.

Cter M: The linearized cyclotide sequence consists of 29 residues with Pro residues at positions 10, 17 and 25. Cter M has a well characterized three-dimensional structure (PDB ID: 2LAM) based on NMR studies, placing it in the Möbius subclass. Cter M contains a single lysine residue at position 21 and was detected in the gas phase from the pod extract as having 2^+, 3^+ and 4^+ charge states, with the 3^+ charge state being the most intense. A comparison of the fragmentation patterns of 4^+ charge states of Cliotide T1 (Figure 4A) and Cter M (Figure 5A) clearly reveals significant differences in the distribution of the major product ions. The most intense ion y_5 (m/z = 563.21) arises by cleavage of L24–P25 bond. Other moderately intense peaks arise from cleavage of the V9–P10 bond (y_{20}^{2+}; m/z = 1184.98) and the cleavage of W16–P17 bond (y_{13}; m/z = 1476.66). A similar pattern of product ion distributions is also observed in the fragmentation pattern of the 3^+ charge state (Figure 5B). Fragmentation of the 2^+ charge state yielded y_5 and y_{13} ions of appreciable intensities (Figure S3). While the y_{20}^{2+} ion was not detectable, the observation of the b_9 ion (m/z = 1056.45) confirms cleavage at V9-P10 bond (Figure S3). The observation of intense internal fragment ions corresponding to the segments [10–16] (m/z = 893.29) and [17–24] (m/z = 914.46) establish that the three X-Pro bond cleavages dominate the fragmentation pattern of this cyclotide sequence.

ctr pep 30: Of the 110 unique cyclotide sequences currently known from *C. ternatea*, 26 were detected in our previously reported transcriptomic analysis.[27] In order to test whether a rapid identification of the new sequences could be achieved in our plant extract, we have chosen one example of a mass spectrometrically detected peptide of mass M_{expt} = 3142.38 Da, which matches the calculated mass for the sequence ctr pep 30 (M_{calc} = 3142.36) from the leaf extract. After reduction, alkylation and endo Glu-C linearization this would result in as mass of M_{calc} = 3508.54. Figure 6 shows the mass spectral fragmentation pattern obtained from the 3^+ and 4^+ charge states of the reduced, alkylated, linear peptide (M_{expt} = 3508.56). The similarity in the distribution of intense product ions of the Cliotide T1 spectra (Figure 4) confirms that this peptide contains a similar placement of the two internal proline residues characteristic of the Bracelet family. Alignment of Cliotide T1 and ctr pep 30 sequences (Figure 6A) shows five positions where sequence changes have occurred (positions 9–11, 18 and 26). The identical m/z values obtained for the b_5 (m/z = 607.29) and y_4 (m/z = 462.17) ions in both the peptides confirms the identity of residues 1–5 and 27–30 (Figure 6B and 6 C). Inspection of the 3^+ charge state fragmentation pattern reveals an intense y_{25}^{2+} ion (m/z = 1452.14) corresponding to cleavage of I5–P6 bond. Interestingly, MS/MS spectrum of the 4^+ charge state reveals an intense y_{25}^{3+} (m/z = 968.43) ion. The observation of intense b_5, y_4 and y_{25} ions and several internal ions confirms the sequence of ctr pep 30. These results establish the experimental detection of the transcriptomically derived sequence of ctr pep 30 in the leaf extract of *C. ternatea*. This peptide could also be detected in low abundance in the pod extract.

Ambiguities in mass spectral analysis

As the database of cyclotides sequences grows, it is increasingly difficult to unambiguously assign sequences based only on measured masses. Figure S4 provides an illustrative example of three nearly identical masses (M = 3083.34 Da, 3083.36 Da and 3083.36 Da) of sequences of *C. ternatea* reported by three different research groups.[17,26,27] The positioning of proline residues in the sequences reveals that ctr pep 43 is clearly distinguishable from Cliotide T1 and Cter O by inspection of MS/MS fragmentation patterns. However, the distinction between Cliotide T1 and Cter O would require a detailed analysis of MS/MS spectra in which a continuous train of sequential product ions is observed. This is often not the case in the fragmentation patterns obtained under CID conditions of peptides of length 20 residues or more containing internal proline residues. Figure 7 shows an illustrative example of a cyclotide (molecular mass M = 3083 Da) detected in stem extract. Initial inspection of fragmentation pattern of 4^+ charge state suggested that this might correspond to ctr pep 43. Figure 7A shows the aligned sequences Cliotide T1 and ctr pep 43 indicating the masses corresponding to internal ions generated between two proline residues. Our initial assignment was supported by the charge state distributions observed for the two cyclotides. In presence of as many three basic residues in Cliotide T1 results in the observation of 3^+, 4^+ and 5^+ charge states (Figure 7B and Figure S5), whereas in the peptide tentatively identified as ctr pep 43 only the 3^+ and 4^+ charge

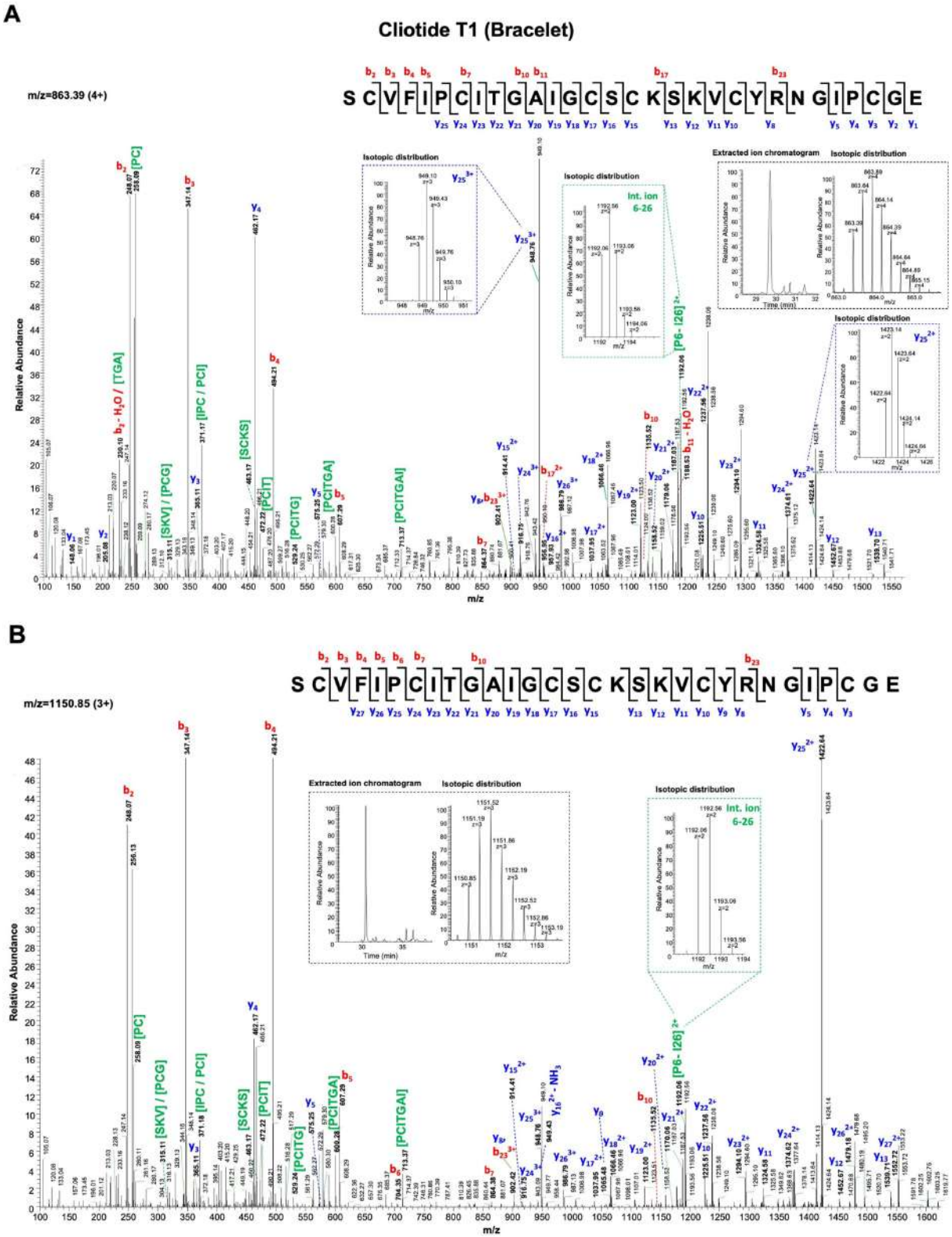


Figure 4. Peptide sequencing of Cliotide T1 (Bracelet) using CID-based MS/MS spectra of the precursors ion with m/z (A) 863.39 (4^+) and (B) 1150.85 (3^+) from reduced, alkylated and endoproteinase Glu-C digests (M = 3449.56 Da) of the *C. ternatea* pod extract. The amino acid sequence was determined by Proteome Discoverer v2.1 software (Thermo Scientific) and confirmed by the manual assignment of N-terminal b- and C-terminal y-ion series. The b-, y- and internal ions are indicated in red, blue and green font, respectively. Masses labeled in the spectra refer to monoisotopic $[M+H]^+$ mass. The peptide sequence deduced from each MS/MS data set is highlighted at the top in the respective panel, where each Cys (C) residue represents a carbamidomethyl-cysteine. Insets with black dashed border indicate the extracted ion chromatogram and isotopic distribution of the precursor ions. Insets with blue and green dashed border indicate the isotopic distribution of multiply charged y-ions and internal ions, respectively.

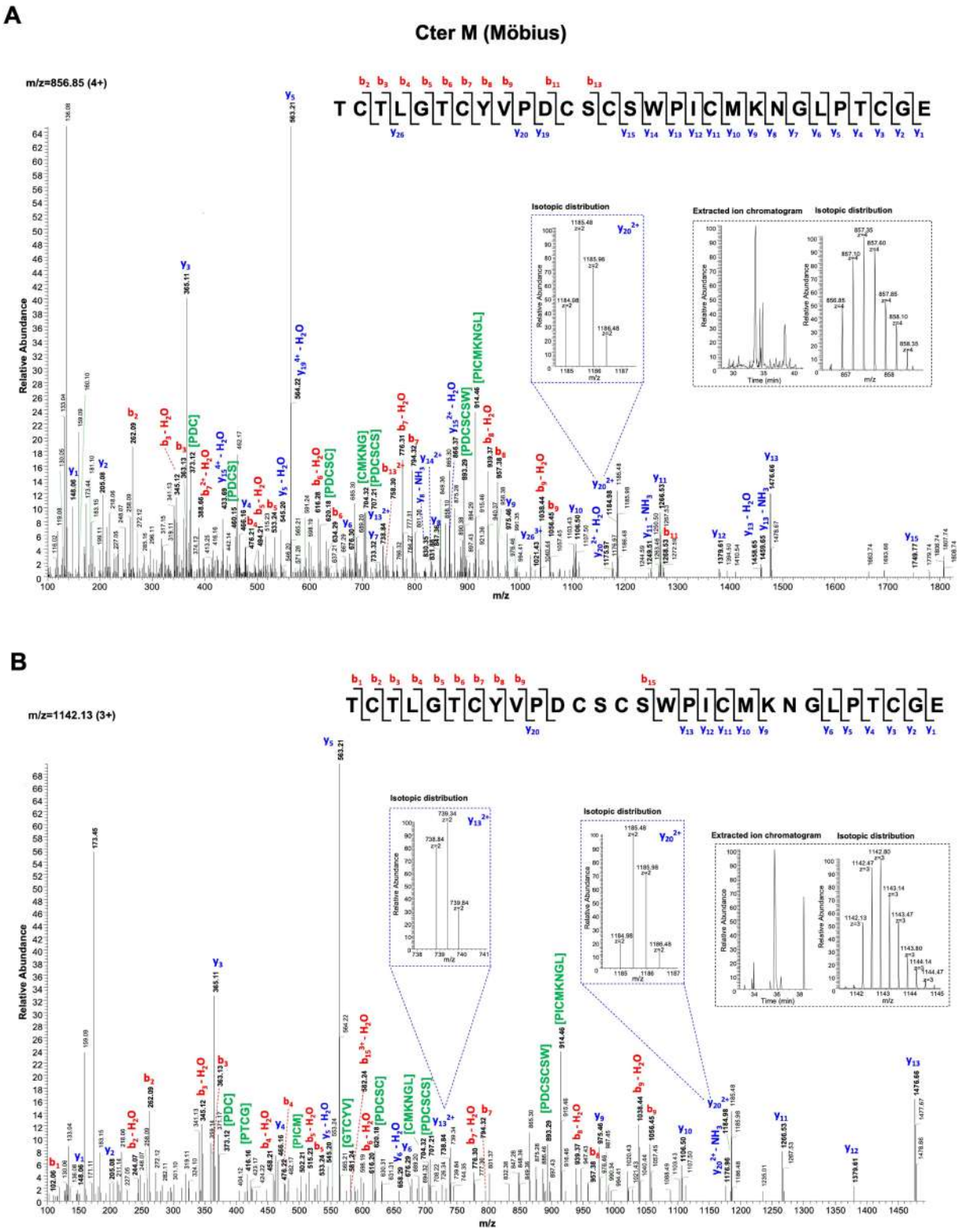


Figure 5. Peptide sequencing of Cter M (Möbius) using CID-based MS/MS spectra of the precursor ions with m/z (A) 856.85 (4^+) and (B) 1142.13 (3^+) from reduced, alkylated and endoproteinase Glu-C digests (M = 3423.39 Da) of the *C. ternatea* pod extract. The amino acid sequence was determined by Proteome Discoverer v2.1 software (Thermo Scientific) and confirmed by the manual assignment of N-terminal b- and C-terminal y-ion series. The b-, y- and internal ions are indicated in red, blue and green font, respectively. Masses labeled in the spectra refer to monoisotopic $[M+H]^+$ masses. The peptide sequence deduced from each MS/MS data set is highlighted at the top in the respective panel, where each Cys (C) residue represents a carbamidomethyl-cysteine. Insets with black dashed border indicate the extracted ion chromatogram and isotopic distribution of the precursor ions. Insets with blue and green dashed border indicate the isotopic distribution of multiply charged y-ions and internal ions, respectively.

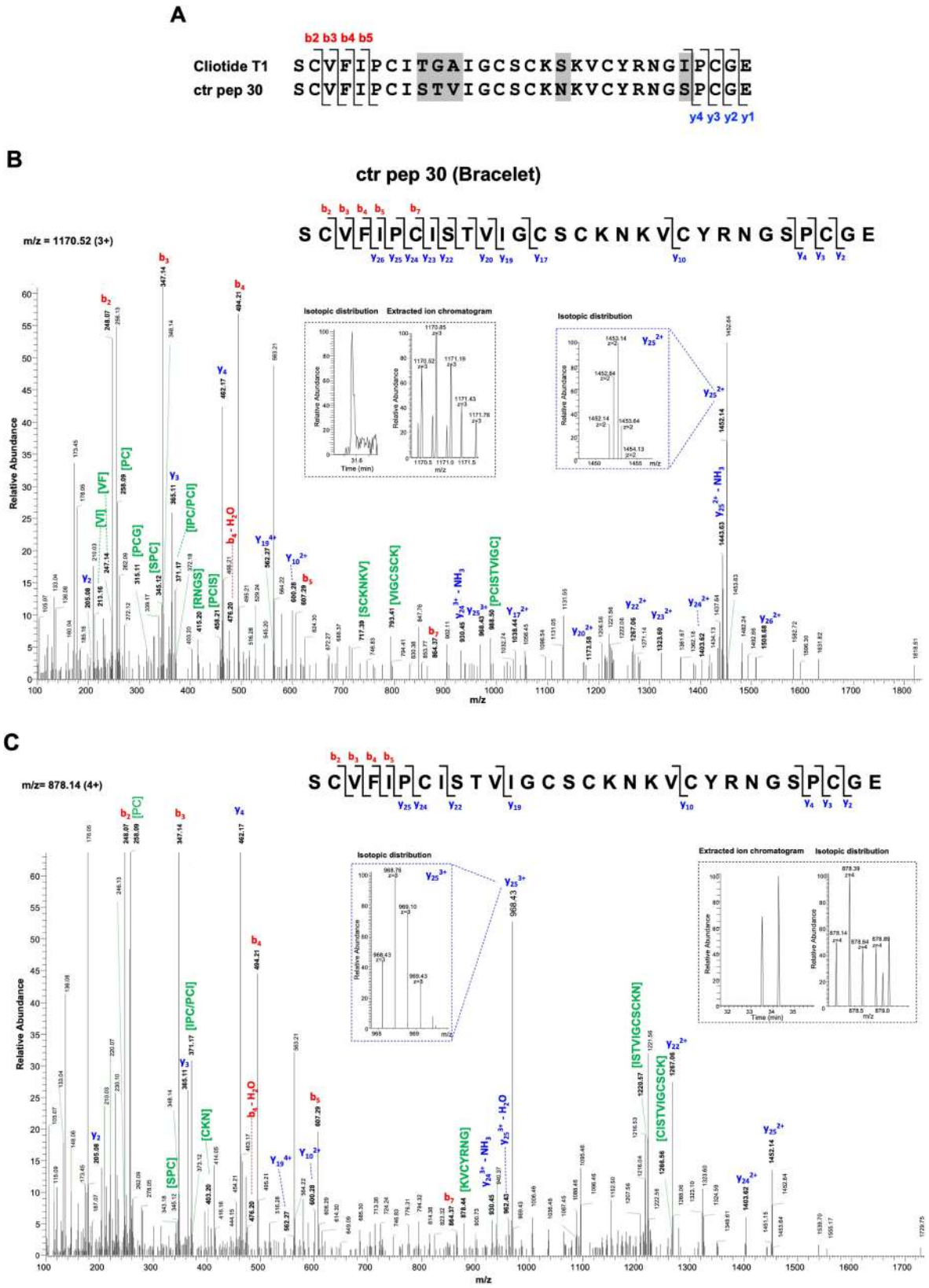


Figure 6. (A) Pairwise sequence alignment of Cliotide T1 and novel cyclotide ctr pep 30. Peptide sequencing of ctr pep 30 (Bracelet) using CID-based MS/MS spectra of the precursor ions with m/z (B) 1170.52 (3^+) and (C) 878.14 (4^+) from reduced, alkylated and endoproteinase Glu-C digests (M = 3508.56 Da) of the *C. ternatea* leaf extract. The amino acid sequence was determined by Proteome Discoverer v2.1 software (Thermo Scientific) and confirmed by the manual assignment of N-terminal b- and C-terminal y-ion series. The b-, y- and internal ions are indicated in red, blue and green font, respectively. Masses labeled in the spectra refer to monoisotopic $[M+H]^+$ masses. The peptide sequence deduced from each MS/MS data set is highlighted at the top in the respective panel, where each Cys (C) residue represents a carbamidomethyl-cysteine. Insets with black dashed border indicate the extracted ion chromatogram and isotopic distribution of the precursor ions. Insets with blue dashed border indicate the isotopic distribution of multiply charged y-ions.

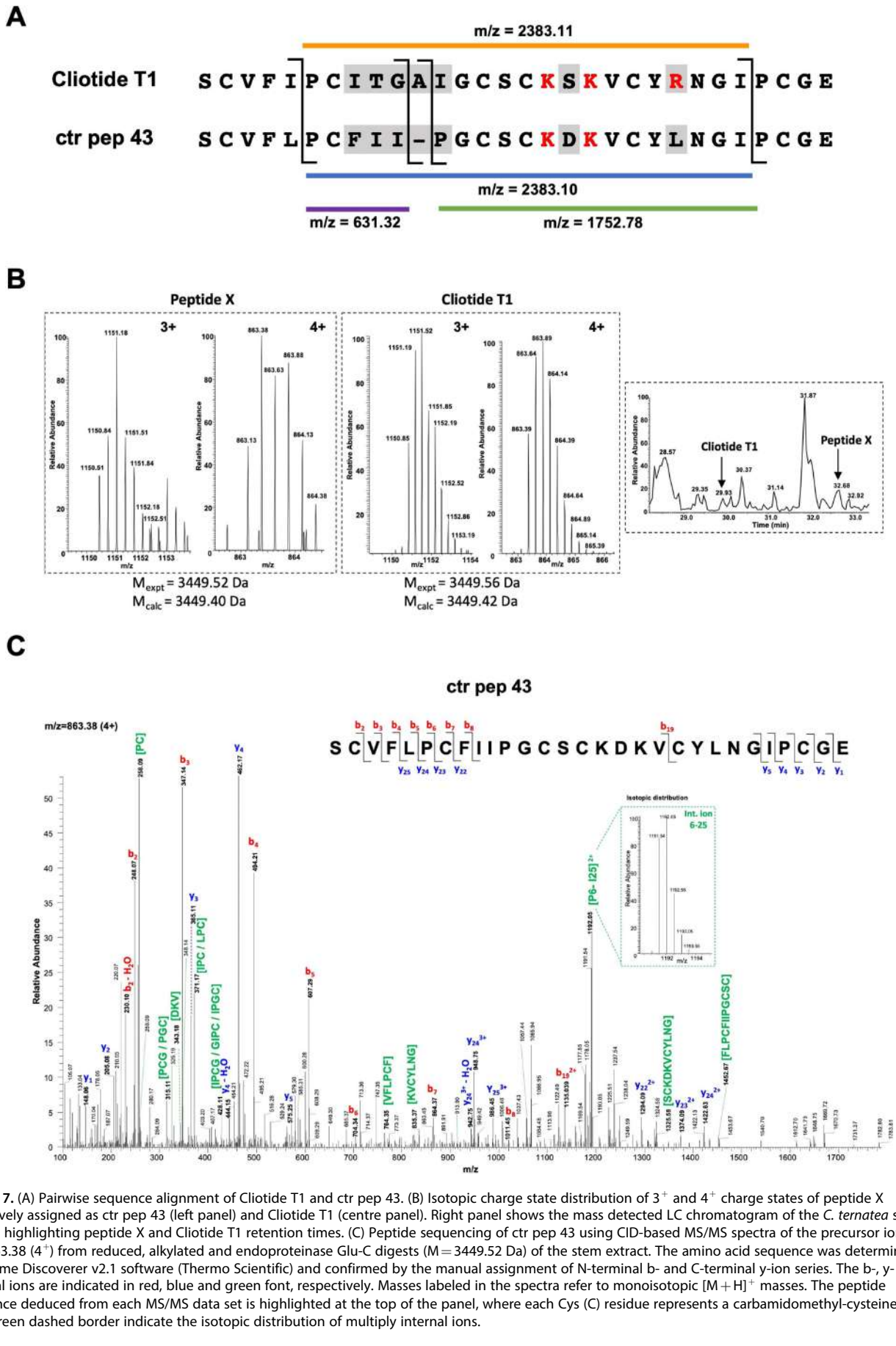


Figure 7. (A) Pairwise sequence alignment of Cliotide T1 and ctr pep 43. (B) Isotopic charge state distribution of 3^+ and 4^+ charge states of peptide X tentatively assigned as ctr pep 43 (left panel) and Cliotide T1 (centre panel). Right panel shows the mass detected LC chromatogram of the *C. ternatea* stem extract highlighting peptide X and Cliotide T1 retention times. (C) Peptide sequencing of ctr pep 43 using CID-based MS/MS spectra of the precursor ions with m/z 863.38 (4^+) from reduced, alkylated and endoproteinase Glu-C digests (M = 3449.52 Da) of the stem extract. The amino acid sequence was determined by Proteome Discoverer v2.1 software (Thermo Scientific) and confirmed by the manual assignment of N-terminal b- and C-terminal y-ion series. The b-, y- and internal ions are indicated in red, blue and green font, respectively. Masses labeled in the spectra refer to monoisotopic $[M+H]^+$ masses. The peptide sequence deduced from each MS/MS data set is highlighted at the top of the panel, where each Cys (C) residue represents a carbamidomethyl-cysteine. Inset with green dashed border indicate the isotopic distribution of multiply internal ions.

states were observed (Figure 7B). A further complication is evident upon comparison of the isotopic multiplet intensities for each charge state. For the peptide shown in Figure 7B, the intensity distribution suggests that a mixture of two overlapping peptides of mass $M_{expt} = 3449.52$ and $M_{expt} = 3449.56$ are present. While the N- and C- terminal b- and y- ions confirm the identity of five residues at N-terminus and four residues at the C- terminus, the accidental degeneracy of masses of the internal segments between P6 and P26/27 preclude unambiguous assignment of the intense internal fragment ion at m/z = 1192. This example emphasizes the need for caution in assigning cyclotide sequences by matching limited MS/MS data to transcriptome derived sequences.

In general, one anticipates a fairly extensive lack of correspondence between transcriptomic datasets and mass spectrometrically derived masses in the case of complex mixtures of very closely related peptide sequences. Plant cyclotide mixtures may provide many examples of ambiguities in sequence determination of minor components. Nevertheless, inspection of mass spectral product ions of already characterized peptide sequences can be leveraged for gaining additional information on novel masses of cyclotides, trypsin inhibitors, and other related cysteine-rich peptides in complex mixtures.[31]

Conclusion

Accurate sequencing of peptides in complex biological mixtures is often challenging due to identical mass species, modifications caused by either sample handling or inherent post-translational modifications (PTMs), low abundant peptides and incomplete MS/MS fragmentation. In the case of macrocyclic proteins such as the cyclotides, these problems are frequent and often laborious. In a recent study by Parsley et al., the authors combined mass spectral fragment ions that are characteristic of referenced cyclotides in CyBase and coupled it to mass shift analysis to determine cyclotides masses in *Viola communis* extracts.[31] Using this approach, the authors were able to identify novel cysteine-rich peptides in *Lagenaria siceraria* plant, namely LSCRP-I, -II, and -III. This is the only other study, apart from the one presented here, that has attempted to establish a method for accelerating the mass-spectral characterization and classification for novel cyclotides in a mixture.[31] Nevertheless, next-generation sequencing (NGS) technologies can be used in combination with tandem MS/MS to fast-track the assignment of sequences.

In the present study, transcriptomic data previously published by our group was used to uncover a suite of cyclotide sequences from multiple tissues of *C. ternatea*. Notable variations in cyclotide mass signals were observed in different plant tissues (pods, stems, leaves, flowers and roots). Several cyclotides were observed specifically in one tissue type, but were absent in the other. For rapid classification of cyclotides in the plant extracts, we initially distinguished the MS/MS spectra obtained for two prototypical cyclotides, Cliotide T1 and Cter M, belonging to two structural subfamilies, Bracelet and Möbius, respectively. We used the mass spectral fragmentation patterns obtained during LC–ESI–MS analysis and the number and position of proline residues present in the sequence to determine signatures of fragmentation for a particular subfamily. The fragmentation of X-Pro bonds for both Cliotide T1 (Bracelet) and Cter M (Möbius) under CID conditions lead to distinct distributions of product ions, which further guided the identification of a novel peptide ctr pep 30. The identical m/z values obtained for the b_5 and y_4 ions in both ctr pep 30 and Clitotide T1 arising from cleavages of two Xxx-Pro bond cleavages at the N- and C-terminus, allowed rapid sequencing of the new cyclotide and classifying it as belonging to the Bracelet family. Identification of mass spectral fragmentation ions of already characterized peptide sequences can be used to quickly discern masses in complex mixtures, aiding in the characterization of novel cyclotides. Unambiguous sequence determination will necessarily depend on the quality of mass spectral fragmentation patterns obtained for individual peptide components.

Experimental Section

Plant material

C. ternatea seeds were sown and grown in the campus of National Centre for Biological Sciences, Bangalore (13°04′13.8″N 77°34′49.9″E). Mature plants were collected and separated into 5 tissues: leaves, stems, pods, flowers and roots and oven dried at 70 °C.

Extraction of cysteine-rich peptides

Cyclotide extraction were performed using established methods.[25,27] The dried tissue samples were individually macerated using kitchen grinder. Peptides were extracted from each tissue using 1:1 (v/v) mixture of CH_2Cl_2 and MeOH under continuous agitation at 25 °C. The green extract was filtered and partitioned in 1:1 (v/v) mixture of CH_2Cl_2 and MeOH: ddH_2O (2:3 (v/v)) by liquid/liquid phase separation. The organic solvents were evaporated and the remaining aqueous phase was further fractionated on a manually packed RP-C_{18} silica gel column eluted (particle size: 40–63 μm; pore size: 60 Å; Polygosil®). A stepwise gradient of 0%, 30%, 50%, 70% and 100% of ACN/ddH_2O solvent was. Each elute was collected and concentrated using a rotary evaporator.

Screening for cysteine-rich peptides in crude extracts

Elutes of 0%, 30%, 50%, 70% and 100% ACN/ddH2O from each plant tissue were screened for cysteine-rich peptides using MALDI-TOF (UltraFlex Bruker Daltonics mass spectrometer) for cyclotide-like masses (2.5–4.0 kDa). MALDI-TOF mass-spectrometry was performed with 50 Hz pulsed nitrogen laser ($\lambda = 337$ nm), in positive ion reflectron mode. The samples were prepared in 1:1 mixture of H_2O-ACN with 0.1% TFA by mixing an equal amount of peptide fractions (0.5 μL) with α-cyano-4-hydroxycinnamic acid (CCA)/2,5-dihydroxy benzoic acid (DHB) matrix (Sigma) and spotted on a stainless-steel plate and air-dried. In each of the tissues, elutes 50%, 70% and 100% showed cyclotide-like masses (2.5–4.0 kDa) and fewer small molecule interferences. These elutes were pooled for each tissue, henceforth referred as "crude extract" and stored in −20 °C for further LC–MS analysis and MS/MS sequencing of peptides.

Sample preparation for LC–MS/MS spectrometry

Reduction, Alkylation: Reduction of disulfide bonds was performed by adding 100 µL of a freshly prepared 0.1 M solution of dithiothreitol (prepared in 0.1 M NH_4HCO_3, pH 8.2) that was added to 100 µL of crude extract of cyclotides, dissolved at a concentration of 5 mg mL^{-1} in 0.1 M NH_4HCO_3 buffer (pH 8.2). The reduction mixture was allowed to incubate for 30 min at 37 °C. The reduced samples were carbamidomethylated for 30 min at 25 °C in the dark by adding 100 µL of 250 mM iodoacetamide (prepared in 0.1 M NH_4HCO_3, pH 8.2) to the mixture and briefly heating to 65 °C for 1 min. The alkylation reaction was stopped by adding 10 µL of 100 mM dithiothreitol. Before LC–MS/MS, samples were desalted using C18 Zip-Tips™ (Millipore).

Enzymatic Digestion: Digestion of the crude extract of cyclotides (after reduction and alkylation) was performed by adding 20 µL of endo Glu-C (400 ng µL^{-1}) (Sigma-Aldrich). All digests were incubated at 37 °C for 12–16 h, quenched with 3% formic acid and stored at −20 °C until further analysis. Before LC–MS/MS, samples were desalted using C_{18} Zip-Tips™ (Millipore).

LC–MS/MS spectrometry

The LC–MS/MS analyses of peptides were performed using LTQ-Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific, SanJose, CA, U.S.A.) Samples were sprayed using an EASY-Spray PepMap RSLC C_{18} column (75 µm × 15 cm, 3 µm particle size, 10 nm pores). The mobile phase was 0.1% formic acid (FA) as buffer A and 80% ACN (with 0.1% FA) as buffer B. A gradient elution was run from 10% to 95% solvent B over 60 min (0–5 min: equilibration time; 5–50 min: 10% B to 65% B; 50–52 min: 65% B to 95% B; 52–60 min: column wash) at a flow rate of 300 nl/min. Full MS scan was done for a mass range between m/z 375 and m/z 2000 on an Orbitrap Mass analyzer with 120 K resolution, followed by HCD-based MS/MS with 60 K resolution, performed in the scan range of m/z 100 and m/z 2000. The mass spectrometer was run in positive, data-dependent, "top speed" MS/MS mode, with two alternating stepped HCD fragmentation energies to optimize fragmentation: 14% ±3% and 25% ±3%. MS/MS was carried out only for peaks above 20,000 intensity (arbitrary units) with an isolation window of 1.2 m/z, an AGC target of 50,000, maximum injection time 118 ms, and 1 microscans. The source voltage was 1,800 V. The acquired data was processed using Thermo Scientific™ Proteome Discoverer™ software version 2.1. MS/MS spectra were searched with the SEQUEST® HT engine against.

Sequencing of cyclotides

Sequences of known cyclotides from *C. ternatea* were retrieved from CyBase database (http://www.cybase.org.au) and a user-defined database was created using the CyBase sequences and our transcriptome derived sequences. The following parameters were used: proteolytic enzyme - GluC; missed cleavages allowed, one; minimum peptide length, 6 amino acids. Carbamidomethylation (+57.021 Da) of cysteine residues was kept as a fixed modification and methionine oxidation (+15.9949 Da) was set as variable modification. The mass tolerances for precursor and fragment ions were kept at 10 ppm and 0.1 Da respectively. Peptides were confidently matched to protein sequences with a maximum false discovery rate of 1% as determined by the Percolator® algorithm. Protein groups were further identified with a criterion of having the presence of at least two unique peptides.

Acknowledgements

The authors thank Prof. R. Sowdhamini (National Centre for Biological Sciences, Bangalore, India) for her valuable suggestions throughout the project. The authors acknowledge the help provided by the Proteomics Facilities at NCBS and the Indian Institute of Science. NK was supported by the Tata Education and Development Trust. PB acknowledges his support by the DST-YOS Chair Professorship awarded by the SERB, Department of Science and Technology, Government of India. Support by SERB, Department of Science and Technology (Early Career Award, Ramanujan Fellowship), Max Planck Partner group program and Department of Biotechnology to RV is acknowledged.

Conflict of Interest

The authors declare no conflict of interest.

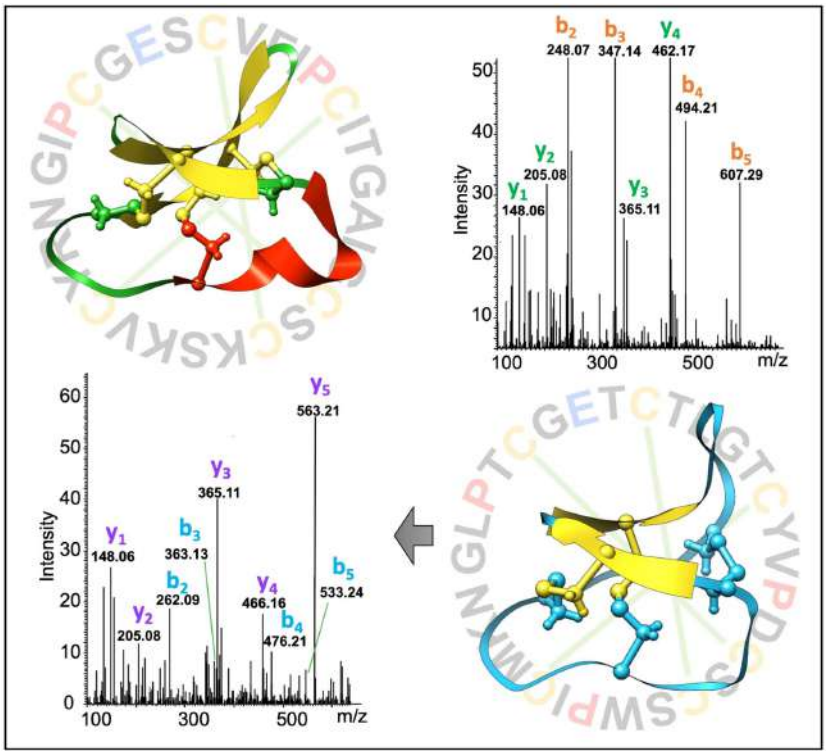
Keywords: Mass spectrometry · cyclotides · MS/MS fragmentation pattern · *Clitoria ternatea* · X-Pro bond cleavage

[1] D. J. Craik, N. L. Daly, T. Bond, C. Waine, *J. Mol. Biol.* **1999**, *294*, 1327–1336.
[2] C. Jennings, J. West, C. Waine, D. Craik, M. Anderson, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10614–10619.
[3] K. J. Rosengren, N. L. Daly, M. R. Plan, C. Waine, D. J. Craik, *J. Biol. Chem.* **2003**, *278*, 8606–8616.
[4] G. K. T. Nguyen, S. Wang, Y. Qiu, X. Hemu, Y. Lian, J. P. Tam, *Nat. Chem. Biol.* **2014**, *10*, 732–738.
[5] M. L. Colgrave, D. J. Craik, *Biochemistry* **2004**, *43*, 5965–5975.
[6] R. Burman, A. Herrmann, R. Tran, J. E. Kivelä, A. Lomize, J. Gullbo, U. Göransson, *Org. Biomol. Chem.* **2011**, *9*, 4306–4314.
[7] M. A. Esmaeili, N. Abagheri-Mahabadi, H. Hashempour, M. Farhadpour, C. W. Gruber, A. Ghassempour, *Fitoterapia* **2016**, *109*, 162–168.
[8] C. K. L. Wang, M. L. Colgrave, K. R. Gustafson, D. C. Ireland, U. Goransson, D. J. Craik, *J. Nat. Prod.* **2008**, *71*, 47–52.
[9] J. P. Tam, Y. A. Lu, J. L. Yang, K. W. Chiu, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 8913–8918.
[10] M. L. Colgrave, A. C. Kotze, D. C. Ireland, C. K. Wang, D. J. Craik, *ChemBioChem* **2008**, *9*, 1939–1945.
[11] L. Gran, *Acta Pharmacol. Toxicol. (Copenh).* **1973**, *33*, 400–408.
[12] K. N. T. Nguyen, G. K. T. Nguyen, P. Q. T. Nguyen, K. H. Ang, P. C. Dedon, J. P. Tam, *FEBS J.* **2016**, *283*, 2067–2090.
[13] T. Grover, R. Mishra, Bushra, P. Gulati, A. Mohanty, *Peptides* **2021**, *135*, 170430.
[14] G. K. Oguis, E. K. Gilding, Y. H. Huang, A. G. Poth, M. A. Jackson, D. J. Craik, *Ind. Crops Prod.* **2020**, *147*, 112214.
[15] P. Claeson, U. Göransson, S. Johansson, T. Luijendijk, L. Bohlin, *J. Nat. Prod.* **1998**, *61*, 77–81.
[16] A. G. Poth, J. S. Mylne, J. Grassl, R. E. Lyons, A. H. Millar, M. L. Colgrave, D. J. Craik, *J. Biol. Chem.* **2012**, *287*, 27033–27046.
[17] A. G. Poth, M. L. Colgrave, R. E. Lyons, N. L. Dalya, D. J. Craik, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 10127–10132.
[18] J. F. Hernandez, J. Gagnon, L. Chiche, T. M. Nguyen, J. P. Andrieu, A. Heitz, T. T. Hong, T. T. C. Pham, D. Le Nguyen, *Biochemistry* **2000**, *39*, 5722–5730.
[19] N. L. Daly, K. J. Rosengren, D. J. Craik, *Adv. Drug Delivery Rev.* **2009**, *61*, 918–930.
[20] A. Herrmann, E. Svangård, P. Claeson, J. Gullbo, L. Bohlin, U. Göransson, *Cell. Mol. Life Sci.* **2006**, *63*, 235–245.
[21] A. Heitz, J. F. Hernandez, J. Gagnon, T. T. Hong, T. Châu Pham, T. M. Nguyen, D. Le-Nguyen, L. Chiche, *Biochemistry* **2001**, *40*, 7973–7983.
[22] E. K. Gilding, M. A. Jackson, A. G. Poth, S. T. Henriques, P. J. Prentis, T. Mahatmanto, D. J. Craik, *New Phytol.* **2016**, *210*, 717–730.

[23] G. K. Oguis, E. K. Gilding, M. A. Jackson, D. J. Craik, *Front. Plant Sci.* **2019**, *10*, 645.
[24] A. Serra, X. Hemu, G. K. T. Nguyen, N. T. K. Nguyen, S. K. Sze, J. P. Tam, *Sci. Rep.* **2016**, *6*, 23005.
[25] A. G. Poth, M. L. Colgrave, R. Philip, B. Kerenga, N. L. Daly, M. A. Anderson, D. J. Craik, *ACS Chem. Biol.* **2011**, *6*, 345–355.
[26] G. K. T. Nguyen, S. Zhang, N. T. K. Nguyen, P. Q. T. Nguyen, M. S. Chiu, A. Hardjojo, J. P. Tam, *J. Biol. Chem.* **2011**, *286*, 24275–24287.
[27] N. V. Kalmankar, R. Venkatesan, P. Balaram, R. Sowdhamini, *Sci. Rep.* **2020**, *10*, 12658.
[28] C. K. L. Wang, Q. Kaas, L. Chiche, D. J. Craik, *Nucleic Acids Res.* **2008**, *36*, D206–D210.
[29] J. P. Mulvenna, C. Wang, D. J. Craik, *Nucleic Acids Res.* **2006**, *34*, D192–D194.
[30] M. L. Colgrave, A. G. Poth, Q. Kaas, D. J. Craik, *Biopolymers* **2010**, *94*, 592–601.
[31] N. C. Parsley, O. L. Williams, L. M. Hicks, *J. Am. Soc. Mass Spectrom.* **2020**, *31*, 1833–1843.
[32] T. Vaisar, J. Urban, *J. Mass Spectrom.* **1996**, *31*, 1185–1187.
[33] L. A. Breci, D. L. Tabb, J. R. Yates, V. H. Wysocki, *Anal. Chem.* **2003**, *75*, 1963–1971.
[34] R. N. Grewal, H. El Aribi, A. G. Harrison, K. W. M. Siu, A. C. Hopkinson, *J. Phys. Chem. B* **2004**, *108*, 4899–4908.
[35] R. Hellinger, J. Koehbach, D. E. Soltis, E. J. Carpenter, G. K. S. Wong, C. W. Gruber, *J. Proteome Res.* **2015**, *14*, 4851–4862.
[36] M. Narayani, R. Babu, A. Chadha, S. Srivastava, *Phytochem. Rev.* **2020**, *19*, 787–825.
[37] C. K. L. Wang, R. J. Clark, P. J. Harvey, K. Johan Rosengren, M. Cemazar, D. J. Craik, *Biochemistry* **2011**, *50*, 4077–4086.
[38] N. L. Daly, R. J. Clark, M. R. Plan, D. J. Craik, *Biochem. J.* **2006**, *393*, 619–626.
[39] J. P. Mulvenna, L. Sando, D. J. Craik, *Structure* **2005**, *13*, 691–701.

Manuscript received: May 31, 2021
Revised manuscript received: July 14, 2021
Accepted manuscript online: July 20, 2021
Version of record online: ■■, ■■■

FULL PAPER



Mass spectral fragmentation patterns of cyclotides are determined by the number of internal proline residues, permitting a quick distinction between the Möbius and Bracelet structural classes in complex natural peptide libraries.

N. V. Kalmankar, P. Balaram*, R. Venkatesan**

Manuscript 5

Proteome-wide discovery of macrocyclic cysteine knot peptides: Identification of cyclo-oligomerized cyclotides from butterfly pea

Neha V. Kalmankar, Ramanathan Sowdhamini, Radhika Venkatesan and Padmanabhan Balaram*

3.3 ABSTRACT

Macrocyclic peptides show diverse biological activities and have become a promising class of drugs. Among the different members of this class, backbone cyclic peptide macrocycles are the most contrained structures that improves the stability of the peptides significantly. Cyclotides represent a category of naturally occuring, ribosomally synthesized, ultra-stable macrocyclic peptides (~30 residues size) that are ligated by a specialized peptidyl asparaginyl ligase (PAL). This ligase belongs to C13 subfamily of cysteine proteases. Butelase 1 is a PAL, first discovered in *Clitoria ternatea* (butterfly pea) plant, that catalyzes the peptide bond formation at an Asx residue in cyclotides. Butelase 1 has been identified to efficiently catalyze with broad specificity, both intra- and intermolecular macrocylization of peptide subtrates containing specifically Asn residue follwed by a 'His-Val' tail at the C-terminal end. In the present study, we discovered cyclic mono-, di-, trimeric forms of cyclotide precursors containing three disulfide-bonds and a head-to-tail macrocyclic backbone coming from multiple tissue extracts of *C. ternatea*. Our work shows that butelase 1 mediates cyclodimerization and trimerization of cyclotides at high substrate concentration. The cyclodimers and cyclotrimers were favored for certain sequences and present in low abundance compared to their corresponding cyclomonomers. With over >100 'cliotides' produced in *C. ternatea*, butelase 1 seems to act as a promiscuous enzyme with broad substrate specificity, therefore making them as valuable biotechnological tools for peptide and protein chemists.

3.4 INTRODUCTION

"Macrocyclic peptides" are produced in both prokaryotes and eukaryotes, and represent a remarkably diverse group of molecules. As the name suggests, they are can be defined as ring-like peptidic structures spanning multiple amino acid residues. Based on the type of cyclization, they can be either sidechain-to-sidechain, head-to-sidechain, and head-to-tail

(backbone) cycles. Among these, the backbone cyclized structures are conformationally the most constrained and rigid type due to the unrotatable, consecutive, unsaturated amide bonds. Naturally occurring backbone macrocyclic peptides are mature products of their linear precursos, predominantly mediated by an enzymatic peptide backbone cyclization process by specific cyclases. One such example are the cyclotides, which are ~30 residue long backbone cyclic gene products emerging from the linear precusors by the action of a specialized asparaginyl endopeptidase (AEP). Cyclotides belong to the class of ribosomally-derived, head-to-tail backbone, macrocyclic peptides that function as the host defense agents in plants. With their rigid structure and enhanced stability, they have become attractive candidates for drug development featuring as molecular grafting scaffolds.

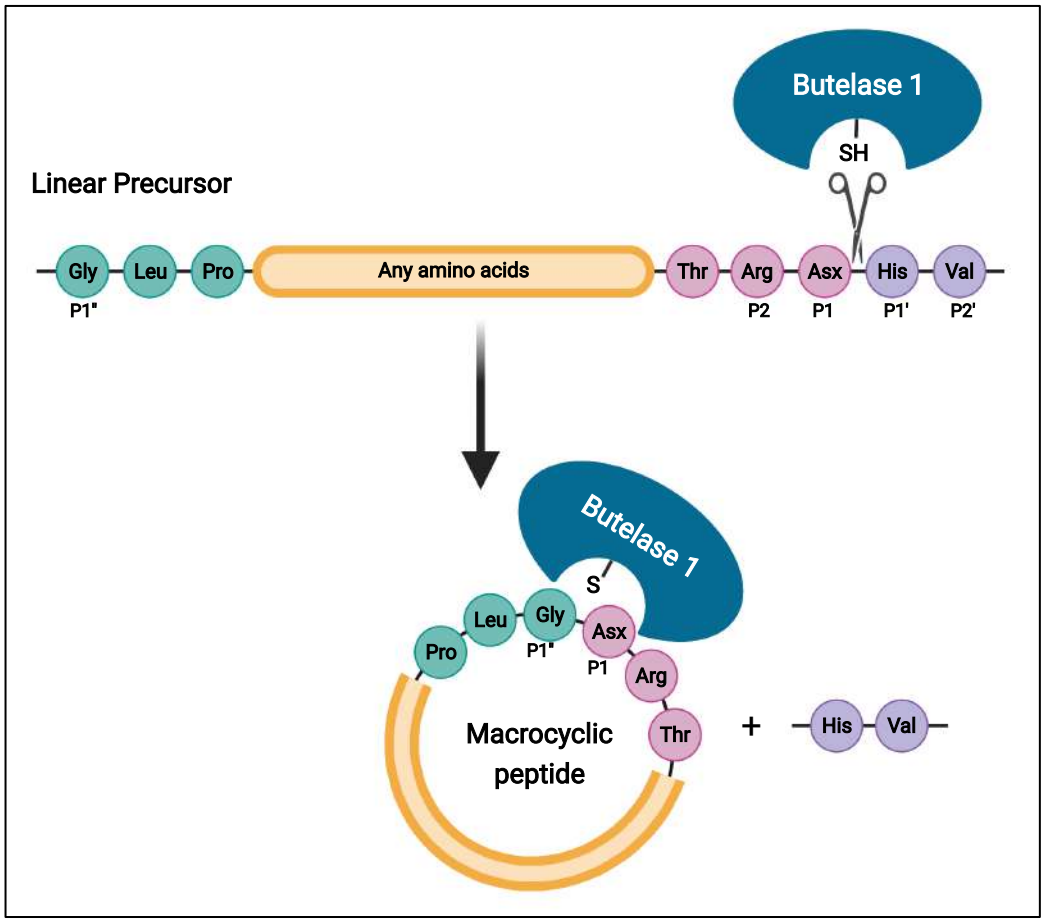


Figure 3.1: Schematic illustration of peptide macrocyclization of by butelase 1.

Butelase 1, a peptide ligase or AEP, isolated from *Clitoria ternatea,* has been discovered to be responsible for the backbone cyclization of cyclotides at an Asx (Asp/Asn) residue ligation site (Nguyen et al., 2014). Butelase 1 recognizes the C-terminal 'Asx-His-Val' sorting signal motif and cleaves between the Asx-His segment and releases His-Val segment from the peptide to form a thioester acyl-enzyme intermediate (**Figure 3.1**). Butelase 1 is known to be the fastest peptide ligase with catalytic efficiencies as high

as 1.3 x 10^6 $M^{-1}s^{-1}$ (Nguyen et al., 2015).

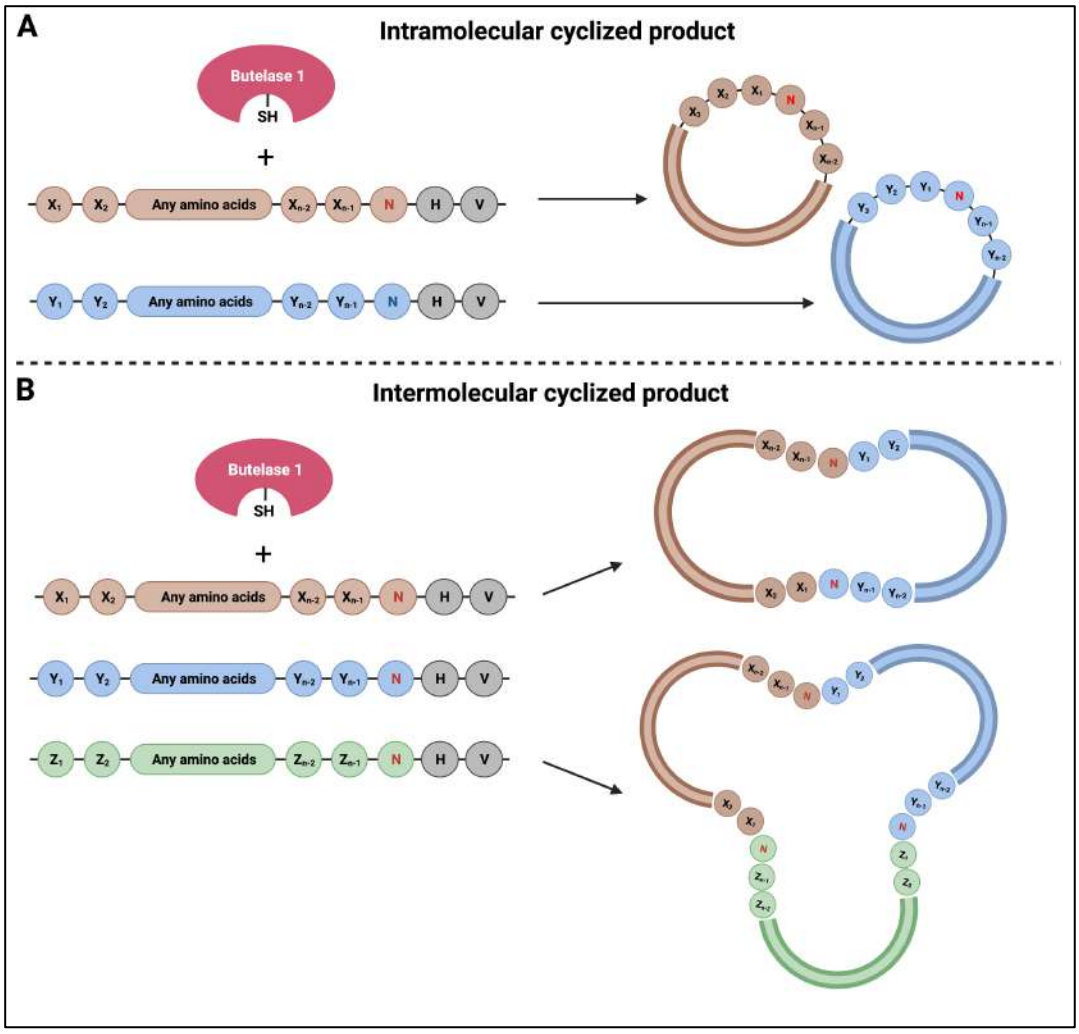


Figure 3.2: Overview of butelase-1 mediated intra- and intermolecular cyclo-oligmerization reactions.

Among membrane pore-forming peptides, self-association is a commonly observed phenomenon. However, this can become challenging to ascertain when working with disulfide-rich peptides that may form non-native disulfide-bond isoforms. For example, theoretically a peptide with three or four disulfide bonds could form 15 or 105 different disulfide-bond isomers, respectively. On the other hand, self-association can be an outcome of an enzymatic process occuring at non-Cys positions along the polypeptide chain. For example, in a study by Nguyen et al., it was observed that butelase 1 mediated cyclodimerization of kalata B1 at high substrate concentrations (>400 μM) (Nguyen et al., 2014). The mass-spectrometric analysis revealed a peculiar peak having a molecular mass of 5,782 Da (monomeric mass M=2891 Da). This work established for the first time that butelase 1 is able to perform intermolecular ligation and cyclization of cysteine-rich peptides of 2-3kDa size (**Figure 3.2**). Soon after it was demonstrated that butelase 1

cyclizes several non-cysteine containing proteins with >200 residues with excellent efficiency (Nguyen et al., 2015). More recently, butelase 1 was reported to efficiently cyclo-oligomerize RLYR-containing antimicrobial peptide in one-step ligase-controlled reaction (Hemu et al., 2019). This reaction yielded cyclic mono-, di-, tri-, and tetramers from peptide precursors containing 3-15 amino acids which specially ended with Asn and a His-Val tail. The observation of distinct dimers and trimers in our current study indicates that cyclotides are capable of undergoing a specific oligomerization process, and this could be is vital for cyclotide bioactivities. Of the cyclotides sequenced from *C. ternatea* through mass spectrometry, few novel homo-dimeric and homo-trimeric peptides were observed and characterized.

3.5 RESULTS

3.5.1 Cyclotide Screening in *C. ternatea*

Aerial and non-aerial parts of *C. ternatea* including leaf, stems, flower, pods and roots were collected, divided and dried, and was screened for the presence of cyclotides using the procedure described in **Figure 3.3**. The initial aqueous extracts of *C. ternatea* tissues was prepared by maceration and liquid/liquid phase separation followed by pre-purification using C_{18} silica-gel column chromatography to remove polar plant constituents. Each of the elutes from the column chromatography were subjected to MALDI-TOF/MS for screening of cyclotide-like masses (2500-4000 Da). Fractions 50, 70 and 100% CAN comprised of cyclotide-like masses (Figure S3.1). These elutes were pooled together and lyophilized to form the "crude extract". The crude extract was then subjected to several rounds of purification using semi-preparative RP-HPLC across a linear acetonitrile gradient. Five fractions were collected for each tissue (fractions A-E) (**Figure 3.4**). Each of the five fractions arising from five plant tissues were monitored by MALDI-TOF/MS for the presence of cyclotide-like masses (**Figure 3.5**). Cyclotide-like peptides were found in four out of five fractions i.e. B-E, eluting between 30-60% ACN, suggesting that the corresponding peptides were mostly hydrophobic in nature. This observation is consistent with previous studies in which cyclotides typically eluted in 25−55% ACN. Fractions B-E were combined to give the "semi-purified extract" for each of the five plant tissues.

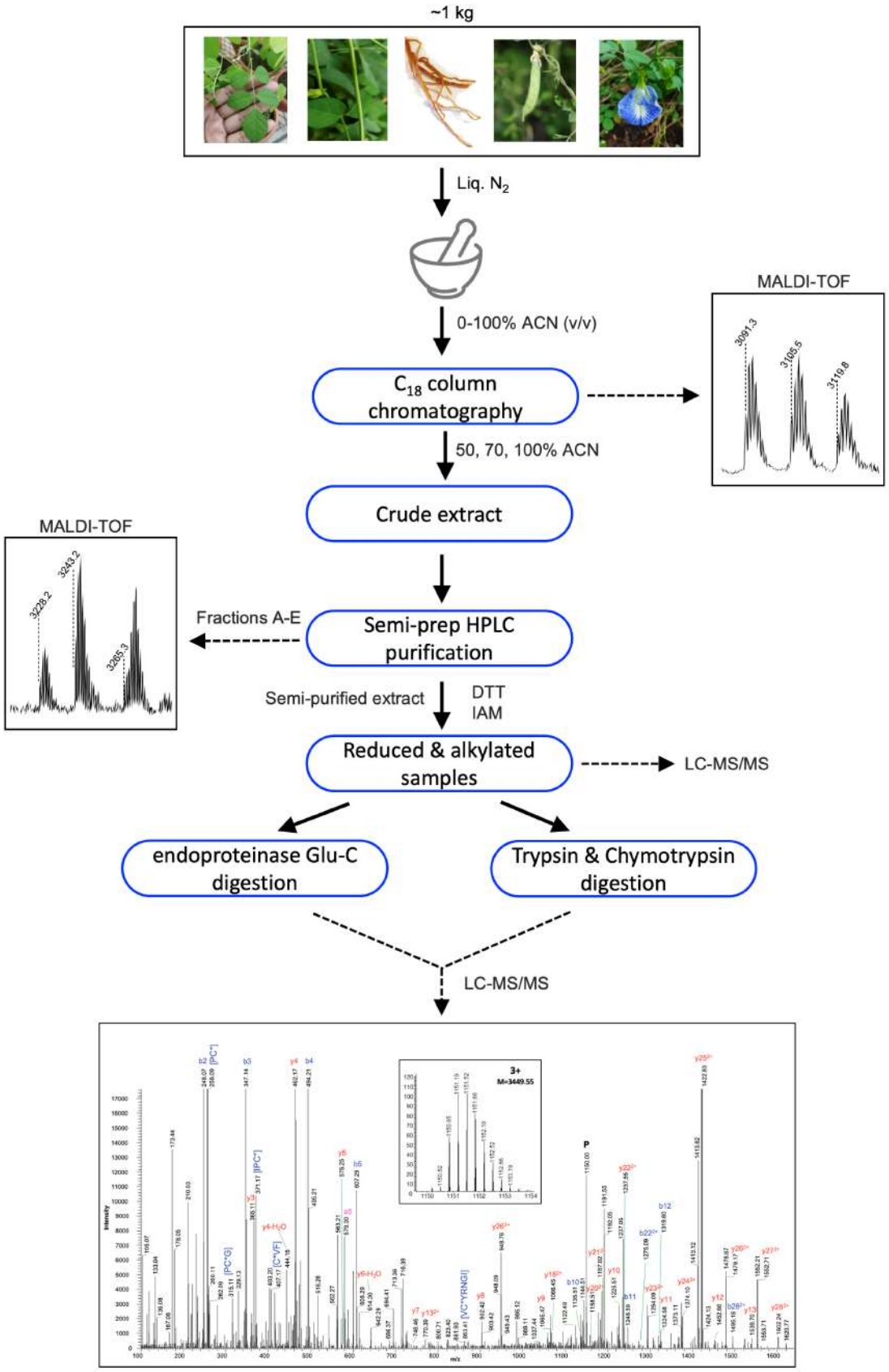


Figure 3.3: Flowchart illustrating the processes of pre-screening for the presence of cyclotides, purification for enrichment of cyclotides and sequencing of individual cyclotides in multiple tissues of *C. ternatea*.

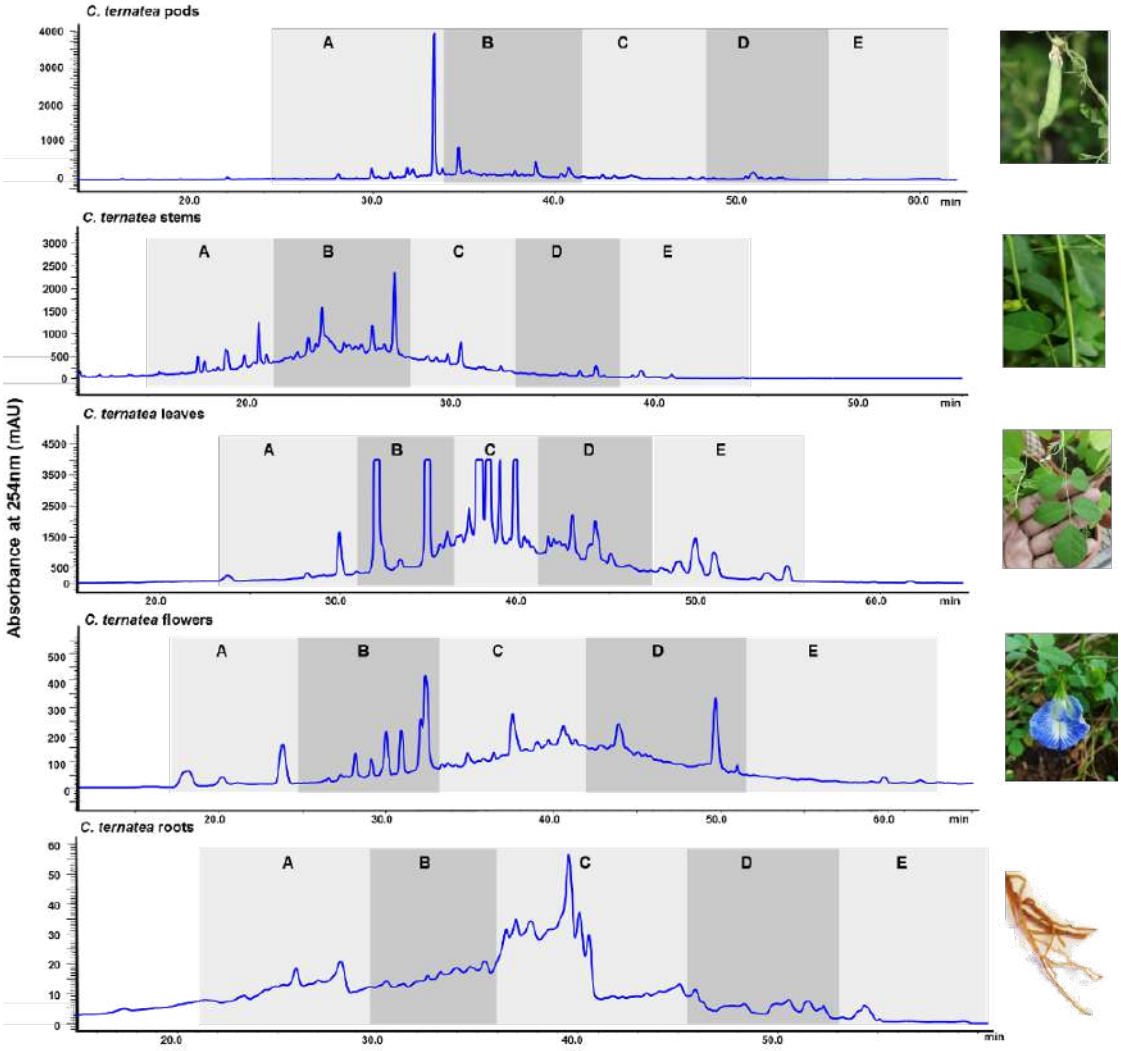


Figure 3.4: UV chromatograms of *C. ternatea* cyclotide crude extracts. Reversed-phase liquid chromatography (RP-HPLC) profiles of pods, stems, leaves, flowers and roots crude extract with magnification of cyclotide fractions (A-E) performed using linear gradients of acetonitrile.

3.5.2 Peptidomic characterization of cyclotides

From the transcriptome stuudy of cyclotides from *C. ternatea*, we understood that cyclotides are expressed in a tissue specific manner (please see Chapter 2 of this thesis). To test this observation at the protein level, we subjected semi-purified extracts of different tissues including that of leaves, pods, flowers, stems, and roots of *C. ternatea* plants to LC-MS for the detection of known and novel cyclotides. Direct sequencing of the cyclotides is not easy and straightforward due to the circular backbone and CCK motif. Therefore, the native cyclotides were chemically modified to yield S-carbamidomethylation of cysteines. After reduction with dithiothreitol, alkylation with iodoacetamide, and digestion of the semi-purified tissue extracts with endoproteinase Glu-C, all peptide-related mass signals were monitored by LC-MS/MS. The presence of cyclotide-like molecules was determined

by a mass shift of 348 Da, which corresponded to the reduction and alkylation of six cysteine residues. Digestion using endoproteinase Glu-C resulted in a mass shift of 18 Da, corresponding to the hydrolysis and enzymatic ring-opening of a backbone cyclized peptide, and further amenable to fragmentation by MS/MS. A suite of cyclotides were obtained and the peptidomic expression of the cyclotides varied depending on the type of tissue (please refer to Figure 2 of Manuscript 3 on page 54).

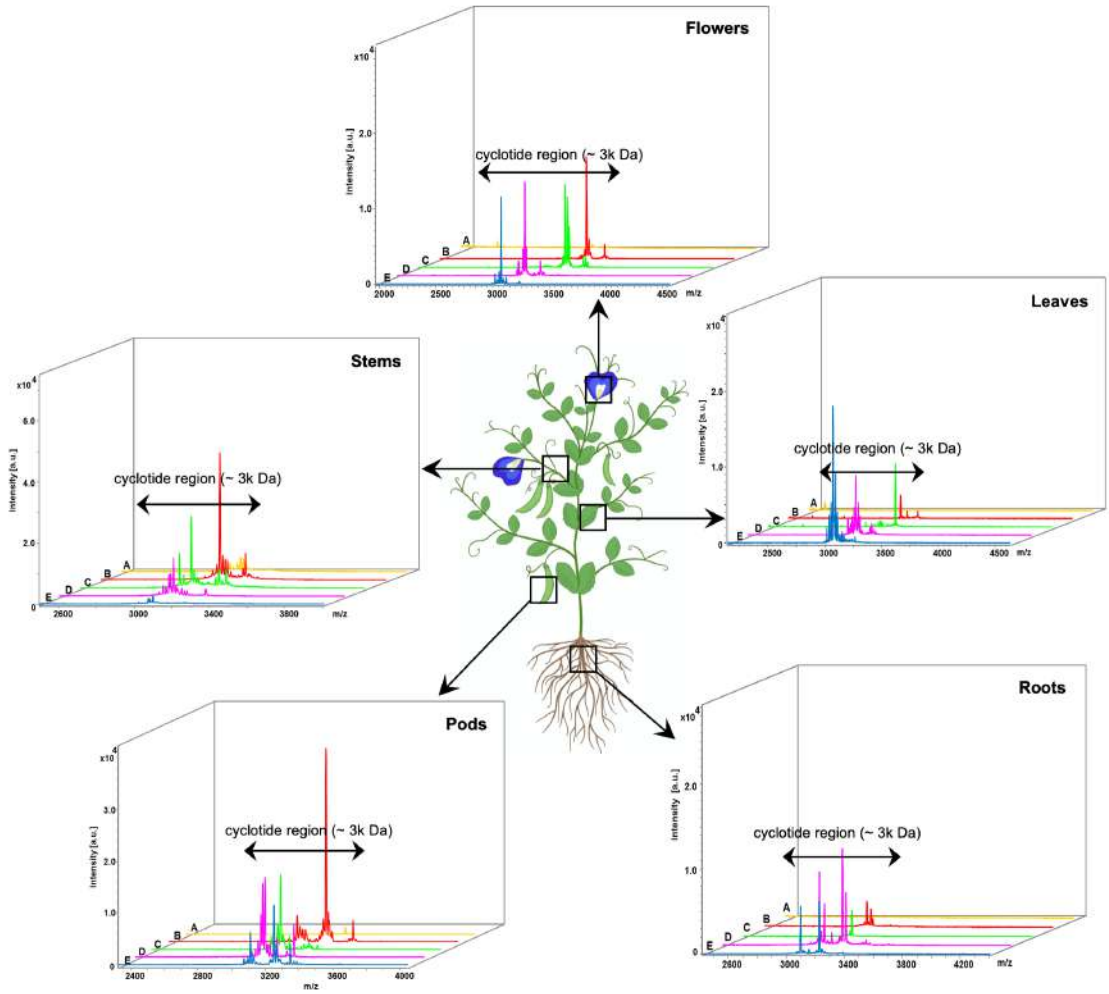


Figure 3.5: Comparative stacked 3D-plot of MALDI-TOF mass spectra of HPLC fractions A-E from five tissues of *C. ternatea* showing the cyclotide region (mass ~3kDa).

Tandem mass-spectrometry can be performed by two approaches: database search and de novo sequencing. In a database search approach, the raw mass spectral data is submitted to a search engine to find a match with a known cyclotide sequence present in a user-defined database. However, this approach is not feasible for unknown cyclotides, in which case de novo sequencing approach is required. We performed both database search and de novo sequencing method in our study. For the database search, we first created a custom database containing full length cyclotide sequences obtained from the CyBase

webserver and our transcriptome derived sequences. MS/MS spectra obtained for each tissue extract from the LC-MS/MS spectrometry experiments were searched against this custom database using SEQUEST® HT search engine from the Thermo Proteome Discoverer v2.1 software. A total of 347 complete sequences were identfied by Thermo Proteome Discoverer v2.1. The highest number of cyclotides were identified from the stems extract (120), followed by pods (109), leaves (63), roots (35) and flower (20). Table S3.2 (Supplementary Information) describes the top 10 known cyclotides hits identified from five tissues of *C. ternatea* from the database search approach.

3.5.3 Discovery of novel arrangements of cyclotides

The first experimentally straightforward evidence on the ligation property of AEP was demonstrated using *C. ternatea* pod fractions. It was termed butelase-1 after the local name of plant and was shown to be a highly efficient enzyme in intra- and intermolecular peptide bond formation, even on non-native-targets. Of the cyclotides sequenced from *C. ternatea* through mass spectrometric, few novel cyclodimeric and cyclotrimeric peptides were observed and characterized. From the pods extract, peak 21 contained monomeric (monoisotopic mass M_1=3073.22 Da; m/z=1025.41 (3+ charge state)) and homo-trimeric (cyclotrimer; monoisotopic mass M_3=9220.60 Da; m/z=1845.12 (5+ charge state)) species (**Figure 3.6**). This could correspond to the Cter M cyclotide (molecular mass=3057.22), with its lone methionine residue in an oxidized state.

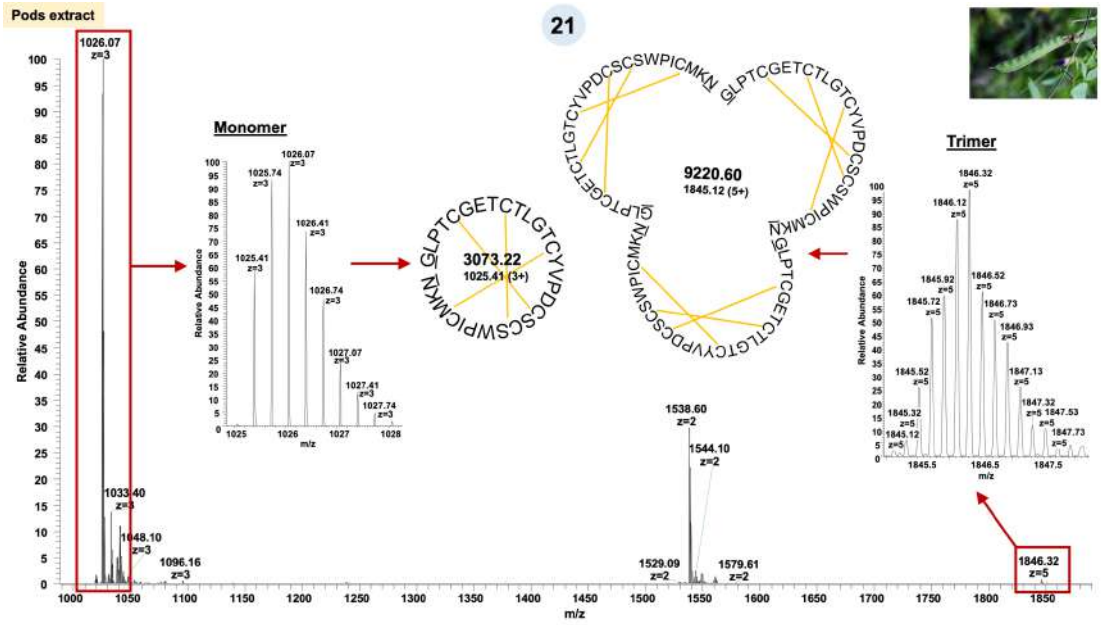


Figure 3.6: Molecular ions corresponding to a peptide extracted from pods of *C. ternatea*, existing as a monomer and cyclotrimer. The peptide sequence, predicted to be Cter M with its methionine in oxidized state, is shown as a circular representation with the cyclization site underlined. Distribution of 3+ and 5+ charge states are shown in the inset.

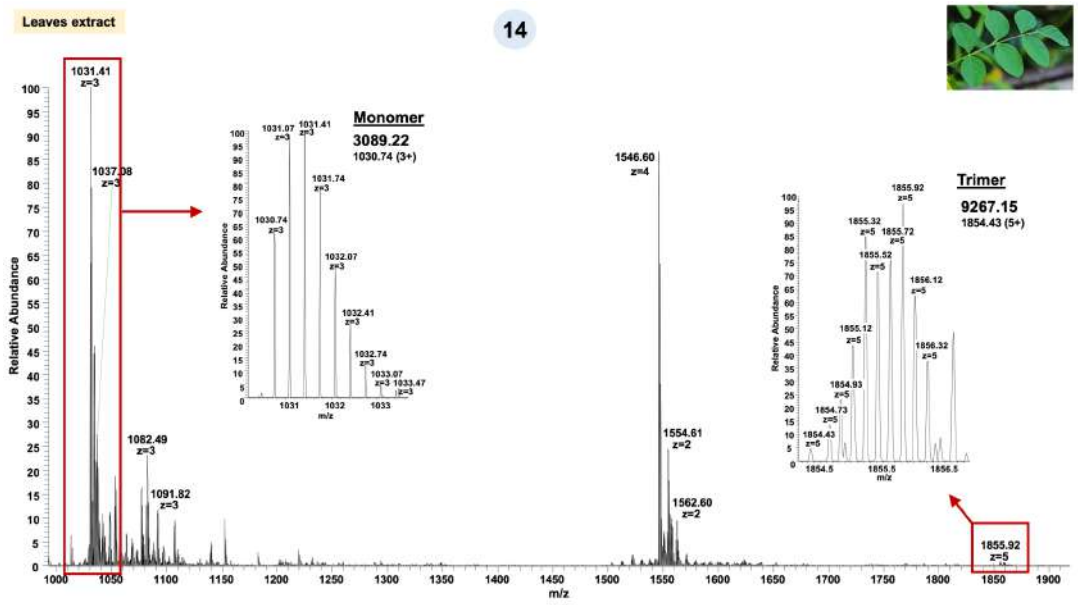


Figure 3.7: Molecular ions corresponding to a novel peptide extracted from leaves of *C. ternatea*, existing as a monomer and cyclotrimer. Distribution of 3+ and 5+ charge states are shown in the inset.

Similar to the pods extract, the leaves extract, peak 14 showed monomeric (monoisotopic masses M_1=3089.20 Da; m/z=1030.74 (3+ charge state)) and homo-trimeric (cyclotrimer; monoisotopic masses M_3=9267.15 Da; m/z=1854.43 (5+ charge state)) species were observed (**Figure 3.7**). However, this mass did not correspond to any known masses of cyclotides, hence represents a novel cyclotide sequence.

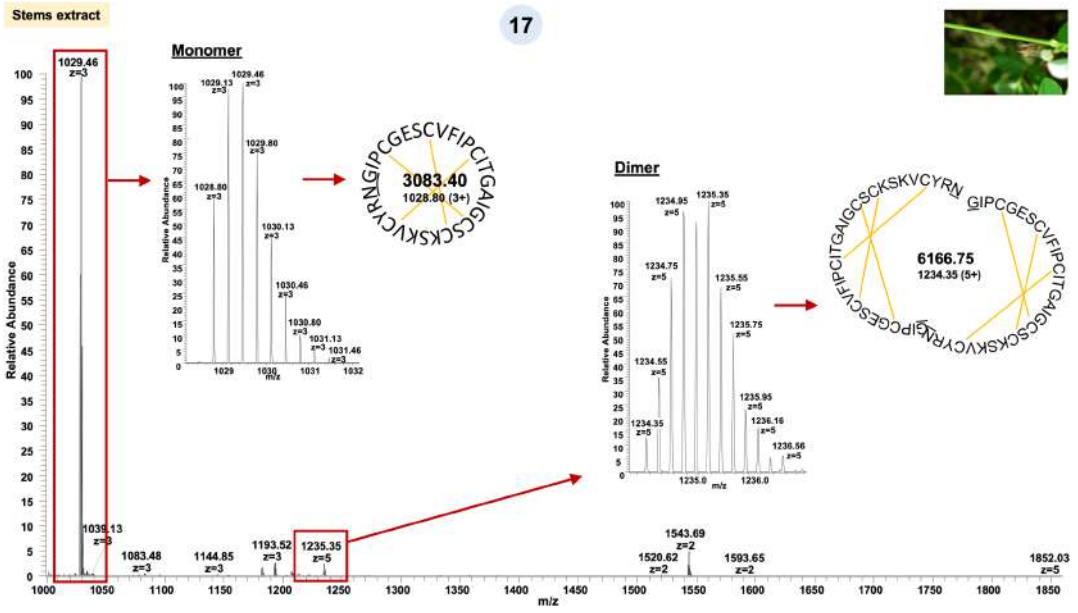


Figure 3.8: Molecular ions corresponding to a peptide extracted from stems of *C. ternatea*, existing as a monomer and cyclodimer. The peptide sequence, predicted to be Cliotide T1, is shown as a circular representation with the cyclization site underlined. Distribution of

3+ and 5+ charge states are shown in the inset.

Monomeric (monoisotopic masses M_1=3083.37 Da; m/z=1028.80 (3+ charge state)) and homo-dimeric (cyclodimer; monoisotopic masses M_2=6166.75 Da; m/z=1234.35 (5+ charge state)) species was also observed from peak 17 of the stems extract (**Figure 3.8**). However, cyclotrimeric masses were absent in this tissue extract. This corresponds to the Clitotide T1 cyclotide (molecular mass M=3083.40).

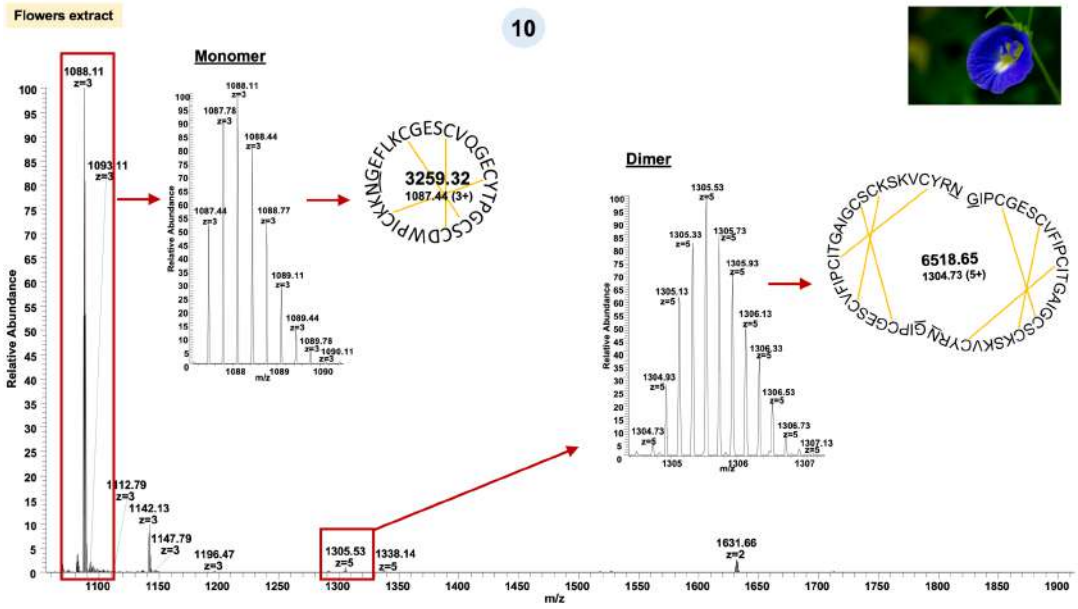


Figure 3.9: Molecular ions corresponding to a peptide extracted from flowers of *C. ternatea*, existing as a monomer and cyclodimer. The peptide sequence, predicted to be Cliotide T2, is shown as a circular representation with the cyclization site underlined. Distribution of 3+ and 5+ charge states are shown in the inset.

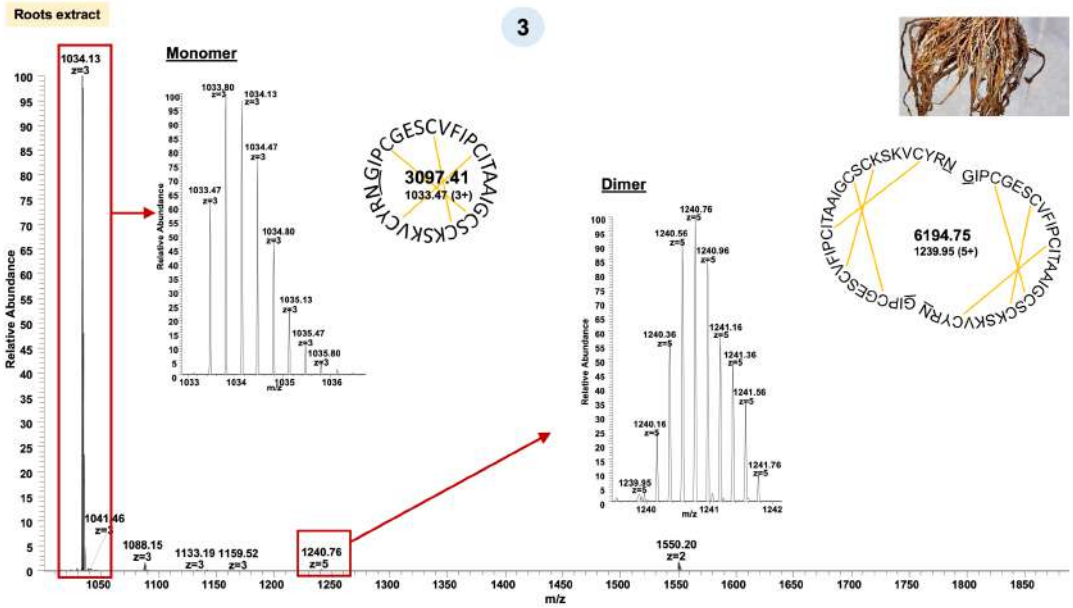


Figure 3.10: Molecular ions corresponding to a peptide extracted from roots of *C. ternatea*,

existing as a monomer and cyclodimer. The peptide sequence, predicted to be Cter P, is shown as a circular representation with the cyclization site underlined. Distribution of 3+ and 5+ charge states are shown in the inset.

Peak 10 from the flower extract contained monomeric (monoisotopic mass M_1=3259.32 Da; m/z=1087.44 (3+ charge state)) and homo-dimeric (cyclodimer; monoisotopic mass M_2=6518.65 Da; m/z=1304.73 (5+ charge state)) species (**Figure 3.9**). This could possibly correspond to Cliotide T2. Similarly, peak 3 from the root extract, showed monomeric (monoisotopic mass M_1=3097.41 Da; m/z=1033.47 (3+ charge state)) and homo-dimeric (cyclodimer; monoisotopic mass M_2=6194.75 Da; m/z=1239.95 (5+ charge state)) species, corresponding to Cter P cyclotide (**Figure 3.10**).

3.6 DISCUSSION AND CONCLUSIONS

Asparaginyl endopeptidases (AEPs) are a class of endo/lysosomal cysteine proteases that recognizes asparagine residue at the P1 site for peptide bond hydrolysis. Many plant specific AEPs have been identified that target seed storage precursors and cleave them to form mature seed storage proteins. It was only recently uncovered that AEPs not only perform cleavage mechanisms, but also has efficient enzymatic capability of ligating linear peptides, therefore also referred to as peptidyl asparaginyl ligases (PAL) (Hemu et al., 2020; Nguyen et al., 2014, 2015). Plants that lacked this specific ligase-type enzyme produced acyclic or linear versions of cyclotide gene products. Traditionally AEPs perform peptide bond cleavage by recognizing the peptide substrate at an Asn residue site and cleaves to form an acyl-enzyme thioester intermediate. In the case of such rare ligase-type AEPs, the N-terminal amine of a precursor peptide acts as a nucleophile to resolve the thioester intermediate, resulting in peptide hydrolysis or ligation. If the nucleophilic amine is from the same peptide substrate, then AEP performs intramolecular ligation and if from a different peptide substrate, resulting in an intermolecular ligation. Butelase 1 was shown to perform intermolecular ligation followed by head-to-tail macrocyclisation resulting in the formation of cyclo-oligomers of several cysteine-rich peptides of diverse lengths (Hemu et al., 2019). Here, we describe the discovery of higher-order oligomeric arrangements of cyclotides from *C. ternatea* tissues established through high-resolution mass-spectrometry. Our study provides fresh evidence that butelase 1, a *C. ternatea* specific AEP responsible for cyclizing native cyclotides in this Fabaceae plant, is likely mediating intermolecular peptide ligation, apart from intramolecular ligation. The observation of distinct dimers and trimers for certain cyclotide sequences in indicates that cyclotides are undergoing a specific oligomerization process at Asx site present at the C-terminal end. Peptide ligases are valuable biochemical tools that occur rarely in nature and butelase 1 holds promise in the

field of protein engineering for its newly discovered ability of chemoenzymatic synthesis of macrocyclic peptides and proteins.

3.7 EXPERIMENTAL SECTION

3.7.1 Extraction of cyclotides from *C. ternatea*

For extraction procedures, please refer to Experimental Section of Manuscript 3 on page 61 of this chapter. **Figure 3.3** illustrates the pre-screening and purification protocol for enrichment of cyclotides, and sequencing of individual cyclotides from multiple tissues of *C. ternatea*.

3.7.2 Purification and MALDI-TOF analysis

HPLC purification of crude extracts was performed using a semi-preparative Phenomenex Proteo C18 column (250 × 10 mm, 10 μm, 110 Å) at flow-rate of 3 ml min^{-1} on a Shimadzu Prominence series equipped with a binary pump, autosampler, PDA detector, and fraction collector. A linear gradient of 1% min^{-1} of 0–95% buffer B (100% acetonitrile, 0.1% trifluoroacetic acid) was applied, and the eluents were monitored at 220, 254, and 280 nm. Early and late-eluting peaks were separated into five fractions (A-E), manually collected for each plant tissue and lyophilized (**Figure 3.4**). Each of the five fractions, per tissue, was subjected to MALDI-TOF mass-spectrometry (UltraFlex Bruker Daltonics) analysis with 50 Hz pulsed nitrogen laser (λ=337 nm), in positive ion reflectron mode. The samples were prepared in 1:1 mixture of H_2O-ACN with 0.1% TFA by mixing an equal amount of peptide fractions (0.5 μl) with α-cyano-4-hydroxycinnamic acid (CCA)/2,5-dihydroxy benzoic acid (DHB) matrix (Sigma) and spotted on a stainless-steel plate and air-dried. In each of the tissue crude extracts, fractions B-E showed cyclotide-like masses (2.5-4.0 kDa) (**Figure 3.5**). These fractions were pooled for each tissue, henceforth referred as "semi-purified extract" and stored in -20 °C for further LC-MS/MS analysis and peptide sequencing.

3.7.3 LC-MS and MS/MS sequencing of cyclotides

For LC-MS/MS spectrometry and sequencing procedure of semi-purified extracts, please refer to Experimental Section of Manuscript 3 on page 62 of this chapter.

3.8 REFERENCES OF CHAPTER 3 (MANUSCRIPT 5)

Breci, L. A., Tabb, D. L., Yates, J. R., and Wysocki, V. H. (2003). Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. *Anal. Chem.* 75. doi:10.1021/ac026359i.

Burman, R., Herrmann, A., Tran, R., Kivelä, J. E., Lomize, A., Gullbo, J., et al. (2011). Cytotoxic potency of small macrocyclic knot proteins: Structure-activity and mechanistic studies of native and chemically modified cyclotides. *Org. Biomol. Chem.* 9. doi:10.1039/c0ob00966k.

Claeson, P., Göransson, U., Johansson, S., Luijendijk, T., and Bohlin, L. (1998). Fractionation protocol for the isolation of polypeptides from plant biomass. *J. Nat. Prod.* 61. doi:10.1021/np970342r.

Colgrave, M. L., and Craik, D. J. (2004). Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: The importance of the cyclic cystine knot. *Biochemistry* 43, 5965–5975. doi:10.1021/bi049711q.

Colgrave, M. L., Kotze, A. C., Ireland, D. C., Wang, C. K., and Craik, D. J. (2008). The anthelmintic activity of the cyclotides: Natural variants with enhanced activity. *ChemBioChem* 9. doi:10.1002/cbic.200800174.

Colgrave, M. L., Poth, A. G., Kaas, Q., and Craik, D. J. (2010). A new "era" for cyclotide sequencing. *Biopolymers* 94, 592–601. doi:10.1002/bip.21400.

Craik, D. J., Daly, N. L., Bond, T., and Waine, C. (1999). Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J. Mol. Biol.* 294, 1327–1336. doi:10.1006/jmbi.1999.3383.

Daly, N. L., Clark, R. J., Plan, M. R., and Craik, D. J. (2006). Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif. *Biochem. J.* 393, 619–626. doi:10.1042/BJ20051371.

Daly, N. L., Rosengren, K. J., and Craik, D. J. (2009). Discovery, structure and biological activities of cyclotides. *Adv. Drug Deliv. Rev.* 61, 918–930. doi:10.1016/j.addr.2009.05.003.

Esmaeili, M. A., Abagheri-Mahabadi, N., Hashempour, H., Farhadpour, M., Gruber, C. W., and Ghassempour, A. (2016). Viola plant cyclotide vigno 5 induces mitochondria-mediated apoptosis via cytochrome C release and caspases activation in cervical cancer cells. *Fitoterapia* 109. doi:10.1016/j.fitote.2015.12.021.

Gilding, E. K., Jackson, M. A., Poth, A. G., Henriques, S. T., Prentis, P. J., Mahatmanto, T., et al. (2016). Gene coevolution and regulation lock cyclic plant defence peptides to their targets. *New Phytol.* 210, 717–730. doi:10.1111/nph.13789.

Gran, L. (1973). On the Effect of a Polypeptide Isolated from "Kalata-Kalata" (Oldenlandia affinis DC) on the Oestrogen Dominated Uterus. *Acta Pharmacol. Toxicol. (Copenh).* 33, 400–408. doi:10.1111/j.1600-0773.1973.tb01541.x.

Grewal, R. N., Aribi, H. El, Harrison, A. G., Siu, K. W. M., and Hopkinson, A. C. (2004). Fragmentation of Protonated Tripeptides: The Proline Effect Revisited. *J. Phys. Chem. B* 108. doi:10.1021/jp031093k.

Grover, T., Mishra, R., Bushra, Gulati, P., and Mohanty, A. (2021). An insight into biological activities of native cyclotides for potential applications in agriculture and pharmaceutics. *Peptides* 135. doi:10.1016/j.peptides.2020.170430.

Heitz, A., Hernandez, J. F., Gagnon, J., Thai Trinh Hong, Châu Pham, T., Tuyet Mai Nguyen, et al. (2001). Solution structure of the squash trypsin inhibitor MCoTI-II. A new family for cyclic knottins. *Biochemistry* 40. doi:10.1021/bi0106639.

Hellinger, R., Koehbach, J., Soltis, D. E., Carpenter, E. J., Wong, G. K. S., and Gruber, C. W. (2015). Peptidomics of circular cysteine-rich plant peptides: Analysis of the diversity of cyclotides from viola tricolor by transcriptome and proteome mining. *J. Proteome Res.* 14, 4851–4862. doi:10.1021/acs.jproteome.5b00681.

Hemu, X., El Sahili, A., Hu, S., Zhang, X., Serra, A., Goh, B. C., et al. (2020). Turning an Asparaginyl Endopeptidase into a Peptide Ligase. *ACS Catal.* 10. doi:10.1021/acscatal.0c02078.

Hemu, X., Zhang, X., and Tam, J. P. (2019). Ligase-controlled cyclo-oligomerization of peptides. *Org. Lett.* 21. doi:10.1021/acs.orglett.9b00151.

Hernandez, J. F., Gagnon, J., Chiche, L., Nguyen, T. M., Andrieu, J. P., Heitz, A., et al. (2000). Squash trypsin inhibitors from Momordica cochinchinensis exhibit an atypical macrocyclic structure. *Biochemistry* 39. doi:10.1021/bi9929756.

Herrmann, A., Svangård, E., Claeson, P., Gullbo, J., Bohlin, L., and Göransson, U. (2006). Key role of glutamic acid for the cytotoxic activity of the cyclotide cycloviolacin O2. *Cell. Mol. Life Sci.* 63. doi:10.1007/s00018-005-5486-4.

Jennings, C., West, J., Waine, C., Craik, D., and Anderson, M. (2001). Biosynthesis and insecticidal properties of plant cyclotides: The cyclic knotted proteins from Oldenlandia affinis. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10614–10619. doi:10.1073/pnas.191366898.

Kalmankar, N. V., Venkatesan, R., Balaram, P., and Sowdhamini, R. (2020). Transcriptomic profiling of the medicinal plant Clitoria ternatea: identification of potential genes in cyclotide biosynthesis. *Sci. Rep.* 10. doi:10.1038/s41598-020-69452-7.

Mulvenna, J. P., Sando, L., and Craik, D. J. (2005). Processing of a 22 kDa precursor protein to produce the circular protein tricyclon A. *Structure* 13. doi:10.1016/j.str.2005.02.013.

Mulvenna, J. P., Wang, C., and Craik, D. J. (2006). CyBase: a database of cyclic protein sequence and structure. *Nucleic Acids Res.* 34. doi:10.1093/nar/gkj005.

Nguyen, G. K. T., Kam, A., Loo, S., Jansson, A. E., Pan, L. X., and Tam, J. P. (2015). Butelase 1: A Versatile Ligase for Peptide and Protein Macrocyclization. *J. Am. Chem. Soc.* 137, 15398–15401. doi:10.1021/jacs.5b11014.

Nguyen, G. K. T., Wang, S., Qiu, Y., Hemu, X., Lian, Y., and Tam, J. P. (2014). Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nat. Chem. Biol.* 10, 732–738. doi:10.1038/nchembio.1586.

Nguyen, G. K. T., Zhang, S., Nguyen, N. T. K., Nguyen, P. Q. T., Chiu, M. S., Hardjojo, A., et al. (2011). Discovery and characterization of novel cyclotides originated from chimeric precursors consisting of albumin-1 chain a and cyclotide domains in the fabaceae family. *J. Biol. Chem.* 286, 24275–24287. doi:10.1074/jbc.M111.229922.

Nguyen, K. N. T., Nguyen, G. K. T., Nguyen, P. Q. T., Ang, K. H., Dedon, P. C., and Tam, J. P. (2016). Immunostimulating and Gram-negative-specific antibacterial cyclotides from the butterfly pea (Clitoria ternatea). *FEBS J.* 283, 2067–2090. doi:10.1111/febs.13720.

Oguis, G. K., Gilding, E. K., Huang, Y. H., Poth, A. G., Jackson, M. A., and Craik, D. J.

(2020). Insecticidal diversity of butterfly pea (Clitoria ternatea) accessions. *Ind. Crops Prod.* 147. doi:10.1016/j.indcrop.2020.112214.

Oguis, G. K., Gilding, E. K., Jackson, M. A., and Craik, D. J. (2019). Butterfly pea (Clitoria ternatea), a cyclotide-bearing plant with applications in agriculture and medicine. *Front. Plant Sci.* 10. doi:10.3389/fpls.2019.00645.

Parsley, N. C., Williams, O. L., and Hicks, L. M. (2020). Exploring the Diversity of Cysteine-Rich Natural Product Peptides via MS/MS Fingerprint Ions. *J. Am. Soc. Mass Spectrom.* 31. doi:10.1021/jasms.0c00078.

Poth, A. G., Colgrave, M. L., Lyons, R. E., Dalya, N. L., and Craik, D. J. (2011a). Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10127–10132. doi:10.1073/pnas.1103660108.

Poth, A. G., Colgrave, M. L., Philip, R., Kerenga, B., Daly, N. L., Anderson, M. A., et al. (2011b). Discovery of cyclotides in the Fabaceae plant family provides new insights into the cyclization, evolution, and distribution of circular proteins. *ACS Chem. Biol.* 6, 345–355. doi:10.1021/cb100388j.

Poth, A. G., Mylne, J. S., Grassl, J., Lyons, R. E., Millar, A. H., Colgrave, M. L., et al. (2012). Cyclotides associate with leaf vasculature and are the products of a novel precursor in Petunia (Solanaceae). *J. Biol. Chem.* 287, 27033–27046. doi:10.1074/jbc.M112.370841.

Rosengren, K. J., Daly, N. L., Plan, M. R., Waine, C., and Craik, D. J. (2003). Twists, knots, and rings in proteins: Structural definition of the cyclotide framework. *J. Biol. Chem.* 278, 8606–8616. doi:10.1074/jbc.M211147200.

Serra, A., Hemu, X., Nguyen, G. K. T., Nguyen, N. T. K., Sze, S. K., and Tam, J. P. (2016). A high-throughput peptidomic strategy to decipher the molecular diversity of cyclic cysteine-rich peptides. *Sci. Rep.* 6. doi:10.1038/srep23005.

Tam, J. P., Lu, Y. A., Yang, J. L., and Chiu, K. W. (1999). An unusual structural motif of antimicrobial peptides containing end-to-end macrocycle and cystine-knot disulfides. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8913–8918. doi:10.1073/pnas.96.16.8913.

Vaisar, T., and Urban, J. (1996). Probing the proline effect in CID of protonated peptides [1]. *J. Mass Spectrom.* 31. doi:10.1002/(SICI)1096-9888(199610)31:10<1185::AID-JMS396>3.0.CO;2-Q.

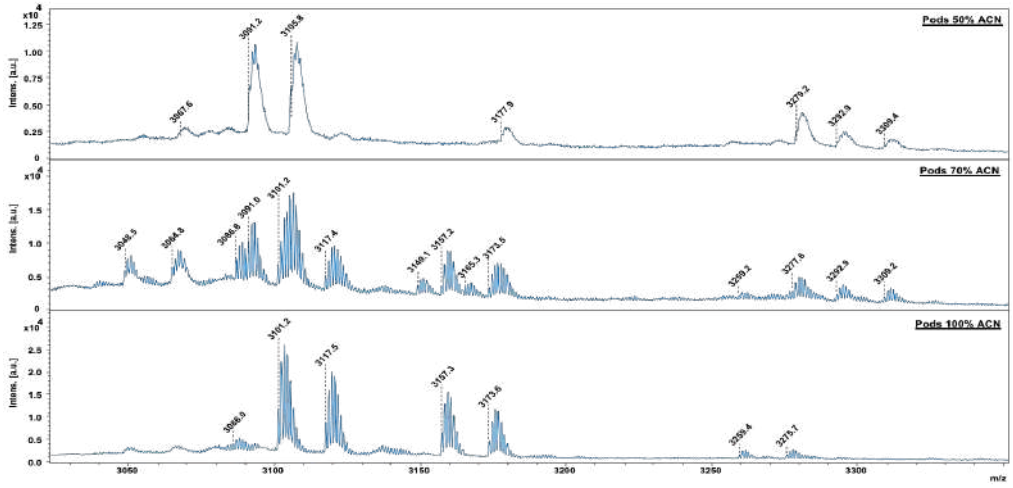
Wang, C. K. L., Colgrave, M. L., Gustafson, K. R., Ireland, D. C., Goransson, U., and Craik, D. J. (2008a). Anti-HIV cyclotides from the Chinese medicinal herb Viola yedoensis. *J. Nat. Prod.* 71. doi:10.1021/np070393g.

Wang, C. K. L., Kaas, Q., Chiche, L., and Craik, D. J. (2008b). CyBase: A database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Res.* 36. doi:10.1093/nar/gkm953.

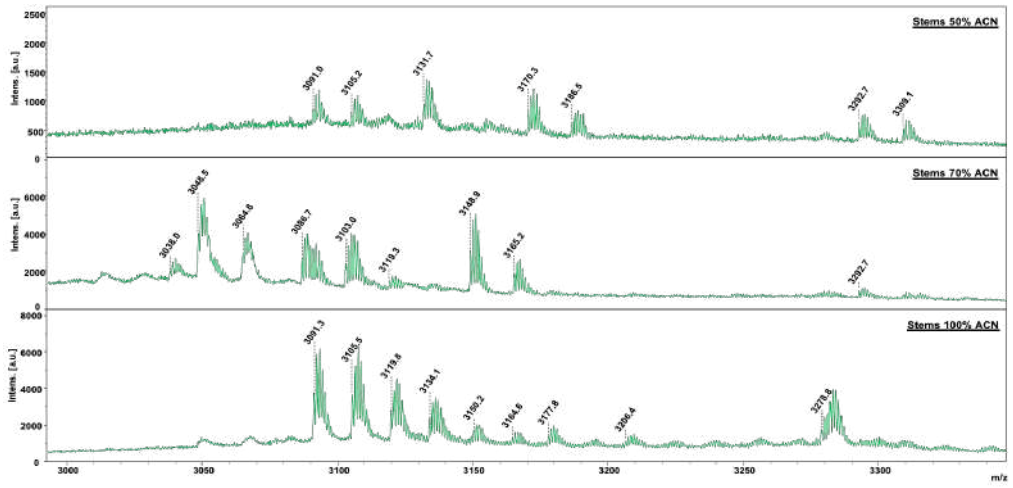
3.9 SUPPLEMENTARY INFORMATION OF CHAPTER 3

Figure S3.1: MALDI-TOF mass spectra of 50%, 70% and 100% eluates of the extracts of *C. ternatea* (A) stem, (B) leaf, (C) pod, (D) flower and (E) root tissues. Region containing cyclotide-like masses (2700-4000 Da) is shown.

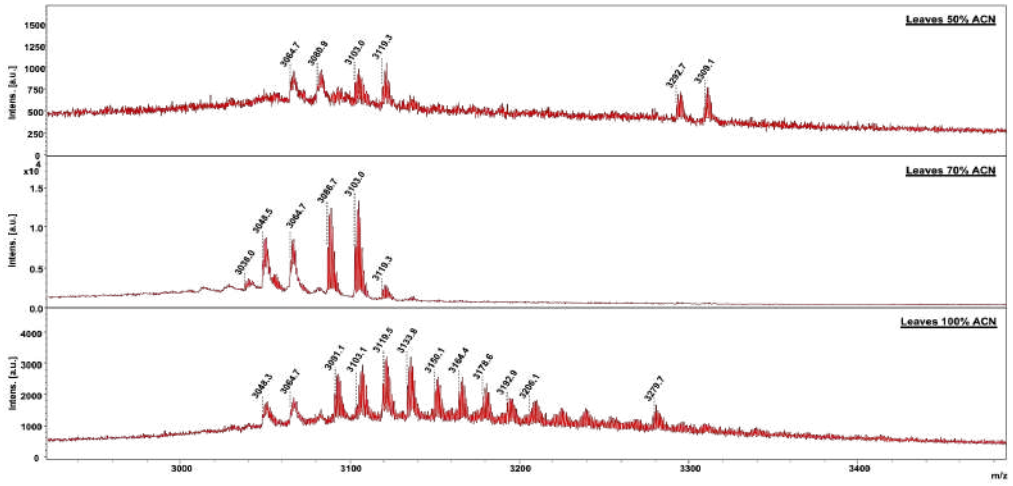
A



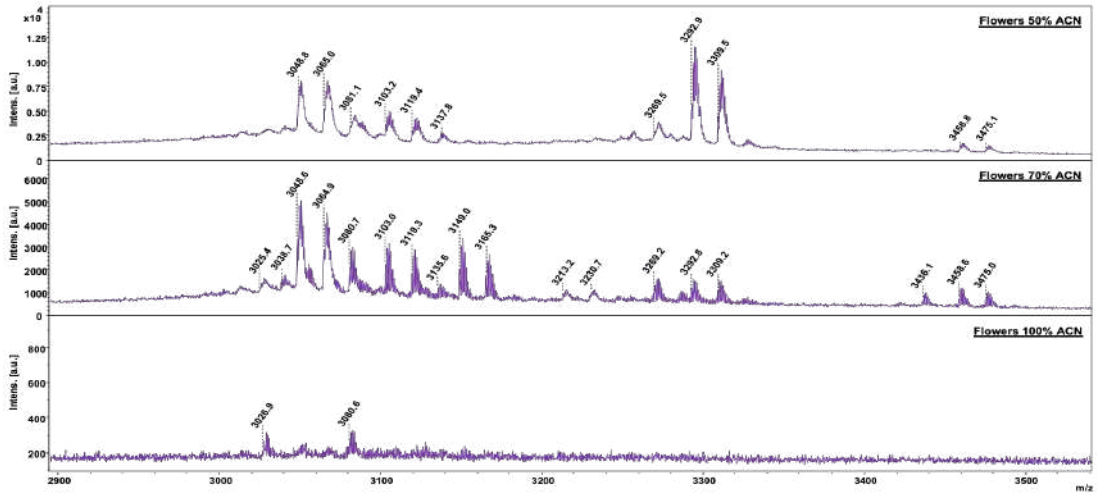
B



C



D



E

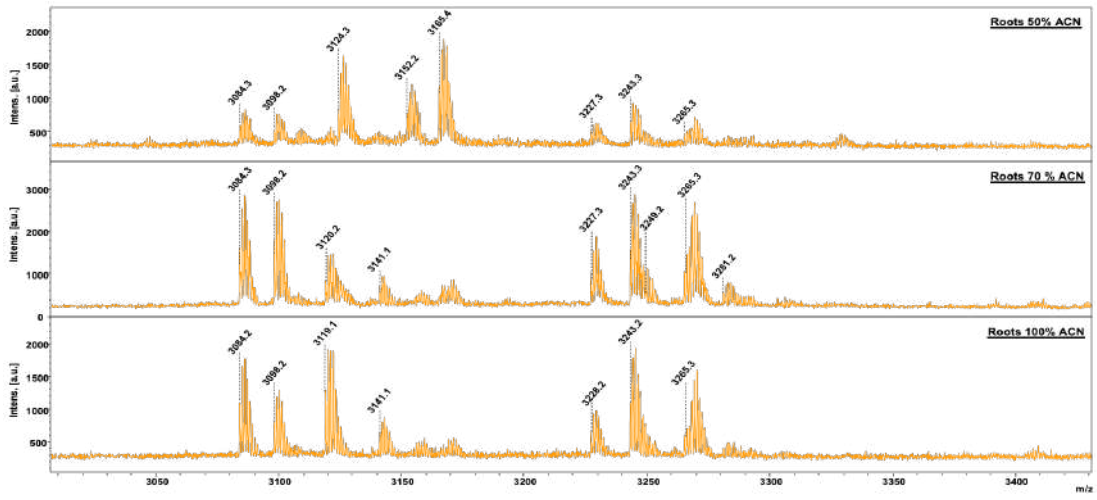


Table S3.1: List of putative cyclotides detected in Indian variety of *Clitoria ternatea*. The native cyclotides gained a mass of 348.18 Da upon reduction and alkylation, and additional 18 Da on linearization with endoproteinase Glu-C.

POD

Peak Number	Expt. Mass of Intact Peptides (Monoiso.) [Da]	Evidence from ESI-MS					
		Multiple Charged States (Monoiso.) (Da) [Expt.]				Red + Alk Monoiso. Mass (Da) [Calc.]	Endo-GluC Monoiso. Mass (Da) [Calc.]
		RT (Intact)	$[M+2H]^{2+}$	$[M+3H]^{3+}$	$[M+4H]^{4+}$		
1	3089.35	27.01		1030.78	773.34	3437.53	3455.53
2	3259.32	27.18		1087.44	815.83	3607.50	3625.50
3	3144.39	27.55		1049.13	787.10	3492.57	3510.57
	3089.34			1030.78		3437.52	3455.52
4	3259.32	27.85		1087.44	815.83	3607.50	3625.50
5	3053.34	28.86	1527.66	1018.78	764.33	3401.52	3419.52
	3107.22		1554.61	1036.74	777.81	3455.40	3473.40
6	3083.37	29.48	1542.69	1028.79	771.85	3431.55	3449.55
	3267.51			1090.17	817.88	3615.69	3633.69

7	3168.45	30.41		1057.15	793.11	3516.63	3534.63
	3083.40		1542.70	1028.80		3431.58	3449.58
8	3097.41	30.87	1549.69	1033.47	775.35	3445.59	3463.59
9	3097.41	31.20		1033.47		3445.59	3463.59
10	3091.23	31.83	1546.61	1031.41		3439.41	3457.41
	3080.91			1027.79		3429.09	3447.09
11	3089.22	32.10	1545.60	1030.74	773.81	3437.40	3455.40
12	3071.34	33.05	1536.66	1024.78	768.84	3419.52	3437.52
13	3074.13	33.23	1537.16	1025.71		3422.31	3440.31
14	3226.47	33.35	1614.23	1076.49	807.62	3574.65	3592.65
15	3126.39	33.63	1564.19	1043.13	782.60	3474.57	3492.57
16	3054.30	33.81	1528.15	1019.10		3402.48	3420.48
	3126.39		1564.18	1043.13		3474.57	3492.57
17	3127.35	34.39	1564.68	1043.46	782.84	3475.53	3493.53
18	3127.35	34.79	1564.68	1043.46		3475.53	3493.53
19	3109.35	34.96	1555.67	1037.45		3457.53	3475.53
20	3109.35	35.45	1556.67	1037.45		3457.53	3475.53
	3235.26		1618.63	1079.42		3583.44	3601.44
21	3073.23	36.36	1537.60	1025.41		3421.41	3439.41
	3118.32		1560.16	1040.44		3466.50	3484.50
22	3074.19	37.34	1538.10	1025.73		3422.37	3440.37
	3057.22		1529.61	1020.07		3405.40	3423.40
23	3250.46	38.82	1626.23	1084.49	813.62	3598.64	3616.64
24	3119.31	39.39	1560.65	1040.77		3467.49	3485.49

STEM

Peak Number	Expt. Mass of Intact Peptides (Monoiso.) [Da]	Evidence from Orbitrap Fusion Mass Spectrometer					
		Multiple Charged States (Monoiso.) [Da]				Red + Alk Monoiso. Mass [Da]	Endo-GluC Monoiso. Mass [Da]
		RT (Intact)	$[M+2H]^{2+}$	$[M+3H]^{3+}$	$[M+4H]^{4+}$		
1	3123.45	18.78	1562.72	1042.15	781.86	3471.63	3489.63
	3151.44		1576.70	1051.47	788.86	3499.62	3517.62
2	3162.45	19.20	1582.23	1055.15	791.62	3510.63	3528.63
3	3152.40	19.50		1051.80	789.10	3500.58	3518.58
	3247.35		1624.67	1083.45	812.84	3595.53	3613.53
4	3124.44	19.75	1563.21	1042.48	782.11	3472.62	3490.62
	3181.44		1591.72	1061.48	796.37	3529.62	3547.62
5	3181.48	20.20		1061.48	796.37	3529.66	3547.66
	3162.48			1055.15	791.62	3510.66	3528.66
	3248.36			1083.78	813.09	3596.54	3614.54
6	3151.44	20.70		1051.48	788.86	3499.62	3517.62
7	3122.44	21.11		1041.81	781.61	3470.62	3488.62
8	3151.44	21.46		1051.48	788.86	3499.62	3517.62
	3278.31			1093.77		3626.49	3644.49

9	3122.44	21.90		1041.81	781.61	3470.62	3488.62
	3301.40			1101.46	826.35	3649.58	3667.58
10	3139.32	22.40	1570.66	1047.45	785.84	3487.50	3505.50
11	3076.17	23.00	1539.08	1026.39		3424.35	3442.35
12	3140.34	24.25		1047.78	786.08	3488.52	3506.52
	3197.34			1066.78	800.34	3545.52	3563.52
13	3259.32	24.61		1087.44		3607.50	3625.50
14	3260.31	25.87	1631.15	1087.77		3608.49	3626.49
15	3097.41	27.59		1033.47		3445.59	3463.59
16	3230.20	28.20	1616.10	1077.73		3577.38	3595.38
	3097.41			1033.47		3445.59	3463.59
17	3083.40	28.65	1542.69	1028.80		3431.58	3449.58
18	3267.51	28.96	1634.75	1090.17		3615.69	3633.69
	3083.40		1542.69	1028.80		3431.58	3449.58
19	3021.24	29.19	1511.62	1008.08		3369.42	3387.42
	3267.51			1090.17		3615.69	3633.69
20	3097.41	29.90		1033.47		3445.59	3463.59
	3021.24		1511.62	1008.08		3369.42	3387.42
21	3229.20	30.90	1615.60	1077.40		3577.38	3595.38
22	3098.40	32.00	1550.20	1033.80		3446.58	3464.58
23	3271.44	33.00		1091.48		3619.62	3637.62
24	3077.22	33.44	1539.60	1026.74		3425.40	3443.40
25	3243.51	34.41	1622.75	1082.17		3590.69	3608.69
26	3227.46	35.94		1076.82		3575.64	3593.64
27	3074.19	37.18	1538.10	1025.73		3422.37	3440.37

LEAF

Peak Number	**Expt. Mass of Intact Peptides (Monoiso.) [Da]**	**Evidence from Orbitrap Fusion Mass Spectrometer**					
		Multiple Charged States (Monoiso.) [Da]				**Red + Alk Monoiso. Mass [Da]**	**Endo-GluC Monoiso. Mass [Da]**
		RT (Intact)	**$[M+2H]^{2+}$**	**$[M+3H]^{3+}$**	**$[M+4H]^{4+}$**		
1	3181.48	19.92		1061.49	796.37	3529.66	3547.66
	3271.48			1091.50	818.87	3619.66	3637.66
2	3275.31	21.44		1092.77		3623.49	3641.49
	3291.30			1098.10		3639.48	3657.48
3	3139.36	22.40		1047.45	785.84	3487.54	3505.54
	3263.31			1088.77	816.83	3611.49	3629.49
	3195.44			1066.15	799.86	3543.62	3561.62
4	3079.17	22.75		1026.39		3427.35	3445.35
5	3197.36	22.40		1066.78	800.34	3545.54	3563.54
	3287.40			1096.79	822.85	3635.58	3653.58
6	3020.19	24.72	1511.08	1007.73		3368.37	3386.37
	3259.32			1087.44		3607.50	3625.50
7	3259.32	25.35	1630.66	1087.44		3607.50	3625.50
8	3317.34	26.37		1106.78		3665.52	3683.52
9	3260.31	26.90	1631.15	1087.77		3608.49	3626.49
10	3083.40	27.85	1542.69	1028.80		3431.58	3449.58

	3267.51		1634.75	1090.17		3615.69	3633.69
11	3097.41	28.18	1549.70	1033.47		3445.59	3463.59
	3168.45		1585.22	1057.15		3516.63	3534.63
12	3083.40	28.51	1542.69	1028.80		3431.58	3449.58
13	3141.39	28.81	1571.69	1048.13		3489.57	3507.57
	3231.42		1616.71	1078.14		3579.60	3597.60
	3201.42		1601.70	1068.14		3549.60	3567.60
	3083.36		1542.68	1028.79		3431.54	3449.54
	3185.43		1593.70	1062.81		3533.61	3551.61
14	3089.22	29.55	1545.60	1030.74		3437.40	3455.40
15	3155.40	29.74		1052.80		3503.58	3521.58
	3245.43		1623.71	1082.81		3593.61	3611.61
	3111.39		1556.69	1038.13		3459.57	3477.57
16	3155.40	30.31	1578.70	1052.80		3503.58	3521.58
17	3097.41	30.80	1549.69	1033.47		3445.59	3463.59
18	3125.38	31.82	1563.69	1042.79		3473.56	3491.56
	3155.40		1578.70	1052.80		3503.58	3521.58
	3245.43		1623.71	1082.81		3593.61	3611.61
	3077.22		1539.61	1026.74		3425.40	3443.40
19	3155.40	32.80	1578.70	1052.80		3503.58	3521.58
	3245.43		1623.71	1082.81		3593.61	3611.61
20	3125.43	33.50	1563.72	1042.81		3473.61	3491.61
	3226.50			1076.50		3574.68	3592.68
21	3073.22	35.85	1537.61	1025.41		3421.40	3439.40
22	3057.22	36.70	1529.61	1020.07		3405.40	3423.40
	3118.32		1560.16	1040.44		3466.50	3484.50
	3071.19		1536.60	1024.73		3419.37	3437.37
23	3112.23	37.36	1557.11	1038.41		3460.41	3478.41
	3241.41		1621.70	1081.47		3589.59	3607.59
	3227.38		1614.69	1076.80		3575.56	3593.56

FLOWER

Peak Number	Expt. Mass of Intact Peptides (Monoiso.) [Da]	Evidence from Orbitrap Fusion Mass Spectrometer					
		Multiple Charged States (Monoiso.) [Da]				Red + Alk Monoiso. Mass [Da]	Endo-GluC Monoiso. Mass [Da]
		RT (Intact)	$[M+2H]^{2+}$	$[M+3H]^{3+}$	$[M+4H]^{4+}$		
1	3275.31	21.34	1638.66	1092.77	819.83	3623.49	3641.49
2	3291.33	21.78	1646.65	1098.11	823.83	3639.51	3657.51
3	3275.31	22.10	1638.66	1092.77	819.83	3623.49	3641.49
4	3202.32	22.47	1602.15	1068.44		3550.50	3568.50
	3291.30			1098.10	823.84	3639.48	3657.48
5	3263.31	22.89	1632.66	1088.77	816.83	3611.49	3629.49
	3413.37		1707.68	1138.79		3761.55	3779.55
	3139.36			1047.45		3487.54	3505.54
6	3264.30	23.62		1089.10	817.08	3612.48	3630.48
7	3421.38	24.35		1141.46		3769.56	3787.56
	3287.31			1096.77		3635.49	3653.49
8	3421.38	24.80	1711.68	1141.46		3769.56	3787.56
9	3259.32	25.62	1630.66	1087.44		3607.50	3625.50

	3421.38		1711.68	1141.46		3769.56	3787.56
10	3259.32	25.85	1630.66	1087.44		3607.50	3625.50
11	3260.31	27.50	1631.15	1087.77		3608.49	3626.49
12	3107.22	28.04	1554.61	1036.74		3455.40	3473.40
13	3095.22	28.71	1548.61	1032.74		3443.40	3461.40
14	3021.24	29.25	1511.62	1008.09		3369.42	3387.42
	3034.20		1518.10	1012.40		3382.38	3400.38
	3091.28		1546.61	1031.41		3439.46	3457.46
15	3091.28	30.87	1546.61	1031.41		3439.46	3457.46
16	3105.20	31.40	1553.60	1036.07		3453.38	3471.38
	3089.20		1545.60	1030.74		3437.38	3455.38
	3122.22		1562.11	1041.74		3470.40	3488.40
17	3121.20	32.00	1561.60	1041.40		3469.38	3487.38
	3138.24		1570.11	1047.08		3486.42	3504.42
	3086.22		1544.10	1029.74		3434.40	3452.40
	3242.46			1081.82		3590.64	3608.64
18	3077.22	33.13	1539.60	1026.74		3425.40	3443.40
	3226.50			1076.50		3574.68	3592.68
19	3209.43	34.09		1070.81		3557.61	3575.61
	3227.46			1076.82		3575.64	3593.64
	3235.26		1618.63	1079.42		3583.44	3601.44
20	3227.46	34.30	1614.73	1076.82		3575.64	3593.64
21	3073.22	36.00	1537.61	1025.40		3421.40	3439.40
22	3118.32	36.65	1560.16	1040.44		3466.50	3484.50
	3135.34		1568.67	1046.11		3483.52	3501.52
23	3074.19	37.10	1538.10	1025.73		3422.37	3440.37
	3057.21		1529.61	1020.07		3405.39	3423.39
	3091.23		1546.61	1031.41		3439.41	3457.41
24	3073.22	38.00	1537.60	1025.41		3421.40	3439.40
	3034.17		1518.08	1012.39		3382.35	3400.35

ROOT

Peak Number	Expt. Mass of Intact Peptides (Monoiso.) [Da]	Evidence from Orbitrap Fusion Mass Spectrometer					
		Multiple Charged States (Monoiso.) [Da]				Red + Alk Monoiso. Mass [Da]	Endo-GluC Monoiso. Mass [Da]
		RT (Intact)	$[M+2H]^{2+}$	$[M+3H]^{3+}$	$[M+4H]^{4+}$		
1	3114.39	26.68		1039.13		3462.57	3480.57
2	3268.50	29.20	1635.24	1090.50		3616.68	3634.68
3	3097.41	29.60	1549.70	1033.47		3445.59	3463.59
4	3083.36	30.41		1028.79		3431.54	3449.54
5	3268.50	30.77	1635.24	1090.50	818.12	3616.68	3634.68
6	3169.41	31.03	1585.71	1057.47		3517.59	3535.59
7	3098.40	31.71	1550.19	1033.80		3446.58	3464.58
8	3242.46	31.94	1622.23	1081.82		3590.64	3608.64
9	3255.45	32.24	1628.72	1086.15		3603.63	3621.63
10	3243.45	32.65	1622.72	1082.15		3591.63	3609.63
	3154.41		1578.20	1052.47		3502.59	3520.59
11	3550.59	33.12	1776.28	1184.53		3898.77	3916.77
	3388.53		1695.26	1130.51		3736.71	3754.71
12	3226.47	33.48	1614.23	1076.49		3574.65	3592.65

13	3243.45	34.10	1622.72	1082.15		3591.63	3609.63
14	3243.45	34.70		1082.15		3591.63	3609.63
	3226.47			1076.49		3574.65	3592.65
	3156.39		1579.19	1053.13		3504.57	3522.57
15	3112.35	34.86	1557.18	1038.45		3460.53	3478.53
	3126.39		1564.18	1043.13		3474.57	3492.57
16	3083.34	35.07	1542.67	1028.78		3431.52	3449.52
	3243.45			1082.15		3591.63	3609.63
	3296.49		1649.24	1099.83		3644.67	3662.67
17	3227.46	35.63	1614.72	1076.82		3575.64	3593.64
	3073.22		1537.60	1025.41		3421.40	3437.40
18	3301.47	36.92	1651.73	1101.49		3649.65	3667.65
19	3301.47	37.40	1651.72	1101.49		3649.65	3667.65
	3285.45		1643.73	1096.15		3633.63	3651.63
20	3285.45	39.46	1643.73	1096.15		3633.63	3651.63
21	3119.31	39.83	1560.65	1040.77		3467.49	3485.49
	3136.35		1569.17	1046.45		3484.53	3502.53
	3153.36			1052.12		3501.54	3519.54
	3141.30		1571.65	1048.10		3489.48	3507.48
22	3359.46	41.90		1120.82		3707.64	3725.64

Table S3.2: Validation of top 10 known cyclotides identified in this study (reduced, alkylated and endo Glu-C linearized samples) using SEQUEST tool of Proteome Discoverer v2.1 (Thermo Scientific)

Tissue	**Accession**	**Annotated Sequence**	**Expt. Intact Mass (Da) [Calc.]**	**Reduced, Alkylated, Endo Glu-C MH^+ (Da) [Expt.]**	**RT [Expt.]**	**Cover-age**	**# PSMs**	**# Unique Peptides**	**Score Sequest HT**
Pod	CBP00595 Cter M	[E].TcTLGTcYVPDcScSWPIcMKNGLPTcGE.[T]	3057.1775	3424.3575	35.74	87.931	19	4	88.36
	CBP00805 caripe 1	[E].ScVFIPcISTVIGcScKDKVcYRNGVIPcGE.[S]	3268.4622	3635.6422	32.42	88.710	11	3	52.30
	P86903 Cter R	[E].ScVFIPcTVTALLGcScKDKVcYKNGIPcGE.[S]	3226.4542	3593.6342	33.15	90.323	8	3	43.44
	CBP00975 cliotide T1	[E].ScVFIPcITGAIGcScKSKVcYRNGIPcGE.[S]	3083.3743	3450.5543	29.54	90.000	7	2	43.39
	P86848 Cter H	[E].ScVFIPcITTVVGcScKNKVcYNDGLPcGE.[S]	3127.3663	3494.5463	33.31	90.000	11	2	39.31
	P84641 circulin C	[E].ScVFIPcITSVAGcScKSKVcYRNGIPcGE.[S]	3099.3690	3466.5490	29.42	90.000	8	2	37.27
	P86841 Cter A	[E].ScVFIPcISTVIGcScKNKVcYRNGVIPcGE.[S]	3267.4937	3634.6737	31.76	88.710	7	1	32.87
	P86902 Cter P	[E].ScVFIPcITAAIGcScKSKVcYRNGIPcGE.[S]	3097.3747	3464.5547	30.52	90.000	5	2	32.22
	P86904 Cter Q	[E].ScVFIPcISTVIGcScKNKVcYRNGIPcGE.[S]	3168.4110	3535.5910	31.02	90.000	6	1	31.96
	G1CWH7 cliotide T9	[E].ScVFIPcITTVVGcScKNKVcYNNGLPcGE.[S]	3126.3874	3493.5674	32.69	90.000	9	2	31.85
Stem	CBP00973 cliotide T20	[E].ScLLGKcYTPGcTcDRPIcKKNGSAIRcGE.[S]	3151.3986	3518.5786	19.14	86.667	11	3	59.34

	A0A0S1RGA2 Cter 30	[E].TcFKTKcYTPGcScSYPVcKKNGFPIcGE.[T]	3139.3238	3506.5038	23.01	87.931	6	2	36.69
	CBP00972 cliotide T19	[E].ScLLGKcYTPGcTcSRPIcKKNGSVIKcGE.[S]	3123.4345	3490.6145	19.29	86.667	6	2	35.56
	CBP00975 cliotide T1	[E].ScVFIPcITGAIGcScKSKVcYRNGIPcGE.[S]	3083.3658	3450.5458	29.96	90.000	5	1	30.76
	CBP00971 cliotide T18	[E].TcFTGTcYTPGcTcSYPVcKKNGLPIcGE.[T]	3021.2318	3388.4118	27.87	87.931	6	2	30.18
	G1CWH8 cliotide T12	[E].ScVFIPcITGAIGcScKSKVcYRDGIPcGE.[S]	3084.3465	3451.5265	30.76	90.000	6	2	29.74
	CBP00969 cliotide T16	[E].TcLRGRcYTPGcTcDHGIcKKNGSVIGcGE.[T]	3106.3123	3473.4923	18.10	75.000	4	2	22.91
	P84641 circulin C	[E].ScVFIPcITSVAGcScKSKVcYRNGIPcGE.[S]	3099.3648	3466.5448	33.54	50.000	5	1	22.01
	CBP00805 caripe 1	[E].ScVFIPcISTVIGcScKDKVcYRNGVIPcGE.[S]	3268.4688	3635.6488	32.57	88.710	5	3	18.81
	CBP01082 cliotide T27	[E].ScVFIPcITGAIGcScKSKVcYRNGVIPcGE.[S]	3182.4167	3549.5967	21.62	88.710	5	1	17.97
Leaf	CBP00595 Cter M [M20 Ox]	[E].TcTLGTcYVPDcScSWPIcmKNGLPTcGE.[T]	3073.1888	3440.3688	33.61	87.931	18	5	85.29
	P86902 Cter P	[E].ScVFIPcITAAIGcScKSKVcYRNGIPcGE.[S]	3097.3734	3464.5534	30.69	90.000	5	2	30.43
	P86850 Cter J	[E].ScVFIPcITGIAGcScKNKVcYIDGTVPcGE.[S]	3155.3819	3522.5619	33.50	50.000	8	1	29.78
	CBP00972 cliotide T19	[E].ScLLGKcYTPGcTcSRPIcKKNGSVIKcGE.[S]	3123.4230	3490.6030	19.60	86.667	4	2	25.71
	CBP00503 Vpf-1	[E].ScVFIPcLTAAIGcScRSKVcYRNGIPcGE.[S]	3125.3690	3492.5490	33.82	90.000	5	2	23.92
	G1CWH8 cliotide T12	[E].ScVFIPcITGAIGcScKSKVcYRDGIPcGE.[S]	3084.3441	3451.5241	30.68	90.000	4	2	21.48
	P86841 Cter A	[E].ScVFIPcISTVIGcScKNKVcYRNGVIPcGE.[S]	3267.4788	3634.6588	31.95	88.710	4	1	21.23
	A0A0S1RGA2 Cter 30	[E].TcFKTKcYTPGcScSYPVcKKNGFPIcGE.[T]	3139.3180	3506.4980	23.09	87.931	4	2	20.78
	CBP00805 caripe 1	[E].ScVFIPcISTVIGcScKDKVcYRNGVIPcGE.[S]	3268.4642	3635.6442	33.70	79.032	5	2	20.66
	CBP01082 cliotide T27	[E].ScVFIPcITGAIGcScKSKVcYRNGVIPcGE.[S]	3182.4127	3549.5927	21.93	88.710	5	2	20.47
Flower	CBP00595 Cter M [M20 Ox]	[E].TcTLGTcYVPDcScSWPIcmKNGLPTcGE.[T]	3073.1924	3440.3724	35.93	87.931	15	5	65.83
	CBP00975 cliotide T1	[E].ScVFIPcITGAIGcScKSKVcYRNGIPcGE.[S]	3083.3651	3450.5451	29.81	90.000	3	2	17.93
	A0A0S1RQN3 Cter 19	[E].ScVYIPcLTTIVGcScKSNVcYSNSIPcGE.[S]	3118.3072	3485.4872	36.04	90.000	5	2	16.25
	CBP00976 cliotide T6	[E].ScVYIPcLTTIVGcScKNSVcYSNSIPcGE.[S]	3118.3072	3485.4872	36.04	90.000	5	2	16.17

	CBP00971 cliotide T18	[E].TcFTGTcYTPGcTcSYPVcKKNGLPIcGE.[T]	3021.2359	3388.4159	28.13	50.000	2	1	14.34
	A0A0S1RGA2 Cter 30	[E].TcFKTKcYTPGcScSYPVcKKNGFPIcGE.[T]	3139.3238	3506.5038	23.22	50.000	2	1	13.24
	P86903 Cter R	[E].ScVFIPcTVTALLGcScKDKVcYKNGIPcGE.[S]	3226.4434	3593.6234	33.32	90.323	2	2	9.51
	P86902 Cter P	[E].ScVFIPcITAAIGcScKSKVcYRNGIPcGE.[S]	3097.3819	3464.5619	30.81	50.000	2	1	9.43
	P86850 Cter J	[E].ScVFIPcITGIAGcScKNKVcYIDGTVPcGE.[S]	3155.3546	3522.5346	35.04	88.710	2	2	5.21
	CBP00977 cliotide T10	[E].ScVYIPcTVTALLGcScKDKVcYKNGIPcGE.[S]	3242.4456	3609.6256	32.37	50.000	1	1	4.8
Root	CBP00973 cliotide T20	[E].ScLLGKcYTPGcTcDRPIcKKNGSAIRcGE.[S]	3151.3995	3518.5795	19.51	86.667	5	3	32.53
	P86903 Cter R	[E].ScVFIPcTVTALLGcScKDKVcYKNGIPcGE.[S]	3226.4417	3593.6217	33.07	90.323	6	3	27.38
	CBP00977 cliotide T10	[E].ScVYIPcTVTALLGcScKDKVcYKNGIPcGE.[S]	3242.4415	3609.6215	32.22	90.323	6	3	25.42
	CBP00972 cliotide T19	[E].ScLLGKcYTPGcTcSRPIcKKNGSVIKcGE.[S]	3123.4293	3490.6093	19.69	86.667	5	2	22.43
	CBP00595 Cter M [M20 Ox]	[E].TcTLGTcYVPDcScSWPIcmKNGLPTcGE.[T]	3073.1978	3440.3778	33.55	68.966	4	2	17.21
	P86902 Cter P	[E].ScVFIPcITAAIGcScKSKVcYRNGIPcGE.[S]	3097.3814	3464.5614	30.56	50.000	3	1	16.54
	CBP00805 caripe 1	[E].ScVFIPcISTVIGcScKDKVcYRNGVIPcGE.[S]	3268.4574	3635.6374	32.71	50.000	3	1	11.29
	CBP00968 cliotide T15	[E].TcFKTKcYTKGcScSYPVcKRNGLPIcGE.[T]	3164.3808	3531.5608	20.66	50.000	2	1	11.07
	CBP00975 cliotide T1	[E].ScVFIPcITGAIGcScKSKVcYRNGIPcGE.[S]	3083.3626	3450.5426	29.80	50.000	2	1	9.15
	P86904 Cter Q	[E].ScVFIPcISTVIGcScKNKVcYRNGIPcGE.[S]	3168.4185	3535.5985	31.08	50.000	2	1	8.84

Table S3.3: Complete list of mass spectrometrically characterized *C. ternatea* cyclotides classified by proline content.

Name	**Alternative name (s)**	**Sequence**	**Length**	**Average Mass (Da)**	**Monoiso-topic Mass (Da)**	**Number of Basic residues (K, R, H)** [a]	**Type** [b]	**Identification Technique**
1 Proline								
Cter N		TCVLGTCYTPDCSCTALVCLKNGSAFCGE	29	2936.37	2934.18	K=1,R=0, H=0	B	MS
Cter 8		TCVLGTCYTPDCSCKAVVCIKNGSAFCGE	29	2955.46	2953.26	K=2,R=0, H=0	B	NGS, MS
Cter 18	cT43, ctr pep 41	TCLHTPCKASVCYCKNAVCYKNDLICSS	28	3044.56	3042.30	K=3,R=0, H=1	H	NGS, MS
ctr pep 18		SCVFIPCITGIAGCSCKNKVCYLNGISCGE	30	3059.61	3057.33	K=2,R=0, H=0	B	NGS, MS

cT16	ctr pep 27	TCLRGRCYTPGCTCDHGIC KKNGSVIGCGE	30	3108.57	3106.31	K=2,R=2, H=1	H	NGS, MS
ctr pep 22		GCLLGRCYRPGCTCVRRIC RRNGSITGCGG	30	3164.73	3162.46	K=0,R=6, H=0	H	NGS, MS
cT33	Cter 34, ctr pep 16	ACVYLPCFSKGCSCFKRQC YKNGFNSCSE	29	3249.73	3247.33	K=3,R=1, H=0	H	MS, NGS
ctr pep 38		SCIYLPCISGVFEGCSCQNK ACYKNGGTACGE	32	3282.72	3280.32	K=2,R=0, H=0	B	NGS, MS
2 Prolines								
cT8	ctr pep 45	SCVFIPCISSVVGCSCKSKV CYNNGIPCGE	30	3073.60	3071.31	K=2,R=0, H=0	B	MS, PcR, NGS
Cter F		SCVFIPCISSVVGCSCKSKV CYLDGIPCGE	30	3073.64	3071.34	K=2,R=0, H=0	B	MS, Amino acid analysis
cT1		SCVFIPCITGAIGCSCKSKV CYRNGIPCGE	30	3085.65	3083.36	K=2,R=1, H=0	B	NGS, PcR, MS, Amino acid analysis
Cter O		SCVFIPCITGIAGCSCKSKV CYRNGIPCGE	30	3085.65	3083.36	K=2,R=1, H=0	B	MS
cT12		SCVFIPCITGAIGCSCKSKV CYRDGIPCGE	30	3086.64	3084.35	K=2,R=1, H=0	B	NGS, MS, PcR
Cter P	cT4	SCVFIPCITAAIGCSCKSKV CYRNGIPCGE	30	3099.68	3097.38	K=2,R=1, H=0	B	NGS, MS, PcR
cT21	Cter 17	TCVHSPCIGPCYCKHGLIC YRNDLQCAE	28	3103.59	3101.29	K=1,R=1, H=2	H	NGS, MS
Cter K		SCVFIPCITTVVGCSCKNK VCYNHEPCGE	29	3110.62	3108.31	K=2,R=0, H=1	B	MS
ctr pep 4		TCFGGTCYTPGCSCDYPIC KKNGDALKCGE	30	3111.52	3109.22	K=3,R=0, H=0	M	NGS, MS
Cter L		SCVFIPCITTVVGCSCKNK VCYDHEPCGE	29	3111.60	3109.29	K=2,R=0, H=1	B	MS
Cter 19	cT6, ctr pep 49	SCVYIPCLTTIVGCSCKSNV CYSNSIPCGE	30	3120.61	3118.30	K=1,R=0, H=0	B	MS, NGS
cT6		SCVYIPCLTTIVGCSCKNSV CYSNSIPCGE	30	3120.61	3118.30	K=1,R=0, H=0	B	MS
cT19	Cter 26	SCLLGKCYTPGCTCSRPIC KKNGSVIKCGE	30	3125.72	3123.43	K=4,R=1, H=0	M	MS, NGS
cT9	ctr pep 2	SCVFIPCLTTVVGCSCKNK VCYNNGIPCGE	30	3128.68	3126.36	K=2,R=0, H=0	B	MS, PcR, NGS
Cter G		SCVFIPCITTVVGCSCKNK VCYNNGLPCGE	30	3128.68	3126.36	K=2,R=0, H=0	B	MS, Amino acid analysis
Cter H		SCVFIPCITTVVGCSCKNK VCYNDGLPCGE	30	3129.66	3127.34	K=2,R=0, H=0	B	MS, Amino acid analysis
ctr pep 30		SCVFIPCISTVIGCSCKNKV CYRNGSPCGE	30	3144.68	3142.36	K=2,R=1, H=0	B	NGS, MS
cT20		SCLLGKCYTPGCTCDRPIC KKNGSAIRCGE	30	3153.69	3151.40	K=3,R=2, H=0	M	NGS, MS
cT17		SCVFIPCITGIAGCSCKNKV CYLNGTVPCGE	31	3156.73	3154.39	K=2,R=0, H=0	B	NGS, MS
Cter I		SCVFIPCITGIAGCSCKNKV CYINGTVPCGE	31	3156.73	3154.39	K=2,R=0, H=0	B	MS
Cter J		SCVFIPCITGIAGCSCKNKV CYIDGTVPCGE	31	3157.72	3155.37	K=2,R=0, H=0	B	MS
Cter 25	cT41, ctr pep 9	TCFFQKCYTPGCSCDAVIC TNNGNPIVCGE	30	3164.58	3162.25	K=1,R=0, H=0	H	MS, NGS
cT15	Cter 24, ctr pep 7	TCFKTKCYTKGCSCSYPVC KRNGLPICGE	29	3166.73	3164.38	K=4,R=1, H=0	M	MS, NGS
cT32	ctr pep 25	TCFGGTCYTPGCSCDYPIC KNNGDLFKCGE	30	3173.54	3171.20	K=2,R=0, H=0	M	MS, NGS
cT27		SCVFIPCITGAIGCSCKSKV CYRNGVIPCGE	31	3184.79	3182.43	K=2,R=1, H=0	B	NGS, MS

ctr pep 8		TCFKTKCYTKGCSCSYPVC KRNGRPTCGE	29	3197.70	3195.36	K=4,R=2, H=0	M	NGS, MS
cT40	Cter 14, ctr pep 1	SCVFIPCTITALLGCSCKSK VCYKNGIPCGE	31	3214.85	3212.47	K=3,R=0, H=0	B	MS, NGS
Cter R	cT7, ctr pep 34	SCVFIPCTVTALLGCSCKD KVCYKNGIPCGE	31	3228.84	3226.45	K=3,R=0, H=0	B	MS, PcR, NGS
cT11	Cter 21, ctr pep 39	SCVFIPCTITALLGCSCKDK VCYKNGIPCGE	31	3242.86	3240.46	K=3,R=0, H=0	B	MS, Amino acid analysis, NGS
Cter B	ctr pep 51	SCVWIPCTVTALLGCSCKD KVCYLNGVPCAE	31	3252.86	3250.45	K=2,R=0, H=0	B	MS, PcR, Amino acid analysis, NGS
Cter C		SCVWIPCTVTALLGCSCKD KVCYLDGVPCAE	31	3253.84	3251.43	K=2,R=0, H=0	B	MS, Amino acid analysis
cT2		SCVQGECYTPGCSCDWPIC KKNGEFLKCGE	30	3261.70	3259.30	K=3,R=0, H=0	M	PcR, NGS, MS
cT23		SCVFIPCTVTALLGCSCKD KVCYKNGFPCGE	31	3262.85	3260.43	K=3,R=0, H=0	B	NGS, MS
Cter 32		SCVWIPCISSILGCSCKDKV CYHNKIPCGE	30	3262.86	3260.44	K=3,R=0, H=1	B	MS, NGS
Cter D		SCVWIPCTVTALLGCSCKD KVCYLNGIPCAE	31	3266.89	3264.46	K=2,R=0, H=0	B	MS, Amino acid analysis
Cter E		SCVWIPCTVTALLGCSCKD KVCYLDGIPCAE	31	3267.87	3265.44	K=2,R=0, H=0	B	MS, Amino acid analysis
Cter A	ctr pep 17	SCVFIPCISTVIGCSCKNKV CYRNGVIPCGE	31	3269.89	3267.48	K=2,R=1, H=0	B	MS, Amino acid analysis, NGS
cT54	ctr pep 46	SCVYIPCTVTALLGCSCKN KVCYRNGIPCGE	31	3271.87	3269.46	K=2,R=1, H=0	B	MS, NGS
cT55		SCVFIPCISTLIGCSCKNKV CYRNGVIPCGE	31	3283.92	3281.50	K=2,R=1, H=0	B	NGS, MS
ctr pep 20		SRVFIPCITGAIGCSCKSKV CYRNCVIPCGE	31	3283.92	3281.51	K=2,R=2, H=0	B	NGS, MS
cT13	Cter23, ctr pep 15	SCVWIPCVSSIVGCSCQNK VCYQNDTTPCGE	31	3300.73	3298.33	K=1,R=0, H=0	B	MS, NGS
cT38	ctr pep 13	SCVWIPCFTSAFGCYCQSK VCYHSKIPCGE	30	3347.92	3345.43	K=2,R=0, H=1	B	MS, NGS
ctr pep 35		SCVYIPCITSIVGCSCQNKV CYNNNIFPCGE	31	3348.92	3346.45	K=1,R=0, H=0	B	NGS, MS
cT14	ctr pep 32	SCVWIPCISSILGCSCKDKV CYHNDTIPCGE	31	3350.88	3348.42	K=2,R=0, H=1	B	NGS, MS
Cter 10	cT34, ctr pep 40	SCVYIPCTVTALLGCSCSNK VCYKNSYIPCGE	32	3395.96	3393.47	K=2,R=0, H=0	B	MS, NGS
3 or more Prolines								
ctr pep 12		TCTLGTCYVPDCSCSWPIC MKNLPTCGE	28	3002.56	3000.22	K=1,R=0, H=0	M	NGS, MS
cT18	Cter 6, ctr pep 28	TCFTGTCYTPGCTCSYPVC KKNGLPICGE	29	3023.49	3021.23	K=2,R=0, H=0	M	MS, NGS
cT28	Cter 16	SCVFLPCFLPGCSCKSSVCY LNGGSIPCGE	30	3047.56	3045.27	K=1,R=0, H=0	H	NGS, MS
Cter M	cT3	TCTLGTCYVPDCSCSWPIC MKNGLPTCGE	29	3059.54	3057.20	K=1,R=0, H=0	M	NGS, MS, PcR
ctr pep 43		SCVFLPCFIIPGCSCKDKVC YLNGIPCGE	29	3085.69	3083.35	K=2,R=0, H=0	B	NGS, MS
ctr pep 3		TCFKTKCYTPGCSCSYPIC KKNGRPTCGE	29	3152.66	3150.33	K=4,R=1, H=0	M	NGS, MS
Cter 2	cT29	SCFAGKCYTPGCTCEYPIC MNNGDPLKCGE	30	3181.63	3179.21	K=2,R=0, H=0	M	NGS, MS

[a] K: Lysine; R: Arginine; H: Histidine. [b] B: Bracelet; M: Möbius; H: Hybrid.

Figure S3.2: Peptide sequencing of 5+ charge state precursor ion of Cliotide T1 (m/z= 690.92) from reduced, alkylated and endoproteinase Glu-C digests (M = 3449.56 Da). The

amino acid sequence was confirmed by the manual assignment of b- and y- product ions. The b- and y- ions are indicated in red and blue font, respectively. Sequences of internal ions are indicated in green font. The labelled masses refer to monoisotopic $[M+H]^+$ mass. The Cys (C) residue represents a carbamidomethyl-cysteine. Insets with black dashed border indicate the extracted ion chromatogram and isotopic distribution of the precursor ions.

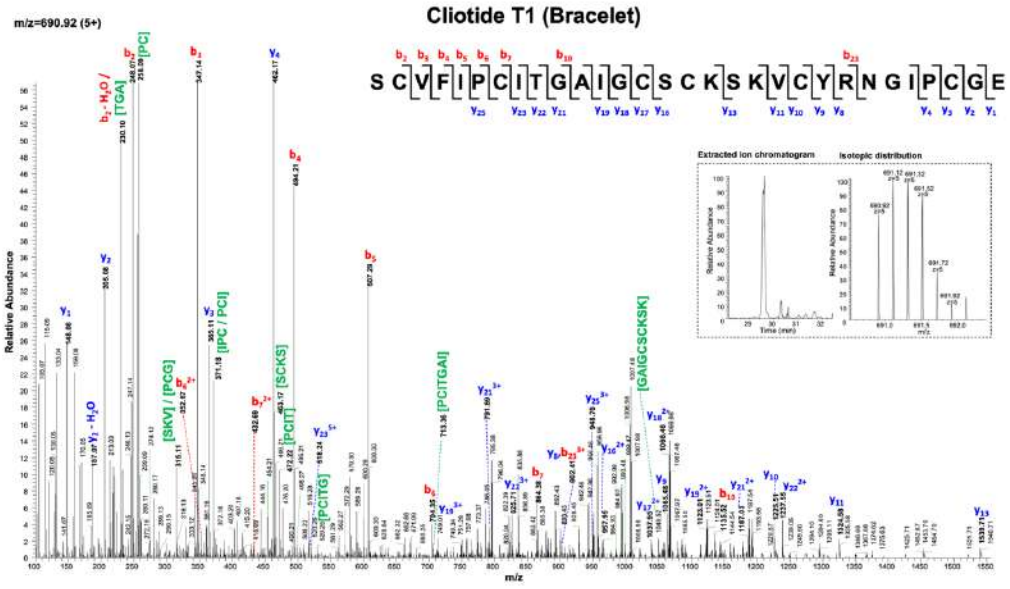


Figure S3.3: Peptide sequencing of 2+ charge state precursor ion of Cter M (m/z= 1712.70) from reduced, alkylated and endoproteinase Glu-C digests (M = 3423.39 Da). The amino acid sequence was confirmed by the manual assignment of b- and y- product ions. The b- and y- ions are indicated in red and blue font, respectively. Sequences of internal ions are indicated in green font. The labelled masses refer to monoisotopic $[M+H]^+$ mass. The Cys (C) residue represents a carbamidomethyl-cysteine. Insets with black dashed border indicate the extracted ion chromatogram and isotopic distribution of the precursor ions.

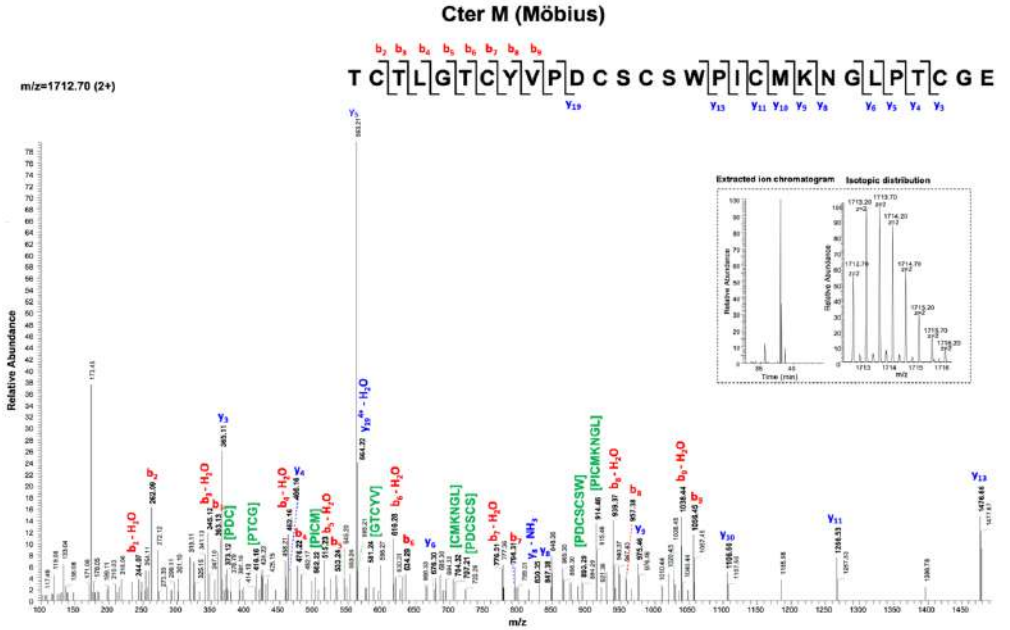
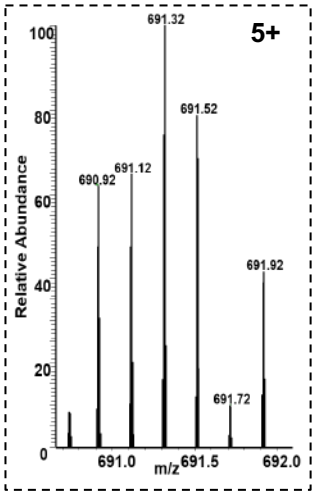
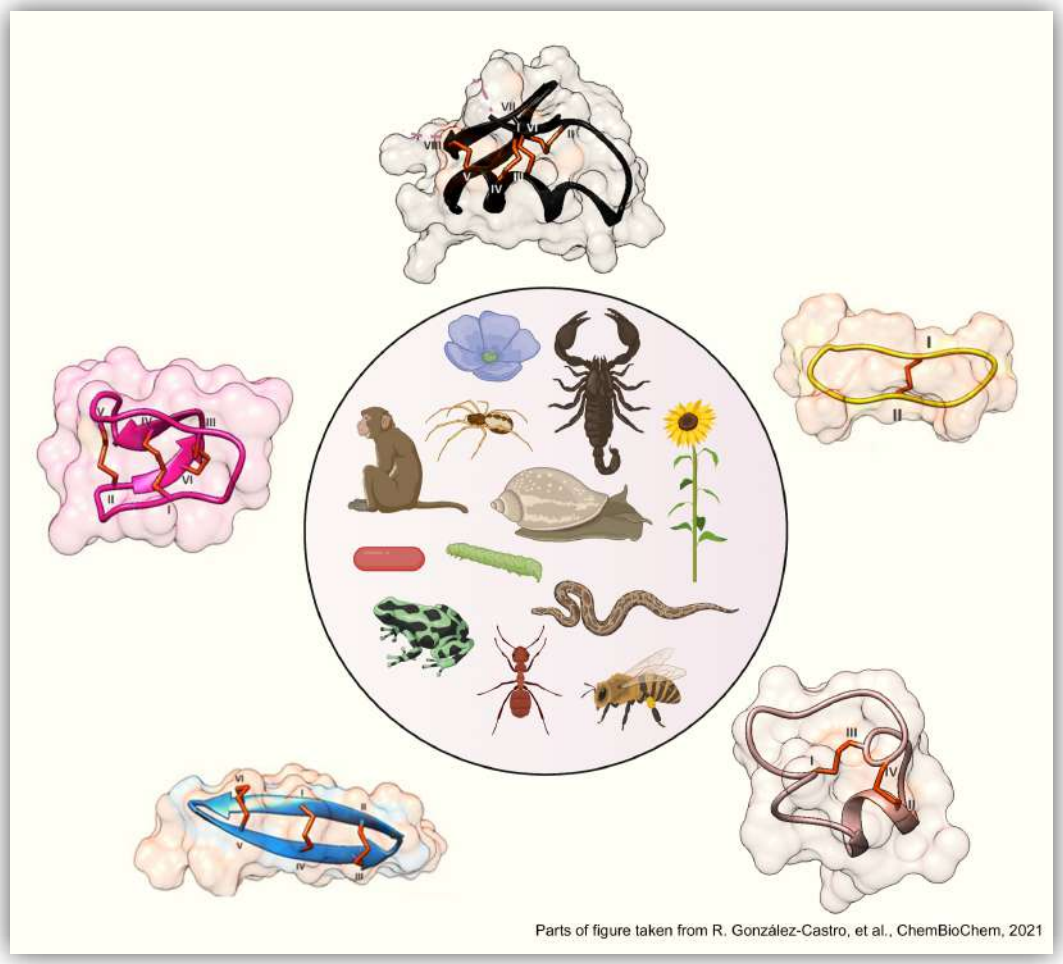


Figure S3.4: Sequences of three nearly identical masses of sequences of *C. ternatea* reported by three different research groups (Kalmankar et al., 2020; Nguyen et al., 2011; Poth et al., 2011a). Proline residues are highlighted in grey box.

Name	Sequence	Monoisotopic Mass (Da)
ctr pep 43	SCVFLPCFIIPGCSCKDKVCYLNGIPCGE	3083.35
cliotide T1	SCVFIPCITGAIGCSCKSKVCYRNGIPCGE	3083.36
Cter O	SCVFIPCITGIAGCSCKSKVCYRNGIPCGE	3083.36

Figure S3.5: Isotopic charge state distribution of 5+ charge state of Cliotide T1 from the stem extract of *C. ternatea*.





Parts of figure taken from R. González-Castro, et al., ChemBioChem, 2021

4

DSDBASE 2.0: Updated Version of DiSulphide DataBASE

Neha V. Kalmankar[1,2], Murugavel Pavalam[1], Sowmya Indrakumar[1] and Ramanathan Sowdhamini[1*]

Manuscript 4 (In preparation for submission to Nucleic Acid Research)

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore 560065, Karnataka, India.
[2]The University of Trans-Disciplinary Health Sciences and Technology (TDU), #74/2, Jarakabande Kaval, Post Attur, Via Yelahanka, Bangalore 560064, Karnataka, India

*Correspondence:
Prof. Ramanathan Sowdhamini
mini@ncbs.res.in

Manuscript 4

DSDBASE 2.0: Updated Version of DiSulphide DataBASE

Neha V. Kalmankar, Murugavel Pavalam, Sowmya Indrakumar and Ramanathan Sowdhamini*

4.1 ABSTRACT

Disulphide bonds are stabilizing crosslinks in proteins and serve to enhance their thermal stability. In proteins that are small and rich in disulphide bonds, they could be the major determining factor for the choice of conformational state since their constraints on appropriate backbone conformation can be substantial. Such crosslinks and their positional conservation could itself enable protein family and functional association. Despite the importance of the field, there is no comprehensive database on disulphide crosslinks that is available to the public. Herein we provide information on disulphides in DSDBASE2.0, an updated and significantly expanded database that is freely available, fully annotated and manually curated database on native and modelled disulphides. The web-interface also provides several useful computational tools that have been specifically developed for proteins containing disulphide crosslinks. The modelling of disulphide crosslinks is performed using stereochemical criteria, coded within our MODIP algorithm. The inclusion of modelled disulphides potentially enhances the loop database substantially, thereby permitting the recognition of compatible polypeptide segments that could serve as templates for immediate modelling. The DSDBASE2.0 database has not only been updated to include 153,944 PDB entries, 216,096 native and 20,153,850 modelled disulphide bond segments from PDB January 2021 release, the current database provides resource to user-friendly search for multiple disulphide bond containing loops, along with annotation of their function using GO and subcellular localization of the query. Further, it is now also possible to obtain three-dimensional models of disulphide-rich small proteins using an independent algorithm, RANMOD that generates and examines random, but allowed backbone conformations to the polypeptide. DSDBASE2.0 still remains as the largest open-access repository that organises all disulphide bonds of proteins in single platform. The database can be accessed from http://caps.ncbs.res.in/dsdbase2.

4.2 OVERVIEW

Disulphide (S-S) bonds, *i.e.* the covalent crosslinks between thiol groups of two cysteine residues, are recognized means of stabilizing native and folded proteins (Thornton, 1981). Such disulphide crosslinks are known to entropically destabilize the unfolded states of a polypeptide, limit the mobility and thereby increase the stability of the folded form and the native state of a polypeptide (Dombkowski et al., 2014; Flory, 1956). Approximately 20% of human proteins contain disulphide bonds (Martelli et al., 2004). Proteins with such domains are often referred to as disulphide-rich domains and they can exist both as single domain structures and as substructures within larger polypeptides. Such domains are typically present in extracellular proteins and peptides that include growth factors, hormones, enzymes, etc. Several small, disulphide-rich systems are also popularly known as bioactive peptides ranging from toxins, cyclotides, insulin and related growth factors, trypsin inhibitors to vasoconstrictors, antibacterial peptides, oxytocin, etc. Despite their abundance in sequence databases and biomedical importance, the three-dimensional structures are not available for many such proteins and peptides. Structural information can be derived by finding homologous entries in protein structure databanks based on conserved cysteine connectivity patterns. However, using existing methods such as PSI-BLAST (Altschul et al., 1997) generally tend to be unreliable for small peptides due to high sequence identity to wide variety of structural hits, thereby increasing the rate of false positives. Here, we introduce an updated database - DSDBASE2.0, that provides information on native disulphides, along with modelled disulphides, and those which are stereochemically possible between pairs of residues in a protein. We had previously developed the database for this purpose - DSDBASE (Vinayagam et al., 2004), which has been widely used and remains as one of the few available resources for such purposes. The current database has not only been substantially enhanced to include more recent PDB entries (**Table 4.1 & 4.2**), it also provides a resource for user-friendly search of multiple disulphide bond containing loops, along with their annotation of function using GO and active site residue information. Furthermore, it is also now possible to obtain three-dimensional models of disulphide-rich peptides using an independent algorithm, RANMOD (Sowdhamini et al., 1993) that generates and examines random, but allowed backbone conformations that a polypeptide can assume.

Table 4.1: Distribution of disulphide bonds in DSDBASE version 2.0

Dataset	No. of Proteins	Native Disulphides	Modelled Disulphides	Total Disulphides
Non-redundant_30 (Jan - 2021)	26,638	20,932	2,282,752	2,303,684
Non-redundant_90 (Jan - 2021)	43,200	51,932	5,012,608	5,064,540
Full database (Jan - 2021)	153,944	216,096	20,153,850	20,369,946

Table 4.2: Distribution of disulphide bonds in the DSDBASE previous version

Dataset	No. of Proteins	Native Disulphides	Modelled Disulphides	Total Disulphides
Non-redundant_25 (April-2001)	1,131	950	46,568	47,518
Full Database (PDB April-2001)	13,513	22,588	1,491,597	1,514,185
Non-redundant_25 (April-2002)	1,758	1,413	70,887	72,300
Full database (PDB July-2002)	17,163	28,141	1,973,898	2,002,039
Non-redundant_25 (April-2003)	2,849	2,170	147,722	149,892
Full database (PDB April-2003)	19,612	31,657	2,353,960	2,385,617

In theory, stability of proteins can be reduced by removal of native disulphide bonds. On the other hand, structural stability or rigidity of proteins can be improved by engineering new disulphide bonds. Several strategies, both experimental and computational, and empirical design rules have been proposed to introduce non-native disulphide bonds into protein structures in an attempt to improve its properties. However, successfully engineering a disulphide bond to increase stability still remains as a significant challenge. Computationally, tools such as Disulfide by Design (Craig and Dombkowski, 2013), SSBOND (Hazes and Dijkstra, 1988), MODIP (Sowdhamini et al., 1989), and limited others have been developed for identification of 'hot-spot' sites suitable for strainless disulphide bond introduction (Burton et al., 2000; Salam et al., 2014). In the updated database, the modelling of disulphides is performed by the MODIP algorithm, which essentially identifies pairs of residues where stereochemically covalent crosslinks can be accommodated without any strain. These crosslinks are further graded (A, B or C) according to their stereochemical quality. Moreover, it is observed that it is more effective to search in the current database for suitable templates and structural homologues to model disulphide-rich systems. The disulphide bond connectivity information is supplied by the user and the tool searches for substructures in the database to propose 3D models of

disulphide-rich polypeptides, in turn allowing the design of site directed mutants for enhanced stability of the protein. This forms another useful application of the database.

4.3 STRUCTURE OF THE UPDATED DATABASE

DSDBASE2.0 can be accessed from http://caps.ncbs.res.in/dsdbase2. The present release considers all entries in PDB-Jan 2021 release. The current database can be employed to search for loops with multiple disulphide bonds, along with annotation of their biological processes and molecular function, and subcellular localization of the query. A keyword search, PDB code search and EC number search options are made available. All possible pairs of residues which can strainlessly accommodate a disulphide bond and the stereochemical parameters of the disulphide bond are listed. Furthermore, there is a possibility to download all sub-structural motifs having specific loop sizes as well. Modelling peptides, an online search program to identify the nearest sub-structural homologue is provided. The input for that is the disulphide bond connectivity information for probing the database - and all sub-structural motifs with similar disulphide bonding pattern that can accommodate the restraints will be recognized. The modelling of disulphides is performed using MODIP algorithm. For the missing PDB entries, modelling can be performed by providing a PDB file as the input. **Figure 4.1** shows various screenshots from the database.

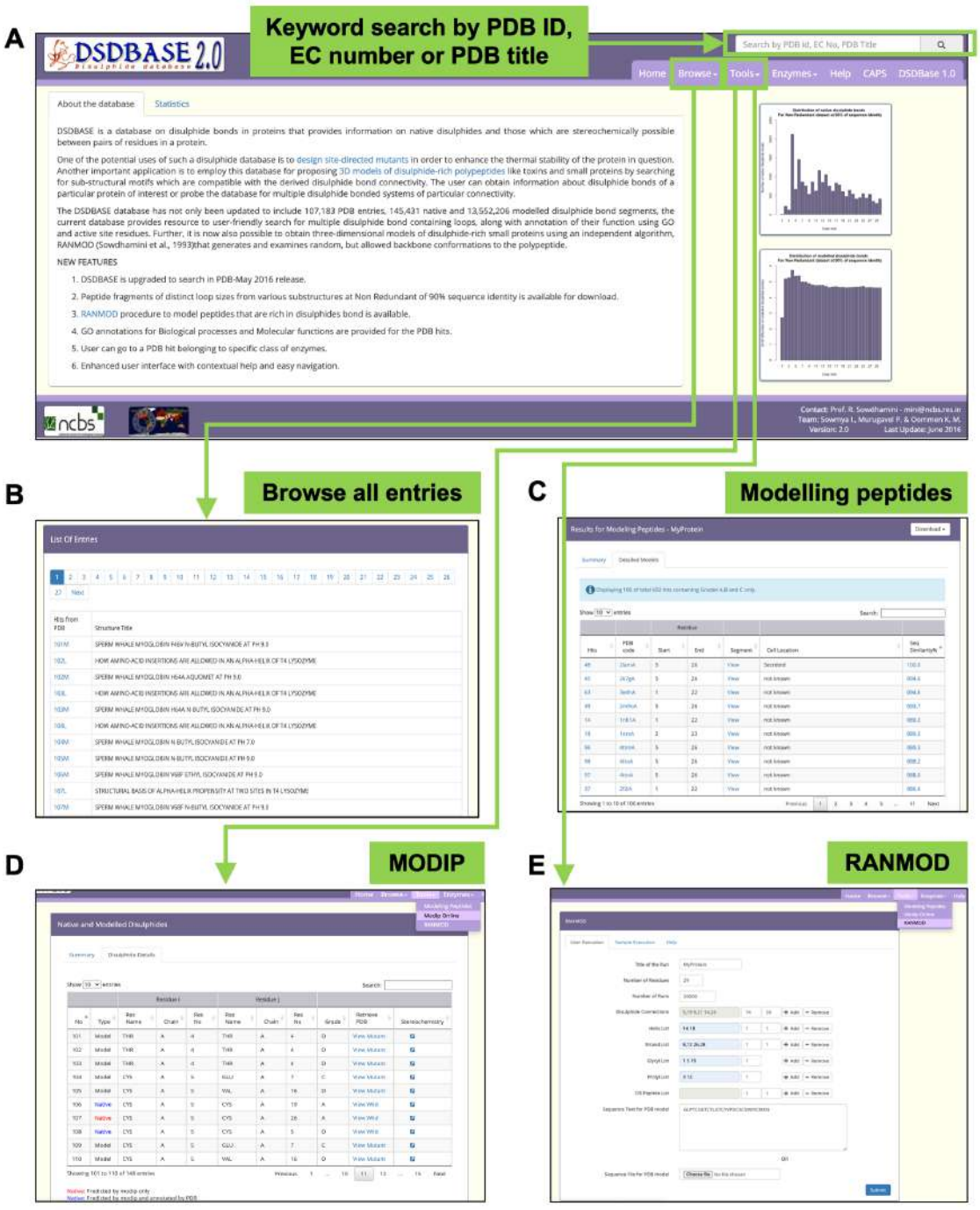


Figure 4.1: DSDBASE 2.0 database organisation and features. Different features of the database have been shown here. (A) Database browser. The users can browse through the different modules of DSDBASE. The user can also perform keyword search by PDB ID, EC number or PDB title. The distribution of native and modelled disulphide bonds based on loop sizes is provided. (B) List of all entries. The user can also download peptide fragments of distinct loop sizes from various substructures from nr-db-90% (90% sequence identity). (C) Modelling peptide feature. (D) RANMOD feature.

4.4 DATABASE STATISTICS

The DSDBASE2.0 database has been updated to include 153,944 PDB entries, 216,096 native and 20,153,850 modelled disulphide bond segments (**Table 4.1**). DSDBASE2.0 have also been updated to record 203,69,946 protein substructures that have stereochemical compatibility to accommodate disulphide bonds (**Table 4.3 & 4.4**)

Table 4.3: Number of substructures recorded in DSDBASE version 2.0

	Full database	**Non-redundant_30 (nr_30) database**	**Native of nr_30 database**
No. of proteins	153,944	26,638	4,544
No. of SS bonds	20,369,946	2,303,684	20,932
Grade A	26,25,229 (13%)	291,084 (13%)	12,702 (61%)
Grade B	54,31,859 (26%)	609,450 (26%)	3,230 (15%)
Grade C	62,20,158 (31%)	713,542 (31%)	3,027 (14%)
Grade D [a]	60,92,700 (30%)	689,608 (30%)	1,973 (10%)

[a] Cα–Cα and Cβ–Cβ distances were compatible but sulphur could not be fixed geometrically.

Table 4.4: Number of substructures recorded in DSDBASE previous version.

	Full database	**Non-redundant_30 (nr_30) database**	**Native of nr_30 database**
No. of proteins	19,612	2,849	766
No. of SS bonds	23,85,617	1,49,892	2,170
Grade A	3,07,003 (12%)	19,591 (13%)	1,321 (61%)
Grade B	6,33,431 (26%)	39,804 (27%)	341 (16%)
Grade C	7,28,374 (30%)	90,497 (60%)	508 (23%)
Grade D [a]	7,16,809 (29%)	-	-

[a] Cα–Cα and Cβ–Cβ distances were compatible but sulphur could not be fixed geometrically.

4.5 FEATURES AND INTERFACED TOOLS

4.5.1 MODIP online - Designing site-directed mutants

MODIP (Modelling Of Disulphides In Proteins) is available online and can be used to examine sites where disulphides can be introduced strainlessly (Sowdhamini et al., 1989). **Figure 4.2** depicts the output page of the MODIP run for a test protein for which stability data for disulphide engineered mutants were available. After the input of N, Cα and Cβ coordinates from the protein crystal structure, the program evaluates all possible Cα-Cα distances and the corresponding Cβ-Cβ distances. Those pairs of residues whose Cα-Cα distance is $\leq$ 6.5Å and Cβ-Cβ distance is $\leq$ 4.5Å are selected for fixation of a sulphur atom

at the two Cβ atom positions. Furthermore, the program calculates the S-S bridge parameters. A gradation scheme is used to evaluate the modelled disulphide, rank them based on certain stereochemical parameters such as dihedral angles and S–S bond distance (Sowdhamini et al., 1989). Thus, a disulphide-bridge modelled with dihedral angles and S-S bond distance within the accepted range is graded as 'A'. Grade 'B' are those disulphide-bridges where it is geometrically suitable to form the S–S covalent bond but with somewhat distorted stereochemistry. Sites which are spatially too close to form a disulphide bridge is graded as 'C'.

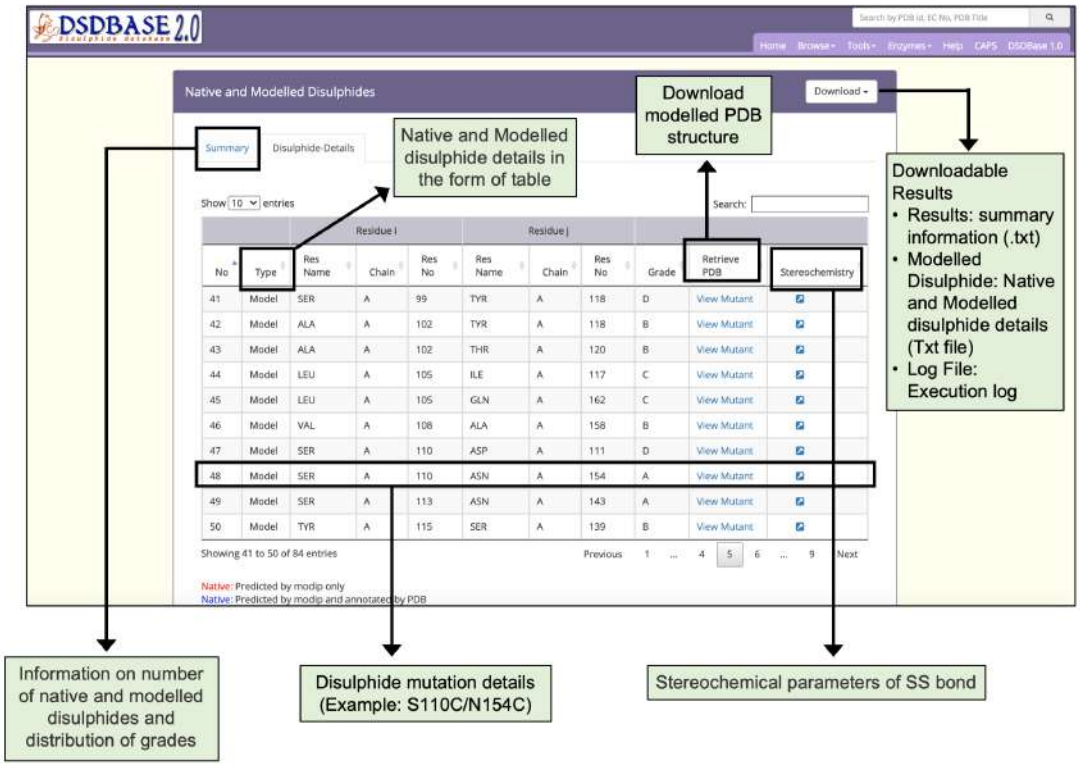


Figure 4.2: Results page for modelling of disulphide bonds in proteins (MODIP). A snapshot of the search output page shows the summary of native and modelled disulphides along with the grades and details of native and modelled disulphide in the form of table. The user should provide a structure (PDB format) for running MODIP procedure.

4.5.2 RANMOD - Random Conformation to Polypeptide Backbone

The independent algorithm, RANMOD, has been added as a new feature (Sowdhamini et al., 1993). **Figure 4.3** depicts the output page of the RANMOD run for a test protein. This method assigns a large number of permitted backbone conformations in a random fashion to the polypeptide, and identifies models which could accommodate strainless disulphide bridges. It uses the MODIP program to model the disulphide bridges (Sowdhamini et al., 1989). This method is beneficial for modelling disulphide-rich peptides for which there is no structure reported. This procedure has been applied on two and three disulphide-bond

containing peptides and found to give rise to several stereochemically acceptable structures. The search is time efficient, for instance, it takes less than 30 seconds to examine two hundred thousand conformations before finally converging at a stereochemically acceptable structure. Depending on the input seed number, RANMOD assigns random values to each of the backbone torsion angles, at each residue along the polypeptide chain. This implies various possible conformations with distinct seed number. The present version can distinguish glycine, proline and cysteine residues and rest of the residues are treated as alanine residues. It is up to the user to identify the "best" model as the present version does not rank the proposed models.

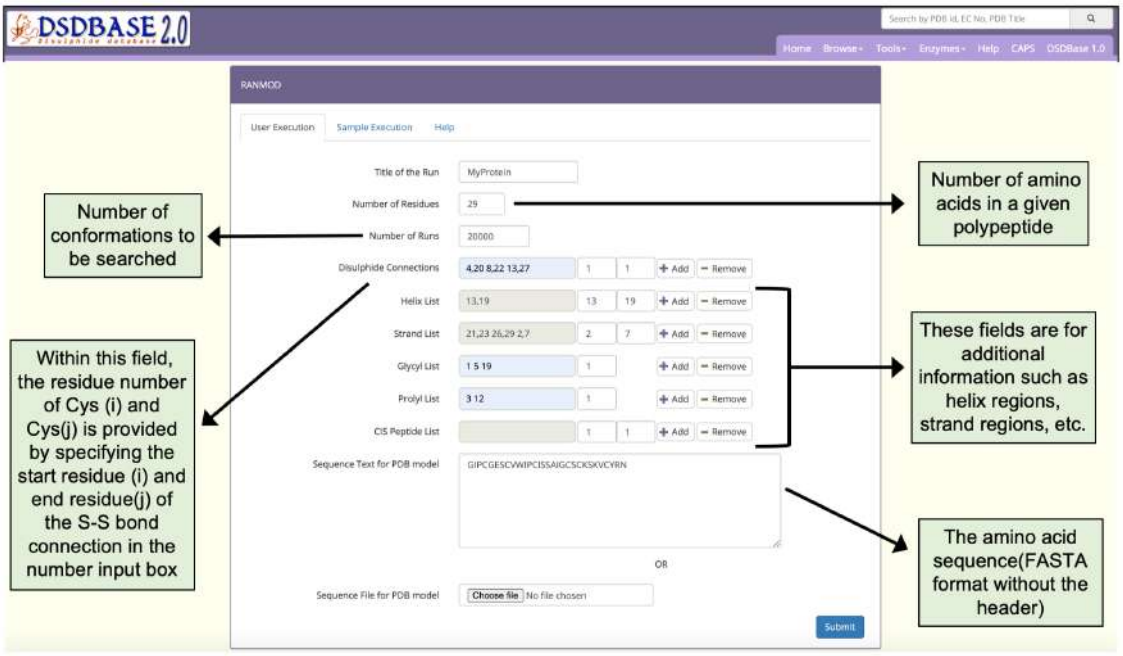


Figure 4.3: Snapshot of the input page for RANMOD (Random Conformation to Polypeptide Backbone) program. The user should provide a sequence (FASTA format) and disulphide connections information for running RANMOD procedure.

4.5.3 Modelling polypeptides

A search engine for modelling disulphide rich peptides has been added as a feature (Thangudu et al., 2005). Users can probe the database for multiple disulphide bonded systems of particular connectivity and get possible conformations for that segment with the option to provide partial or no sequence information. **Figure 4.4** depicts the output page of the search engine. This tool searches against a non-redundant database for homologues based on disulphide connectivity and/or sequence similarity details input by the user. The searches can be performed against the full PDB database (fulldb) or the non-redundant PDB database (nr-db-30%; 30% sequence identity and nr-db-90%; 90% sequence identity) or a subset of native disulphides of the non-redundant set of proteins

(nr-native-30%). The output generates protein or peptides hits or sub-structural motifs of a protein that shares similar connectivity pattern as that of the query. The predicted subcellular location of the query is also provided, and the native disulphide bonds and redox-active disulphide bonds are highlighted. The option to download the coordinate files of the hits and a graphical view of the segment via JSmol are also made available.

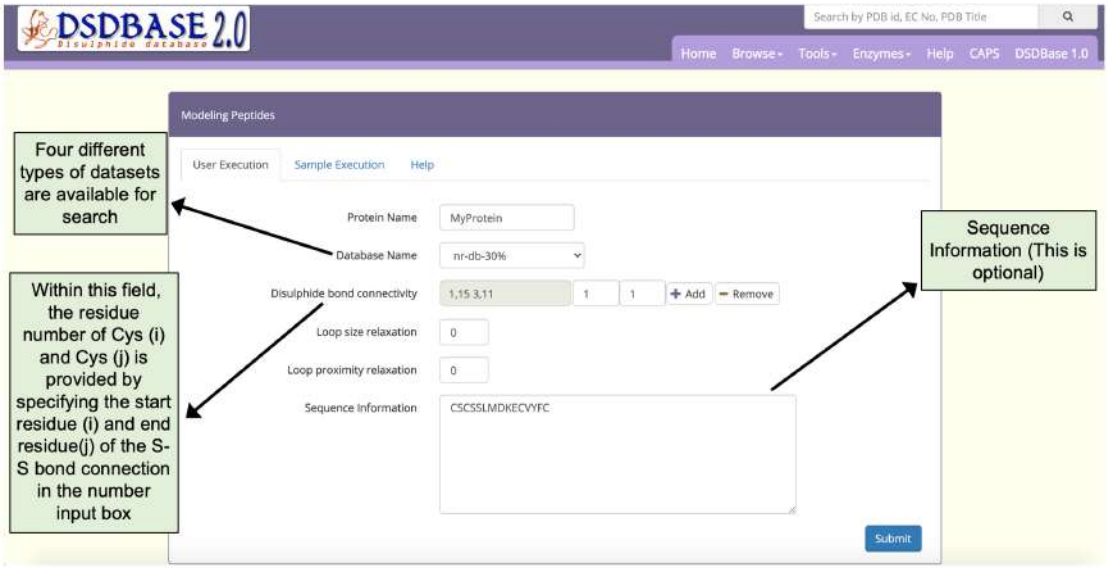


Figure 4.4: Snapshot of the input page for Modelling peptides. The user can probe the database for multiple disulphide bonded systems of particular connectivity and get possible conformations for that segment with the option to provide partial or no sequence information.

4.6 SCOPE OF THE DATABASE/CASE-STUDIES

4.6.1 MODIP

Xylanase: Xylanases are the most important xylan-degrading enzymes that are responsible for the hydrolysis of β-1,4 bonds in plant xylan, the main component of the hemicellulose. MODIP program was used to predict sites in an endo-β-1,4-xylanase from *Trichodrema reesei* (PDB ID: 1XYP) where cysteine mutagenesis can be sterically introduced thereby imparting stability to the protein (Törrönen and Rouvinen, 1995). The analyses revealed a key site of mutagenesis at positions S110C-N154C (**Figure 4.2**). The introduction of a disulphide bond at these sites is known to increase the thermodynamic stability of the protein and our prediction from the list supports this assertion (Xiong et al., 2004). The disulphide bridge was modelled with dihedral angles within an ideal range of values and was assigned grade 'A' based on the stereochemical parameters such as dihedral angles and S-S bond distance (**Figure 4.5**)

```
SER A 110    ASN A 154  CA -- CA =   6.392   CB -- CB =   3.852

Coordinates of              b(S-S)  ChiSS   Chi1I   Chi1J   Chi2I  Chi2J  Grade

SI   19.001  26.730   2.864
SJ   21.354  25.125   1.661  3.091   138.2   -45.4    53.3  -141.9   125.0   C

SI   19.001  26.730   2.864
SJ   19.070  26.724   1.254  1.611   -36.6   -45.4   164.4  -101.4  -130.4   C

SI   20.309  24.085   2.873
SJ   21.354  25.125   1.661  1.908   -78.0  -164.7    53.3   153.5  -177.5   A

SI   20.309  24.085   2.873
SJ   19.070  26.724   1.254  3.334   172.6  -164.7   164.4   114.8  -152.1   C
```

Figure 4.5: Stereochemistry of the predicted mutation site (S110C-N154C) for disulphide bridge modelling in the case of xylanase (PDB ID: 1XYP).

4.6.2 RANMOD

Cyclotide: Cyclotides are plant-derived, gene-encoded, triple disulphide, macrocyclic peptides of 26-37 residues length. They are characterized by a cyclic cystine-knot (CCK) motif, making them ultra-stable peptides against thermal, chemical and enzymatic degradation (Craik et al., 1999, 2002). A single model was generated using the RANMOD tool for the cycloviolacin O2 sequence (PDB ID: 2KNM) (Göransson et al., 2009) with 20,000 number of runs and no secondary structural clues provided in the input. The grades of the disulphide bridges for the model is as follows: Cys4-Cys20 Grade B, Cys8-Cys22 Grade C and Cys13-Cys27 Grade C. As cyclotide is naturally a cyclic peptide, the N- and C-terminal residues had to be cyclized. As this feature is not available within the RANMOD tool, we could generate the cyclizable conformation of cyclotide sequence externally using SYBYL software package version 7.1 (Tripos International, St. Louis, USA). The model generated by RANMOD was further energy minimized using Tripos forcefield using steepest descent method. As evident in **Figure 4.6**, the Cα deviation between RANMOD model and the known NMR model of cycloviolacin O2 sequence is bare minimum and the models showed an overall topological RMSD of ~6.7Å.

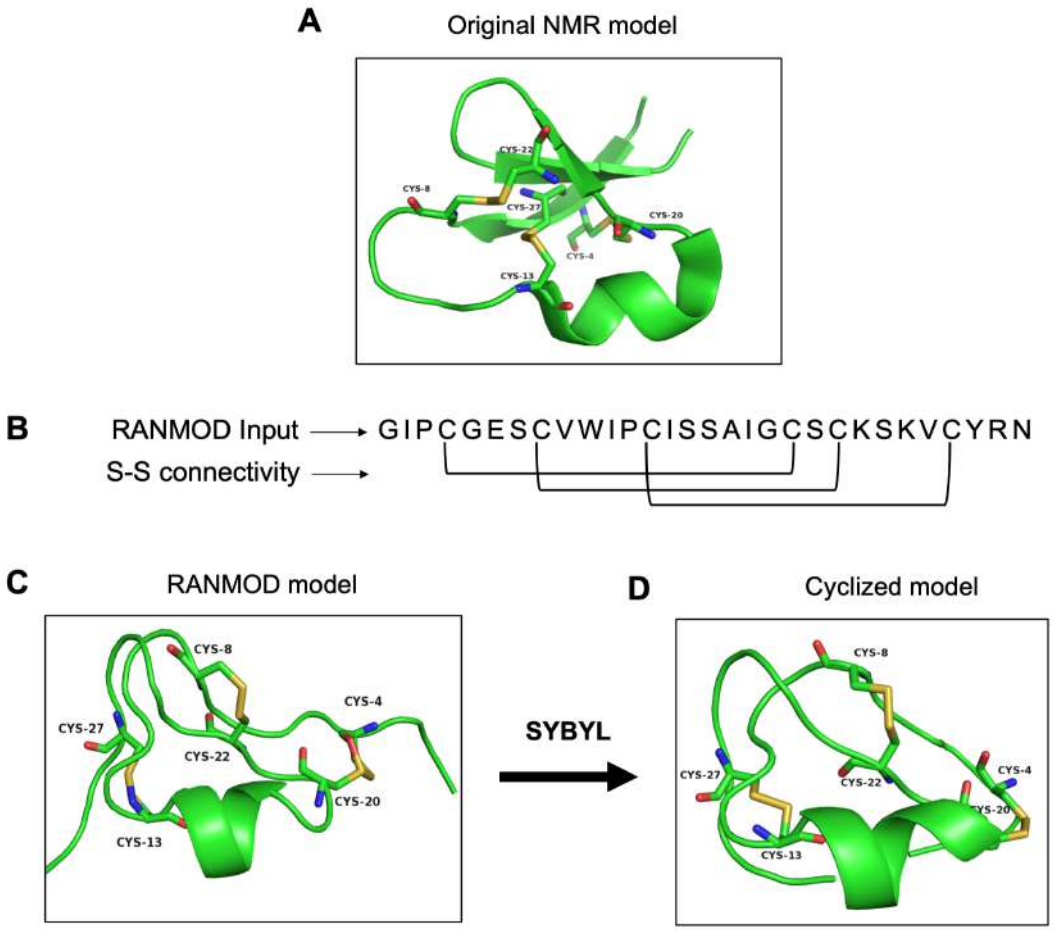


Figure 4.6: Model generated using RANMOD program for test case of cycloviolacin O2. (A) The NMR model (the first model of the ensemble) is shown. (B) The input sequence of cycloviolacin O2 from *Violoa odorata* (PDB ID: 2KNM) and disulphide connectivity information. (C) Cartoon/Ribbon view of the RANMOD generated model (D) The cyclized and energy minimized model.

Conotoxin: Conotoxin is a group of disulphide-rich, neurotoxic peptides present in the venom of marine cone snails. Their size typically ranges from 8-45 amino acid residues and can possess one to five disulphide bonds (Jin et al., 2019). Here, we demonstrate the structural modelling of a novel conotoxin, pc16a (PDB ID: 2LER), the first peptide characterised from a South-African cone snail *Conus pictus* of 11 residues sequence length (Van Der Haegen et al., 2012). Conotoxin pc16a contains two disulphide bonds (Cys2-Cys10 and Cys4-Cys11), which make the interlocked loops. In the first attempt of modelling (preliminary), no secondary structural information was provided, except the number of residues, the disulphide bond connectivity information and the full sequence of the peptide. This procedure, which involved 10,000 number of conformations to be searched, generated nine plausible random conformations for conotoxin pc16a. In the second attempt (advanced modelling), along with the sequence information, disulphide bond connectivity and peptide length information, any known secondary structural

information was also provided prior to the run. This procedure (10,000 runs) generated a single plausible random conformation for conotoxin pc16a. The superposed Cα-traces in the aligned regions of modelled structures obtained by the RANMOD procedure and native X-ray/NMR structures is shown in **Figure 4.7**. As the additional input parameters to the modelling procedure increased, structural similarity to the native template structure increased (**Figure 4.7**). It should be noted that generating conformations having interlocked disulphide loops is more difficult and complex compared to independent/loop-within-loop disulphide bridges. It was satisfying to note that several local structural features identified in native structures were indeed observed in the conformations obtained by the RANMOD procedure. These findings suggest strong agreement of the overall folding of the conotoxin model to the native structure originally proposed.

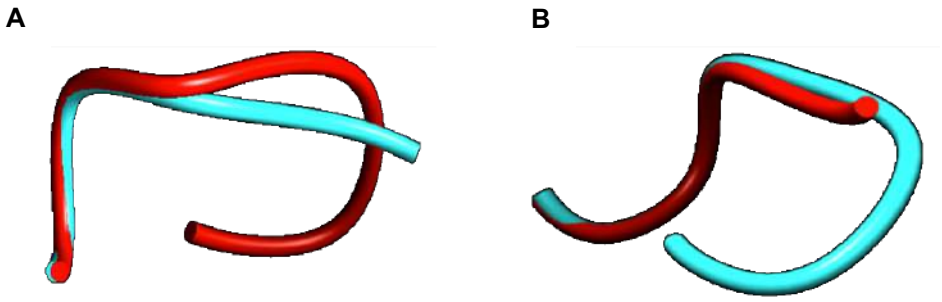


Figure 4.7: Superposed Cα-traces in the aligned regions for conotoxin - pc16a. (A) Structural superposition of RANMOD procedure generated model structure and native template structure of conotoxin, pc16a (PDB ID: 2LER). (B) Structural superposition of the model generated using more parameters (advanced modelling) and the native template structure. Red and cyan ribbon corresponds to backbone conformation of modelled structure and native template structure, respectively.

Buckwheat trypsin inhibitor: Trypsin inhibitors are types of serine protease inhibitors that reduces the biological activity of trypsin. A new peptide trypsin inhibitor (PDB ID: 2LQX) named BWI-2c was obtained from buckwheat (*Fagopyrum esculentum*) seeds by sequential affinity, ion exchange and reversed-phase chromatography (Oparin et al., 2012). The three-dimensional structure of the peptide was determined by NMR spectroscopy. The peptide is 41 amino acid residues long and a larger system compared to the previous cases (conotoxin and cyclotide). It contains four cysteines involved in two intramolecular disulphide bonds at (Cys11-Cys32 and Cys15-Cys2), forming the smaller disulphide loop within the bigger loop. We performed advanced modelling i.e. provided secondary structure clues. Superposed Cα-traces in the aligned regions and full-atom structure in the aligned region are depicted in **Figure 4.8**. Modelled BWI-2c structure when superposed over the native template structure indicates that the modelled structure adopts a similar

fold as native template. The secondary structural elements as specified prior to the run are indeed observed in the conformation obtained by RANMOD.

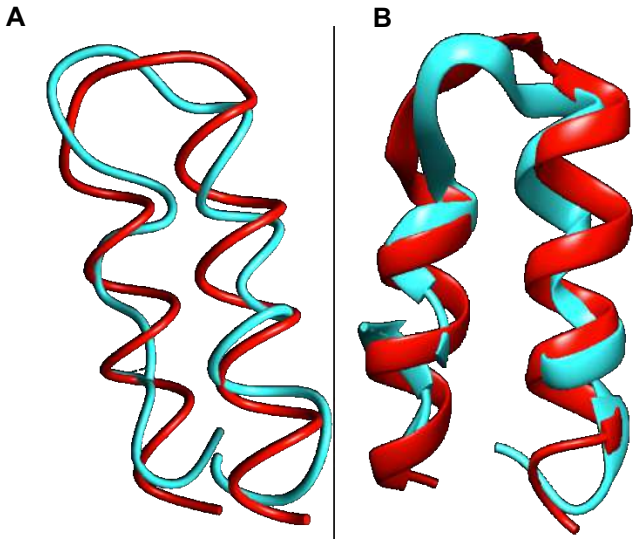


Figure 4.8: BWI-2c modelled structure superposed over native NMR structure. (A) Cα-traces of the aligned region. (B) Full-atom structure of the aligned region. Red and cyan ribbon corresponds to backbone conformation of modelled structure and native template structure, respectively.

4.6.3 Modelling disulphide-rich polypeptides

Disulphide Connectivity 4-19, 9-21 and 14-26: The DSDBASE2.0 server can be used for modelling disulphide-rich peptides with a particular disulphide bond connectivity by searching the query segment against the database for structures bearing similar connectivity. We used the 'Modelling peptides' tool for searching structural homologs having the disulphide connectivity C4-C19, C9-C21 and C14-C26. We made use of the option to provide sequence information and provided the sequence of a cyclotide - Cter M (GLPTCGETCTLGTCYVPDCSCSWPICMKN). We searched against the full PDB database (fulldb). The resulting output displayed a total of 184 hits containing all grades and the top matching hit as PDB ID 2FQA (chain A) with a sequence similarity of 76% with that of the query sequence. The output also provides the subcellular location of the protein hits as extracted from PDB files. The top hit structure we obtained corresponds to a linear cyclotide -Violacin A from *Viola odorata* and contains the exact same disulphide connectivity with that of the query sequence. The ability of the tool to categorically predict a near structural homolog based on the three disulphide-bonded system is noteworthy. **Figure 4.9** shows the segment output feature of the 'Modelling peptides' tool. This feature provides the user with the PDB file for the matched segment and can be viewed using JSmol view.

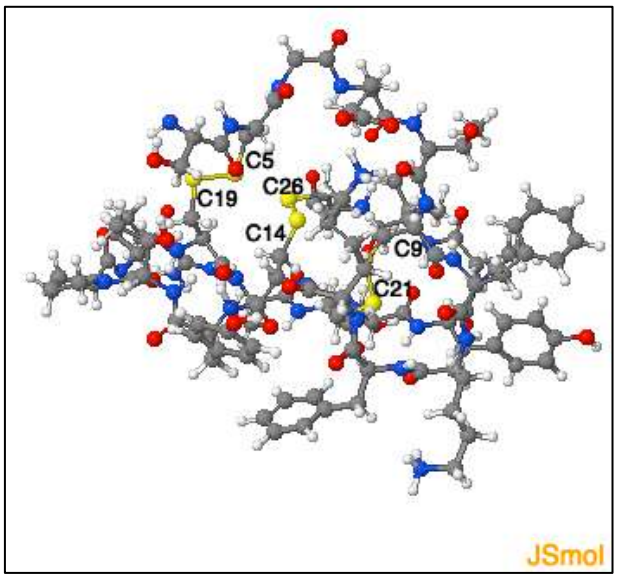


Figure 4.9: Representation of the three-dimensional structure segment of the homolog identified by querying "C4 - C19, C9- C21 and C14 - C26" disulphide connectivity in the 'Modelling peptide' tool. The cysteine residues and the corresponding disulphide bridges are highlighted.

4.7 DISCUSSION AND CONCLUSIONS

To the best of our knowledge, DSDBASE is the first database of its kind that organises all disulphide bonds of proteins in one platform. The DSDBASE2.0 database has not only been updated to include 153,944 PDB entries, 216,096 native and 20,153,850 modelled disulphide bond segments from PDB- Jan 2021 release, the current database provides resource to user-friendly search for multiple disulphide bond containing loops, along with annotation of their function using GO and active site residues. Further, it is now also possible to obtain three-dimensional models of disulphide-rich small proteins using an independent algorithm, RANMOD that generates and examines random, but allowed backbone conformations to the polypeptide. By the very nature of the RANMOD procedure, with constraints on disulphide bond connectivity, multiple conformations are proposed as 3D models of the polypeptide. In the future, efficient computational methods can be applied to fix side chains of all the residues. Using the installed a search engine for modelling disulphide-rich polypeptides, it is also possible to search the non-redundant databases using particular disulphide bond connectivity involving one or more disulphide bonds. The output is a list of proteins or peptides or sub-structural motifs of a protein bearing similar cysteine connectivity patterns. In the future we wish to update the database further by writing algorithms that can generate models using sequence alignment as an input. Furthermore, within the RANMOD program, we can incorporate computational methods that performs energy minimization routines of the proposed models. This will help the users to rank the proposed models.

4.8 REFERENCES OF CHAPTER 4

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* doi:10.1093/nar/25.17.3389.

Burton, R. E., Hunt, J. A., Fierke, C. A., and Oas, T. G. (2000). Novel disulfide engineering in human carbonic anhydrase II using the PAIRWISE side-chain geometry database. *Protein Sci.*

Craig, D. B., and Dombkowski, A. A. (2013). Disulfide by Design 2.0: A web-based tool for disulfide engineering in proteins. *BMC Bioinformatics*. doi:10.1186/1471-2105-14-346.

Craik, D. J., Daly, N. L., Bond, T., and Waine, C. (1999). Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J. Mol. Biol.* 294, 1327–1336. doi:10.1006/jmbi.1999.3383.

Craik, D. J., Simonsen, S., and Daly, N. L. (2002). The cyclotides: Novel macrocyclic peptides as scaffolds in drug design. *Curr. Opin. Drug Discov. Dev.*

Dombkowski, A. A., Sultana, K. Z., and Craig, D. B. (2014). Protein disulfide engineering. *FEBS Lett.* doi:10.1016/j.febslet.2013.11.024.

Flory, P. J. (1956). Theory of elastic mechanisms in fibrous proteins. *J. Am. Chem. Soc.* doi:10.1021/ja01601a025.

Göransson, U., Herrmann, A., Burman, R., Haugaard-Jönsson, L. M., and Rosengren, K. J. (2009). The conserved glu in the cyclotide cycloviolacin O2 has a key structural role. *ChemBioChem* 10. doi:10.1002/cbic.200900342.

Hazes, B., and Dijkstra, B. W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Eng. Des. Sel.* doi:10.1093/protein/2.2.119.

Jin, A. H., Muttenthaler, M., Dutertre, S., Himaya, S. W. A., Kaas, Q., Craik, D. J., et al. (2019). Conotoxins: Chemistry and Biology. *Chem. Rev.* doi:10.1021/acs.chemrev.9b00207.

Martelli, P. L., Fariselli, P., and Casadio, R. (2004). Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics*. doi:10.1002/pmic.200300745.

Oparin, P. B., Mineev, K. S., Dunaevsky, Y. E., Arseniev, A. S., Belozersky, M. A., Grishin, E. V., et al. (2012). Buckwheat trypsin inhibitor with helical hairpin structure belongs to a new family of plant defence peptides. *Biochem. J.* 446. doi:10.1042/BJ20120548.

Salam, N. K., Adzhigirey, M., Sherman, W., Pearlman, D. A., and Thirumalai, D. (2014). Structure-based approach to the prediction of disulfide bonds in proteins. in *Protein Engineering, Design and Selection* doi:10.1093/protein/gzu017.

Sowdhamini, R., Ramakrishnan, C., and Balaram, P. (1993). Modelling multiple disulphide loop containing polypeptides by random conformation generation. The test cases of α-conotoxin gi and edothelin I. *Protein Eng. Des. Sel.* 6, 873–882. doi:10.1093/protein/6.8.873.

Sowdhamini, R., Srinivasan, N., Shoichet, B., Santi, D. V., Ramakrishnan, C., and Balaram, P. (1989). Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng. Des. Sel.* 3, 95–103. doi:10.1093/protein/3.2.95.

Thangudu, R. R., Vinayagam, A., Pugalenthi, G., Manonmani, A., Offmann, B., and

Sowdhamini, R. (2005). Native and modeled disulfide bonds in proteins: Knowledge-based approaches toward structure prediction of disulfide-rich polypeptides. *Proteins Struct. Funct. Genet.* 58, 866–879. doi:10.1002/prot.20369.

Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* doi:10.1016/0022-2836(81)90515-5.

Törrönen, A., and Rouvinen, J. (1995). Structural Comparison of Two Major endo-1,4-Xylanases from Trichoderma reesei. *Biochemistry* 34. doi:10.1021/bi00003a019.

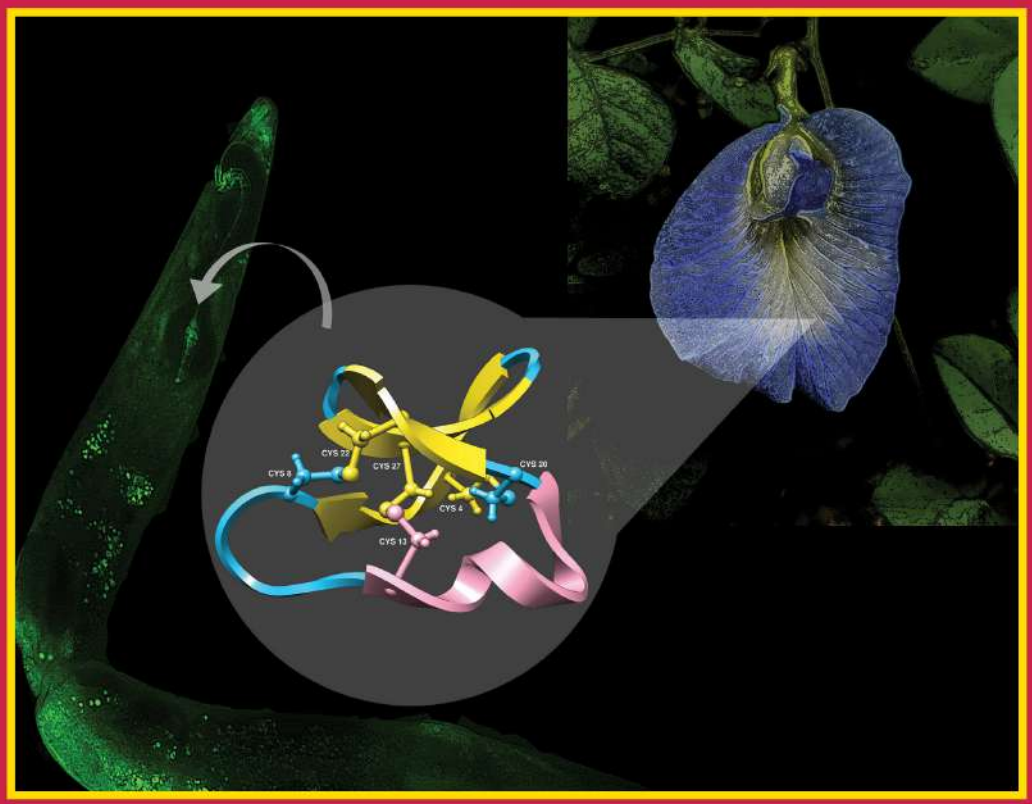
Van Der Haegen, A., Peigneur, S., Dyubankova, N., Möller, C., Marí, F., Diego-García, E., et al. (2012). Pc16a, the first characterized peptide from Conus pictus venom, shows a novel disulfide connectivity. *Peptides* 34. doi:10.1016/j.peptides.2011.10.026.

Vinayagam, A., Pugalenthi, G., Rajesh, R., and Sowdhamini, R. (2004). DSDBASE: A consortium of native and modelled disulphide bonds in proteins. *Nucleic Acids Res.* doi:10.1093/nar/gkh026.

Xiong, H., Fenel, F., Leisola, M., and Turunen, O. (2004). Engineering the thermostability of Trichoderma reesei endo-1,4-β- xylanase II by combination of disulphide bridges. *Extremophiles*. doi:10.1007/s00792-004-0400-9.

June 10, 2021
Volume 64 • Number 11
pubs.acs.org/jmc

Journal of Medicinal Chemistry



ACS Publications
Most Trusted. Most Cited. Most Read.
www.acs.org

5

Screening of Cysteine-Rich Cyclized Peptides as Inhibitors Against β-Amyloid Toxicity

Neha V. Kalmankar[1,2], Hrudya Hari[1], Ramanathan Sowdhamini[1], and Radhika Venkatesan[1,3*]

Manuscript 2
Journal of Medicinal Chemistry, 64, 11, 7422–7433 (2021)
https://doi.org/10.1021/acs.jmedchem.1c00033

Manuscript 6 (In preparation for submission to the Journal of Chemical Information and Modeling)

[1]National Centre for Biological Sciences (TIFR), GKVK Campus, Bangalore 560065, Karnataka, India.
[2]The University of Trans-Disciplinary Health Sciences and Technology (TDU), #74/2, Jarakabande Kaval, Post Attur, Via Yelahanka, Bangalore 560064, Karnataka, India
[3]Department of Biological Sciences, Indian Institute of Science Education and Research, Mohanpur 741246 West Bengal, India.

*Correspondence:
Dr. Radhika Venkatesan
radhika@ncbs.res.in

Manuscript 2

Disulfide-rich, cyclic peptides from *Clitoria ternatea* protect against β-amyloid toxicity and oxidative stress in transgenic *Caenorhabditis elegans*

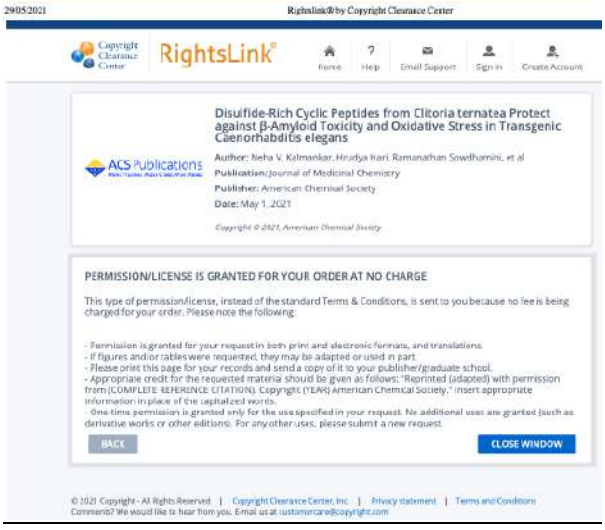
Neha V. Kalmankar, Hrudya Hari, Ramanathan Sowdhamini and Radhika Venkatesan*

5.1 SUMMARY

Cyclotides are disulfide-rich miniproteins found in plants with a unique head-to-tail cyclized backbone and three disulfide bonds forming a cyclic cystine-knot motif. The article provides an account of the peptide-based therapeutic applications of cyclotides against amyloid β-induced toxicity using transgenic *Caenorhabditis elegans*. These peptides show potent antioxidant activity by the reduction of intracellular ROS and inhibits β-amyloid peptides in transgenic *C. elegans* model of Alzheimer's disease.

5.2 RELATED INFORMATION

The published manuscript attached here refers to - Kalmankar NV, Hari H, Sowdhamini R and Venkatesan R. Disulfide-Rich, Cyclic Peptides from *Clitoria ternatea* Protect Against β-Amyloid Toxicity and Oxidative Stress in Transgenic *Caenorhabditis elegans*. J. Med. Chem. 2021, 64, 11, 7422–7433. DOI: 10.1021/acs.jmedchem.1c00033. This article has been incorporated here as it is without any changes and for non-commercial use as per the license conditions according to Copyright (2021) American Chemical Society. Please note: The page numbers indicated within the PDF is independent of rest of the thesis.



Journal of Medicinal Chemistry

pubs.acs.org/jmc Article

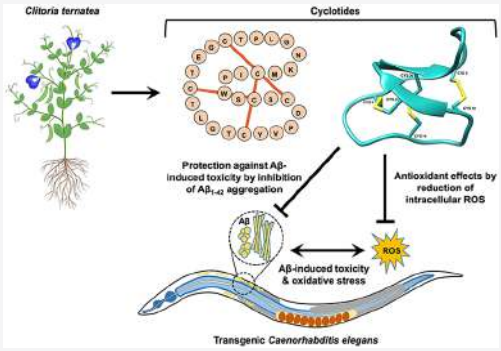
Disulfide-Rich Cyclic Peptides from *Clitoria ternatea* Protect against β-Amyloid Toxicity and Oxidative Stress in Transgenic *Caenorhabditis elegans*

Neha V. Kalmankar, Hrudya Hari, Ramanathan Sowdhamini, and Radhika Venkatesan*

Cite This: *J. Med. Chem.* 2021, 64, 7422−7433 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Neurotoxic aggregation of β-amyloid (Aβ) peptides is a hallmark of Alzheimer's disease and increased reactive oxygen species (ROS) is an associated process. In the present study, we report the neuroprotective effects of disulfide-rich, circular peptides from *Clitoria ternatea* (*C. ternatea*) (butterfly pea) on Aβ-induced toxicity in transgenic *Caenorhabditis elegans*. Cyclotides (∼30 amino acids long) are a special class of cyclic cysteine knot peptides. We show that cyclotide-rich fractions from different plant tissues delay Aβ-induced paralysis in the transgenic CL4176 strain expressing the human muscle-specific $A\beta_{1-42}$ gene. They also improved Aβ-induced chemotaxis defects in CL2355 strain expressing $A\beta_{1-42}$ in the neuronal cells. ROS assay suggests that this protection is likely mediated by the inhibition of Aβ oligomerization. Furthermore, Aβ deposits were reduced in the CL2006 strain treated with the fractions. The study shows that cyclotides from *C. ternatea* could be a source of a novel pharmacophore scaffold against neurodegenerative diseases.



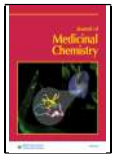
INTRODUCTION

Alzheimer's disease (AD) is a chronic and progressive neurodegenerative disorder characterized primarily by extracellular plaques of beta-amyloid (Aβ) peptide deposits and neurofibrillary tangles in the brain.[1] Majority of the strategies to combat AD have focused on inhibition of amyloid aggregation, inhibition of γ-secretases, stabilization of α-helical conformations, or elimination of aggregated amyloid peptides.[2] Among these strategies, identification of inhibitors of Aβ aggregates has gained prominence as the origin of neurotoxicity is often associated with the aggregation process. Despite studies investigating potential inhibitors of Aβ, only few have been approved for AD treatment, albeit considerable side effects and only symptomatic benefits in mild-to-moderate patients.[3] Plant metabolites such as ginkgolides from *Ginkgo biloba*, buckwheat trypsin inhibitor, cocoa peptide, and others have been tested against Aβ-induced toxicity.[4−6] These studies use transgenic *Caenorhabditis elegans* (*C. elegans*) that express human $A\beta_{1-42}$ as a model to investigate the toxicity of Aβ. *C. elegans* is widely recognized as an excellent model organism to study AD pathology.[7−9] Here, using *C. elegans*, we investigated the potential of cyclic peptides derived from an Indian medicinal plant in alleviating the Aβ-induced toxicity.

Clitoria ternatea (*C. ternatea*, butterfly pea) (Fabaceae) is widely used as a forage crop, natural food colorant, and in traditional Ayurvedic medicine.[10,11] Its extracts are reported to have a wide range of pharmacological activities.[12−14] *C. ternatea* has drawn significant medicinal interest, notably for the production of "cyclotides," a group of ultrastable macrocyclic peptides present in all plant tissues.[15] It is also the only known cyclotide-producing species in Fabaceae. Cyclotides are a novel class of gene-encoded, disulfide-rich, macrocyclic peptides (26−37 residues). They also form a conserved cyclic cystine knot motif by three disulfide bonds (Cys I−Cys IV, Cys II−Cys V, and Cys III−Cys VI).[16,17] This arrangement renders them exceptionally stable against enzymatic, chemical, and thermal degradation.[18] Peptides have few advantages over small molecules due to their low toxicity, nonimmunogenicity, better bioavailability, and convenient synthesis.[19,20] Quite a few studies on neuroprotection by peptides against Aβ toxicity have been reported.[6,21−24] However, none of these peptides have been developed into a clinical therapeutic for AD, but the potential does exist.

Although cyclotides are known to be produced in several plant species, de novo characterization is laborious and challenging due to varied sequences and abundance in different species, poor chromatographic resolution, difficult ionization,

Received: January 8, 2021
Published: May 28, 2021



https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

and deconvolution of complex structural arrangements. Nevertheless, with recent advances in RNA-sequencing technologies, peptide discovery and sequencing are relatively easier. Therefore, a combined peptidomic and transcriptomic approach is more useful and complementary in cyclotide characterization. Our group has recently sequenced and assembled the transcriptome of *C. ternatea* using RNA-seq data and 71 putative cyclotide sequences were identified.[25] Here, we present the peptidome of *C. ternatea*, characterized by matching the mass of cyclotides obtained from matrix-assisted laser desorption/ionization−time-of-flight (MALDI-TOF) experiments with the previously reported transcriptome library. In vivo anti-Aβ activity of cyclotides is shown using transgenic *C. elegans* (CL4176, CL2355, and CL2006) (Table 1) models that exhibit pathological behaviors associated with Aβ.[26] Overall, findings from our study provide novel insights into the potential of cyclotides as Aβ inhibitors.

Table 1. Description of the Transgenic *C. elegans* Used in the Study

strain	transgene	expression	phenotype
N2		control	wild-type movement
CL4176	*myo-3*/ Aβ_{1-42}	inducible in muscles	inducible larval paralysis and rapid paralysis
CL2006	*unc-54*/ Aβ_{1-42}	constitutive in muscles	adult onset paralysis and progressive paralysis
CL2355	*snb-1*/ Aβ_{1-42}	inducible in pan-neuronal	wild-type movement, chemosensory deficits, abnormal thrashing in liquid, and memory deficits

RESULTS

Profiling of Cyclotides from *C. ternatea*. Initial aqueous extracts were prepared from five tissues of *C. ternatea* (pods, stems, leaves, flowers, and roots) and were partially purified by reverse-phase C_{18} column chromatography. Each eluate was analyzed by matrix-assisted laser desorption/ionization−time-of-flight−mass spectrometry (MALDI-TOF-MS) to identify cyclotide-like masses (2500−4000 Da) before further purification. Fractions eluted with 50, 70, and 100% MeCN/ddH_2O solvents showed cyclotide-like masses. These eluates were pooled ("crude extract") for each plant tissue and further purified by preparative reversed-phase high-performance liquid chromatography (RP-HPLC) (Figure 1). Early (hydrophilic) and late (hydrophobic) eluting fractions were manually collected and separated into five fractions (A−E). Each fraction was analyzed for cyclotides via MALDI-TOF-MS using characteristic *m/z* values (2500−4000 Da). Fraction D was found to be richer in cyclotides compared to fractions A, B, C, and E (Figure 1 and Figures S1, S2, Supporting Information). Fraction D for each tissue is henceforth termed as cyclotide-rich fraction (CRF) and used in all the subsequent assays. The MALDI-TOF spectra for the CRFs from five tissues of *C. ternatea* are shown in Figure 2. Using the transcriptome assembly of *C. ternatea*,[25] we identified various cyclotide sequences by correlating the calculated mass of cyclotide genes and experimentally identified monoisotopic masses from MS data (Figure 2 and Table S1, Supporting Information). Out of the 67 cyclotide-like masses identified from all five tissue CRFs, 22 of them could be sequenced using nucleotide information (Table 2). Cyclotides commonly undergo post-translational modifications and processing that lead to increasing isoform occurrences and cyclotide diversity in a single plant species.[27,28] Therefore, it should be noted that some of the cyclotide masses reported in Table 2 cannot be identified.

Effect of Cyclotides on Aβ-Induced Paralysis in Transgenic AD *C. elegans*. CRFs from different tissues of *C. ternatea* were characterized for their effect on Aβ-induced

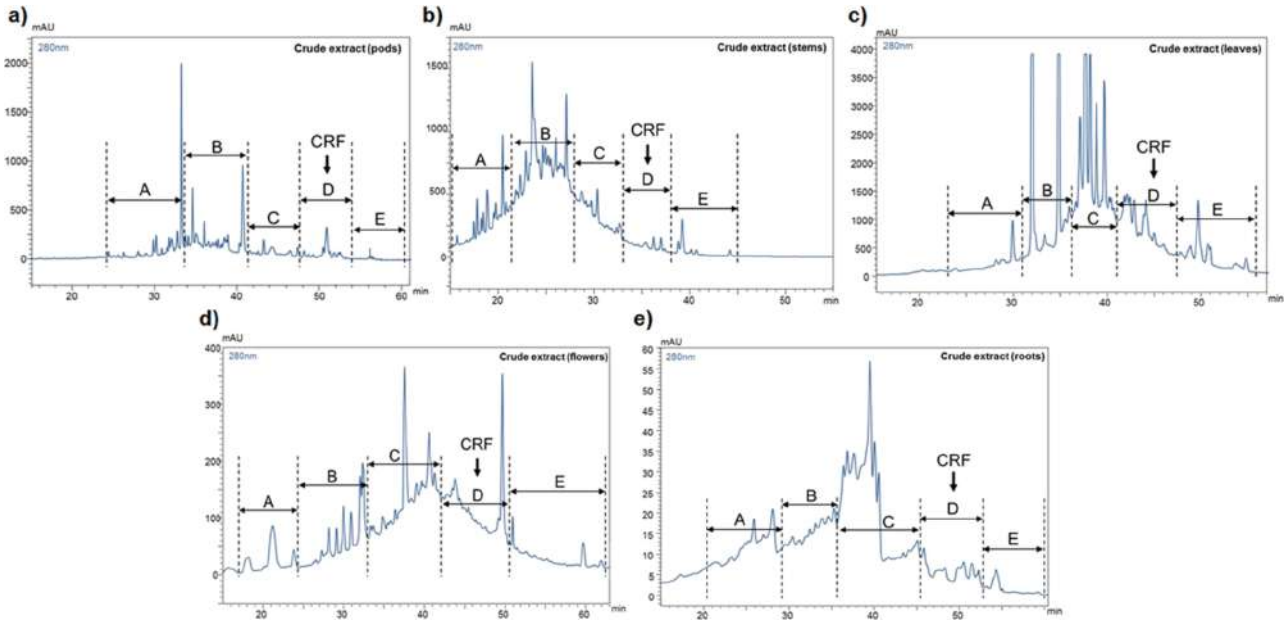


Figure 1. Semi-prep high-performance liquid chromatography (HPLC) profiles of crude cyclotide extracts obtained from (a) pods, (b) stems, (c) leaves, (d) flowers, and (e) roots of *C. ternatea*. A linear gradient of 1% min^{-1} of 0−95% buffer B (100% acetonitrile and 0.1% trifluoroacetic acid) was applied, and the eluents were monitored at 220, 254, and 280 nm. Five fractions (A−E) were collected from each plant tissue sample. The CRF (or fraction D) is highlighted with an arrow.

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

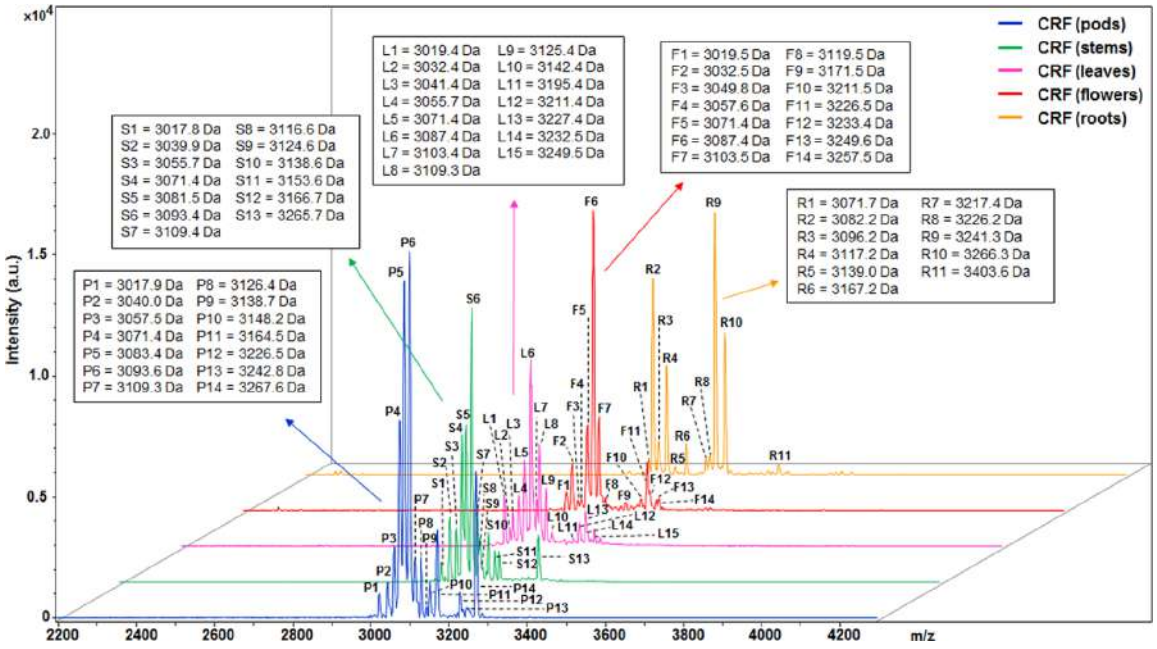


Figure 2. A stacked plot of MALDI-TOF mass spectra of CRF from different tissues of *C. ternatea*. The monoisotopic signals ($M^+ = m/z - H^+$) are labeled with names and their corresponding molecular weights are highlighted in the insets.

paralysis in *C. elegans* strain CL4176. This strain expresses full-length Aβ_{1-42} protein in the body wall muscles and shows a phenotype of Aβ expression and aggregation in muscles upon a temperature upshift from 15 to 23 °C, leading to progressive paralysis. Figure 3a illustrates the duration of treatment and paralysis assays in CL4176 and wild-type strain N2. Figure 3b shows microscopy images of CL4176 worms under normal and paralyzed conditions. Our results show that in an untreated control, the worms show complete paralysis within 100 h of temperature upshift. By comparison, the paralysis rate induced by Aβ was delayed in CL4176 worms treated with CRF by ~40% (Figure 3c–g), irrespective of which plant tissue it was extracted from. Based on previous reports, Vitamin C was chosen as a positive control for all the assays.[34−36] In the condition above, 80% of the untreated worms became paralyzed after 80 h, while only 50% of the Vitamin C-treated worms were paralyzed. CRFs from all the five tissues delayed the onset of paralysis in a dose−response manner. The highest concentration of CRF, that is, 200 μg/mL concentration displayed a significant delay in paralysis compared to the lower concentrations of cyclotides tested (Figure 3h). For example, when worms were treated with 200 μg/mL CRF from pods, ~44% of the worms were paralyzed at 84th h post-temperature upshift, whereas, ~53 and ~63% of the worms were paralyzed when treated with lower concentrations (80 and 20 μg/mL) of CRFs. This trend was observed in all tissue-wise treatments, where 200 μg/mL concentration of CRF showed the maximum desired activity (Figure 3h). As shown in Figure 3h, among CRFs from different tissues, treatment with pod CRF (concentration: 200 μg/mL) showed a significant delay in paralysis rates, (~56%) compared to the positive control (~46%) (p = 0.041 and n = 10) (Table S2, Supporting Information). These observations suggest that cyclotides have the potential to protect against Aβ-induced toxicity due to their ability to suppress paralysis in the whole animal. The wild-type N2 strain showed no paralysis at all at any time, regardless of treatment (Figure 3c–g).

Effect of Cyclotides on Aβ-Induced Chemotaxis Defects in Transgenic AD *C. elegans*. In AD, Aβ accumulation and aggregation are essentially toxic to the neuronal cells and the muscle-specific Aβ-expressing *C. elegans* model cannot specifically reflect the pathology of AD in the brain. Therefore, we assessed the neuroprotective effects of cyclotides on Aβ-induced toxicity in the neurons of *C. elegans*. We used the transgenic strain CL2355 that has an inducible, pan-neuronal expression of Aβ_{1-42} to investigate whether cyclotides can rescue the chemosensory deficits caused by Aβ-induced toxicity. Details of the experiment are illustrated in Figure 4a. Using chemotaxis index (CI) as a measure to evaluate attractive (positive value) and repulsive (negative value) responses, we tested the CRFs from different tissues of *C. ternatea* (Figure 4b–f). CRFs from all the tissues showed significant protection of neuronal cells against amyloid toxicity compared to untreated nematodes (negative control) (Figure 4b–f). Vitamin C which was used as a positive control in the experiment showed the maximum reversal of chemotaxis defects. Among CRF treatment, the most significant rescue was observed when CL2355 worms were treated with 200 μg/mL leaf CRF (CI positive control = 0.27 ± 0.13 and CI leaf CRF = 0.26 ± 0.15) (Table S3, Supporting Information). Additionally, the effect of CRFs on the CI was dose-dependent. A volume of 200 μg/mL concentration shows significant rescue in all except in the stem CRF treatment (Figure 4c). These results suggest that cyclotides could protect neuronal cells against Aβ-induced toxicity, thereby rescuing the associated defects in chemotaxis behavior.

Effect of Cyclotides on Reactive Oxygen Species (ROS) in Transgenic AD *C. elegans*. From the above results, dose-responsive protective effects of CRF on Aβ-induced toxicity are evident. Next, to evaluate the antioxidant effect of cyclotides in vivo, intracellular ROS levels were measured after

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

Table 2. *C. ternatea* Cyclotide Sequences Determined by Proteomic and Transcriptomic Mining

CRF tissue	[a]Protein sequence	Calculated monoisotopic mass (Da)	[b]Previous evidence	[c]RNA-seq evidence
Pods	—	P1=3017.9	Cter 12[29]	
	—	P2=3040.0		
	GISCGESCVFIPCITGIAGCSCKNKVCYLN	P3=3057.5	ctr pep 18[25]	✓
	GIPCGESCVFIPCISSVVGCSCKSKVCYNN	P4=3071.4	Cliotide T8[30]	✓
	GIPCGESCVFLPCFIIPGCSCKDKVCYLN	P5=3083.4	ctr pep 43[25]	✓
	—	P6=3093.6		
	GDALKCGETCFGGTCYTPGCSCDYPICKKN	P7=3109.3	ctr pep 4[25]	✓
	GIPCGESCVFIPCLTTVVGCSCKNKVCYNN	P8=3126.4	Cliotide T9[30]	✓
	—	P9=3138.7		
	—	P10=3148.2		
	GLPICGETCFKTKCYTKGCSCSYPVCKRN	P11=3164.5	Cliotide T15[30]	✓
	GIPCGESCVFIPCTVTALLGCSCKDKVCYKN	P12=3226.5	Cter R[30]	✓
	—	P13=3242.8	Cliotide T10[30]	
	GVIPCGESCVFIPCISTVIGCSCKNKVCYRN	P14=3267.6	Cter A[31]	✓
Stems	—	S1=3017.8	Cter 12[29]	
	—	S2=3039.8		
	—	S3=3055.7		
	GIPCGESCVFIPCISSVVGCSCKSKVCYNN	S4=3071.4	Cliotide T8[30]	✓
	—	S5=3081.5		
	—	S6=3093.4		
	GDALKCGETCFGGTCYTPGCSCDYPICKKN	S7=3109.3	ctr pep 4[25]	✓
	—	S8=3116.6		
	—	S9=3124.6		
	—	S10=3138.6		
	—	S11=3153.6		
	—	S12=3166.7		
	—	S13=3265.7	Cter E[15]	
Leaves	—	L1=3019.4		
	—	L2=3032.4		
	—	L3=3041.4		
	—	L4=3055.7		
	GIPCGESCVFIPCISSVVGCSCKSKVCYNN	L5=3071.4	Cliotide T8[30]	✓
	—	L6=3087.4		

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

Table 2. continued

CRF tissue	[a]Protein sequence	Calculated monoisotopic mass (Da)	[b]Previous evidence	[c]RNA-seq evidence
Leaves	—	L7=3103.4		
	GDALKCGETCFGGTCYTPGCSCDYPICKKN	L8=3109.3	ctr pep 4[25]	✓
	—	L9=3125.4		
	GSPCGESCVFIPCISTVIGCSCKNKVCYRN	L10=3142.4	ctr pep 30[25]	✓
	GRPTCGETCFKTKCYTKGCSCSYPVCKRN	L11=3195.4	ctr pep 8[25]	✓
	—	L12=3211.4		
	—	L13=3227.4		
	—	L14=3232.5		
	—	L15=3249.5		
Flowers	—	F1=3019.5		
	—	F2=3032.5		
	—	F3=3049.8		
	GISCGESCVFIPCITGIAGCSCKNKVCYLN	F4=3057.3	ctr pep 18[25]	✓
	GIPCGESCVFIPCISSVVGCSCKSKVCYNN	F5=3071.4	Cliotide T8[30]	✓
	—	F6=3087.4		
	—	F7=3103.5		
	—	F8=3119.5		
	GDLFKCGETCFGGTCYTPGCSCDYPICKNN	F9=3171.3	Cliotide T32[32]	✓
	—	F10=3211.5		
	GIPCGESCVFIPCTVTALLGCSCKDKVCYKN	F11=3226.5	Cter R[30]	✓
	VDGFCLETCVILPCFSSVAGCYCHGSTCMRG	F12=3233.4	Cliotide T37[32]	✓
	GVPCAESCVWIPCTVTALLGCSCKDKVCYLN	F13=3250.6	Cter B[15]	✓
	—	F14=3257.5		
Roots	GIPCGESCVFIPCISSVVGCSCKSKVCYNN	R1=3071.7	Cliotide T8[30]	✓
	—	R2=3082.2		
	—	R3=3096.2		
	—	R4=3117.2		
	—	R5=3139.0	Cter 30[29]	
	—	R6=3167.2		
	—	R7=3217.4		
	GIPCGESCVFIPCTVTALLGCSCKDKVCYKN	R8=3226.2	Cter R[30]	✓
	—	R9=3241.3		
	—	R10=3266.3		
	—	R11=3403.6		

[a]Sequences with calculated monoisotopic masses shared between the tissues have been underlined. [b]Obtained from *C. ternatea* transcriptome[25] and/or CyBase (a database of cyclic proteins).[33] [c]The existence of reported RNA-seq evidence of cyclotides[25] is shown with check marks.

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

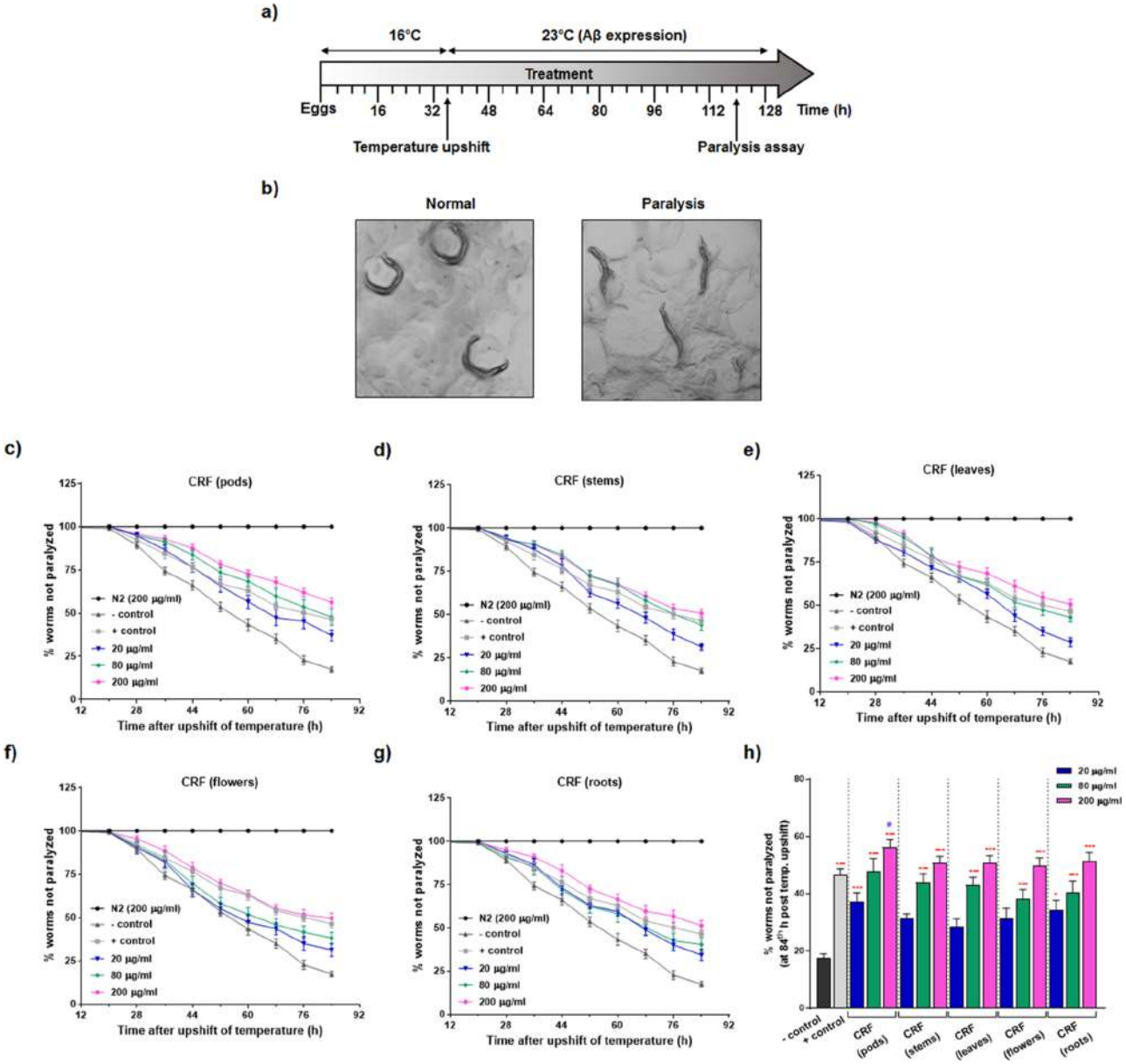


Figure 3. Effect of CRFs from different tissues of *C. ternatea* on Aβ-induced paralysis in *C. elegans* strain CL4176 and wild-type strain N2. (a) Illustrative diagram showing the duration of treatment and paralysis assays in CL4176 and wild-type strain N2 worms. (b) Microscopy images showing CL4176 worms under normal and paralyzed conditions. (c–g) Time course of Aβ-induced paralysis in the transgenic CL4176 strain and wild-type N2 strain treated with a vehicle (negative control) or CRFs (20, 80, and 200 μg/mL) from different tissues or Vitamin C (positive control). (h) The rate of paralysis at 84th h post-temperature upshift for the treated and untreated transgenic worms. Data represent mean ± SE % paralysis from at least 10 independent experiments. Red and blue symbols represent statistically significant differences between CRF treatment and the negative and positive control groups, respectively (*/$^{\#}$, $p < 0.05$; **/$^{\#\#}$, $p < 0.01$; ***/$^{\#\#\#}$, $p < 0.001$).

feeding the transgenic strain CL4176 with 200 μg/mL concentration of CRFs from different tissues. Significantly lower fluorescence intensities were observed in nematodes treated with CRFs compared to untreated control nematodes expressing Aβ (Figure 5a–c and Table S4, Supporting Information). In particular, treatment with 200 μg/mL of CRF from flowers of *C. ternatea* exhibited the most significant reduction of intracellular ROS (~20% reduction from untreated worms) (Figure 5c and Table S4, Supporting Information). This reduction was higher (~7%) compared to the positive control as well. These results suggest indirectly that the delay observed in Aβ-induced paralysis of CL4176 nematodes after treatment with CRFs could be due to reduced accumulation of intracellular ROS.

Effect of Cyclotides on Aβ Deposits in Transgenic AD *C. elegans*. To investigate the effect of cyclotides on Aβ aggregation, we performed thioflavin-S (ThS) fluorescence assay to visualize the formation of amyloid fibrils. The transgenic *C. elegans* strain CL2006, which expresses an Aβ protein fragment involved in the development of AD, was used for this assay. The strain shows a phenotype of constitutive Aβ expression and aggregation in muscles, leading to progressive

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422–7433

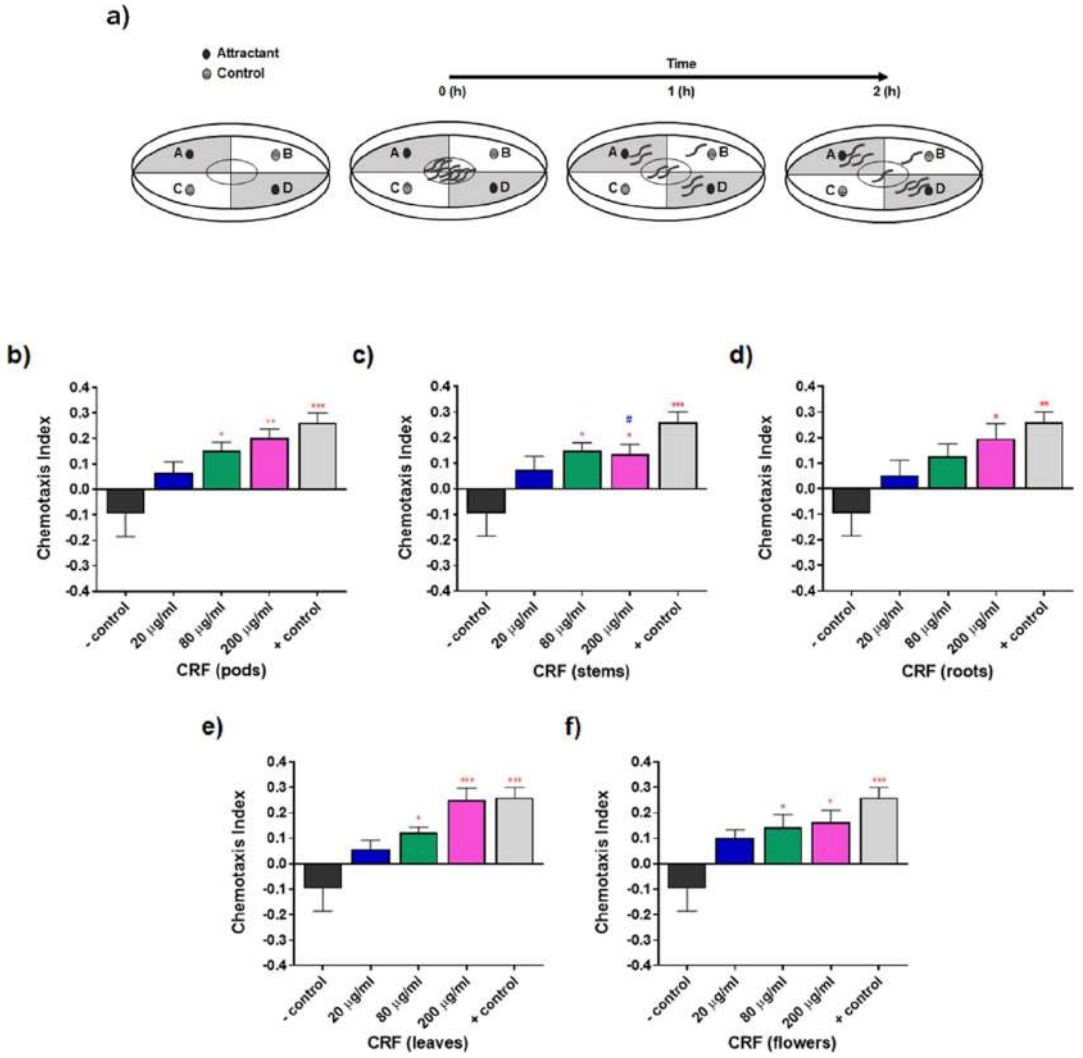


Figure 4. Effect of CRFs from different tissues of *C. ternatea* on Aβ-induced defects in odorant preference in *C. elegans* strain CL2355 and wild-type strain N2. (a) Experimental design for the chemotaxis behavior assay performed with the transgenic *C. elegans* strain CL2355 (neuronal Aβ strain) and wild-type strain N2 using 0.1% benzaldehyde in 100% ethanol as an attractant and 100% ethanol as a control odorant. (b–f) Chemotaxis behavior assay of the transgenic *C. elegans* strain CL2355. Data represent mean ± SE of CI from at least 10 independent experiments. Red and blue symbols represent statistically significant differences between CRF treatment and the negative and positive control groups, respectively (*/$^{\#}$, $p < 0.05$; **/$^{\#\#}$, $p < 0.01$; ***/$^{\#\#\#}$, $p < 0.001$).

paralysis. CL2006 worms were subjected to Aβ staining using ThS at the end of the CRF treatment. Fluorescence images of the head region of the CL2006 worms show significant reduction in Aβ deposits (blue arrowheads) in worms treated with CRFs (concentration: 200 μg/mL) compared to untreated worms (negative control) (Figure 6). The wild-type N2 strain shows no Aβ deposition in the whole animal. These results suggest that cyclotides inhibit Aβ aggregation and deposition in the transgenic worms' muscle cells.

DISCUSSION AND CONCLUSIONS

This study investigates the potential of cyclotides from *C. ternatea* as an inhibitor of Aβ aggregation using *C. elegans*. Synthetically created cyclotides, analogous to the *Oldenlandia affinis* plant-derived peptide, have been shown to suppress multiple sclerosis when administered orally.[37] More recently, antineurodegenerative properties of cyclotides from *Psychotria solitudinum* enable them to act as inhibitors of human prolyl oligopeptidase, a promising target for the treatment of cognitive deficits in several psychiatric and neurodegenerative diseases.[38] However, until now, there have been no reports on pharmacological activities of cyclotides derived from *C. ternatea* for the treatment of neurodegenerative disorders such as AD, Parkinson's disease, or Huntington's disease. In the present study, we isolated CRFs from five tissues of *C. ternatea* and identified 67 cyclotide-like masses from MS experiments, of which 22 were sequenced using the available nucleotide-level information. Using the CL4176 strain, we tested the effects of different CRF concentrations from various tissues. Our data show a significant delay in Aβ-induced paralysis in nematodes treated with CRF, irrespective of the plant tissue, and the effects were concentration-dependent, underscoring the neuroprotective effects of cyclotides. The lower doses of CRF tested (20 and 80 μg/mL) produced less significant protective effects; however, the highest dose tested (200 μg/mL) delayed the paralysis phenotype drastically

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422–7433

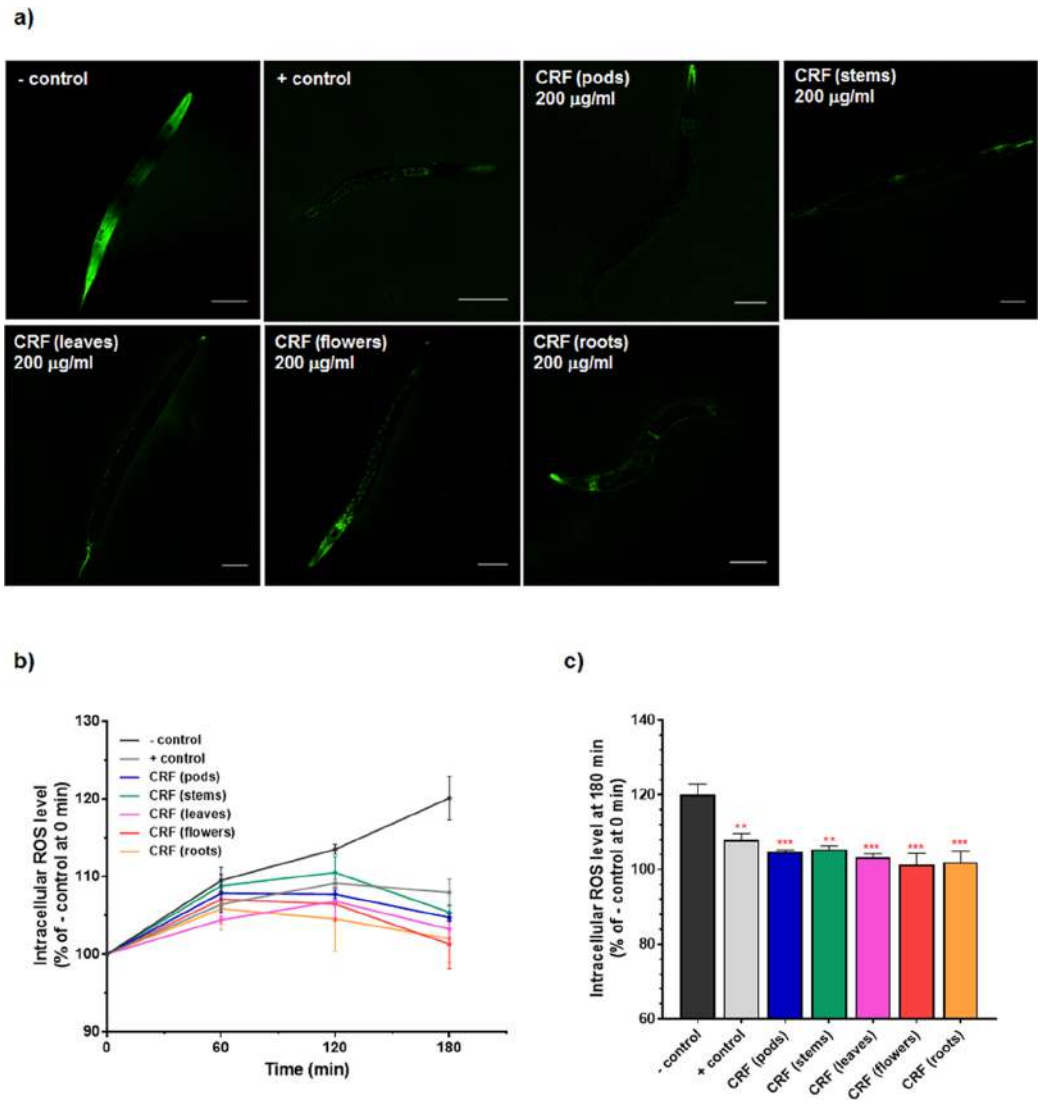


Figure 5. Effect of cyclotides on the intracellular ROS production in the transgenic *C. elegans* CL4176 strain. Age-synchronized worms were treated with or without CRFs followed by a dichlorofluorescein (DCF) assay for ROS estimation. (a) DCF fluorescence images of untreated nematodes (negative control) versus treatment with CRFs from different tissues of *C. ternatea*. Scale bars: 100 μm. (b) Time course of increase in DCF fluorescence detected using a microplate reader at 485 nm excitation wavelength and 530 nm emission wavelength. Results are expressed as percentage of fluorescence (intracellular ROS level) relative to the untreated control (negative control) at t = 0, which is set as 100%. (c) Corresponding % DCF fluorescence observed at t = 180 min. Data represent mean ± SE of % DCF fluorescence from four replicates of experiments. Red asterisks represent statistically significant differences between CRF treatment and the negative control group (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$).

compared to the untreated and positive controls. Many compounds such as scorpion venom peptides,[22] sesamin and sesamolin,[39] and tryptophan-containing heptapeptides[24] exert protective effects against Aβ toxicity tested using the same strain of *C. elegans*. With the development of newer transgenic *C. elegans* models, Aβ expression has been made to express pan-neuronally.[4] The CL2355 strain has inducible neuronal Aβ_{1-42} expression, which leads to defects in 5-HT sensitivity, chemotaxis behavior, and altered movement in liquids.[4] Here, under normal conditions, CL2355 nematodes exhibit chemotaxis dysfunction to volatile cues, but upon CRF treatment, chemotaxis was rescued to a reasonable extent indicating protection against Aβ-induced toxicity in the neuronal cells. The lowest dose of CRF tested (20 μg/mL) was not sufficient to alleviate Aβ toxicity; however, higher doses (80 and 200 μg/mL) improved the rescue of chemotaxis defects. The observed effect could also be due to the presence of other molecules in the CRF; however, we speculate that the effect might be mainly due to cyclotides as CRF contains predominantly cyclotides. Understanding the role of individual cyclotide from CRF warrants further investigation given the multiple challenges with separation and characterization.[40]

Oxidative stress is theorized to be a major contributing factor to AD as it is induced by β-amyloid aggregation.[41−43] Several pieces of evidence correlate AD pathogenesis with increased oxidative stress resulting from increased ROS. Few cyclotide-containing plants such as *Clitoria* spp., *Viola spp.*, *Morinda spp.*, and *Psychotria spp.* have also been reported to

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

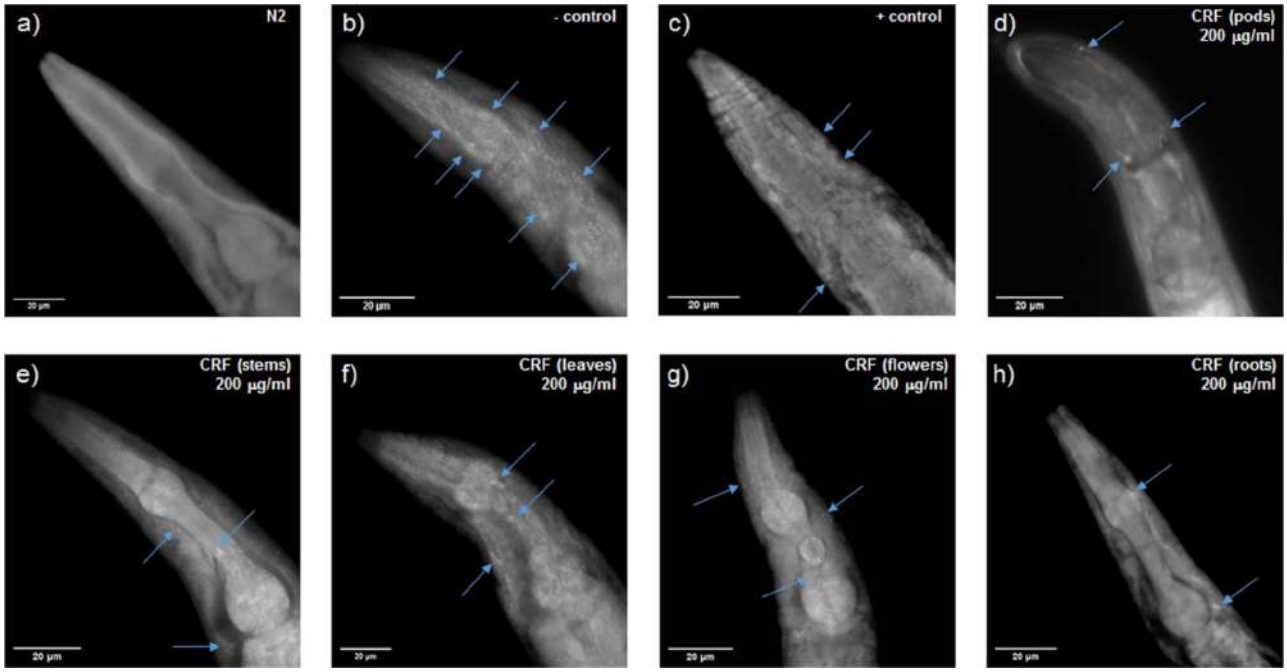


Figure 6. Representative images of Aβ plaques in the head region of the transgenic strain CL2006. Age-synchronized L1 worms were spotted on plates with OP50 at 20 °C for 72 h. Worms were then transferred on nematode growth medium (NGM) plates and treated with vehicle (negative control) or CRFs (concentration: 200 μg/mL) from different tissues or Vitamin C (positive control) for 48 h. Worms were then fixed, permeabilized, and stained with 0.125% ThS. Representative images of (a) wild type, (b) negative control, (c) positive control, (d) CRF from pods, (e) CRF from stems, (f) CRF from leaves, (g) CRF from flowers, and (h) CRF from roots. Arrows indicate Aβ deposits. Scale bars: 20 μm.

exhibit potent antioxidant properties.[44−48] However, to date, there has not been a conclusive study attributing the antioxidant activity solely to cyclotides. These studies have reported high concentrations of other secondary metabolites to co-occur with cyclotides in the isolated fractions.[47] In the present study, feeding CL4176 worms with CRF from flowers of *C. ternatea* exhibited the most significant reduction in intracellular ROS. The fractions used here are mixtures of several cyclotides at a concentration of 200 μg/mL, suggesting that the components of the flower peptide mixture are more active than the known antioxidant, Vitamin C. However, as pure cyclotides were not used, the antioxidant activity observed could be an enhanced effect of other molecules such as anthocyanins or aglycones or other colorless molecules present in the CRF functioning in combination with cyclotides. It has been reported earlier that in cysteine-stabilized peptides, the cysteine and methionine residues have antioxidant activities as they can directly react with free radicals.[49] Here, the six cysteine residues in cyclotides could be contributing to the observed antioxidant activity. Next, using CL2006, we qualitatively assessed the amount of Aβ deposition in the neurons of worms treated with CRF and observed that the deposits were reduced after treatment with the highest dose in line with earlier observations.

In conclusion, our study reports 22 cyclotide sequences in the CRFs using a combined transcriptomic and proteomic approach and provides a systematic activity evaluation of the cyclotides identified. We show that CRFs can (1) attenuate Aβ-induced paralysis in the muscle cells of transgenic CL4176 worms, (2) rescue chemosensory defects associated with Aβ-expression in the neuronal cells of transgenic CL2355 worms, (3) diminish ROS production caused by oxidative stress in transgenic CL4176 worms, and (4) reduce Aβ plaques and oligomers in transgenic CL2006 worms. Taken together, our findings highlight the neuroprotective effects of cyclotides from *C. ternatea* and their immense potential in peptide therapeutics as Aβ inhibitors.

EXPERIMENTAL SECTION

Extraction of Cyclotides from *C. ternatea*. Plant collection and cyclotide extraction were performed using established methods. Briefly, fresh plant parts such as pods, stems, leaves, flowers, and roots were collected and oven-dried at 70 °C. Each tissue was extracted with 1:1 DCM:MeOH (v/v) with overnight stirring. The extract was partitioned with 1:1 mixture of DCM:(MeOH:ddH_2O, 2:3) and fractionated on a packed RP-C_{18} silica gel column (0−100% gradient elution using MeCN/ddH_2O) (particle size: 40−63 μm; pore size: 60 Å; Polygosil). Each eluate was examined using MALDI-TOF (UltraFlex Bruker Daltonics) for cyclotide-like masses (2.5−4.0 kDa). Fractions eluted with 50, 70, and 100% MeCN/ddH_2O showed cyclotide-like masses. These fractions were pooled for each plant tissue and freeze-dried (referred henceforth as "crude extracts") for further RP-HPLC purification.

Purification and MALDI-TOF Analysis. HPLC (Shimadzu Prominence) purification of crude extracts was performed using a semi-preparative Phenomenex Proteo C_{18} column (250 mm × 10 mm, 10 μm, 110 Å) at a flow rate of 3 mL/min with a fraction collector. A linear gradient of 1% min^{-1} of 0−95% buffer B (100% acetonitrile and 0.1% trifluoroacetic acid) was applied, and the eluents were monitored at 220, 254, and 280 nm. Early and late-eluting peaks were separated into five fractions (A−E), manually collected for each plant tissue and lyophilized. Each of the five fractions, per tissue, was subjected to MALDI-TOF-MS (UltraFlex Bruker Daltonics) analysis with a 50 Hz pulsed nitrogen laser (λ = 337 nm) in the positive-ion reflectron mode. The samples were prepared in a 1:1 mixture of H_2O−MeCN with 0.1% TFA by mixing an equal amount of peptide fractions (0.5 μL) with α-cyano-4-hydroxycinnamic acid/2,5-dihydroxy benzoic acid matrix (Sigma) and spotted on a stainless

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

steel plate and air-dried. HPLC fraction D from the crude extracts of each of the five tissues showed maximum cyclotide-like masses (2.5−4.0 kDa), both in terms of number and abundance, and fewer small molecule interferences. The total protein content in the lyophilized fraction D (CRF) was determined using the Bradford protein assay based on a standard protocol. Stock solutions of CRF were prepared by dissolving the lyophilized CRF powder in 70% MeCN (~5 mg/mL) and stored at −20 °C until use. For the assays, dilutions were prepared from the stock solutions to obtain three concentrations of CRF, that is, 20, 80, and 200 μg/mL.

Combined Omics Approach for Cyclotide Identification. Cyclotide precursor and mature sequences were identified using the previously reported transcriptome assembly[25] and their monoisotopic masses were calculated (Table S1, Supporting Information). These calculated monoisotopic masses were then compared to the observed monoisotopic masses from MALDI-TOF spectra of CRFs from five tissues (pods, stems, leaves, flowers, and roots).

Worm Strains and Maintenance. *C. elegans* strains N2, CL4176, CL2006, and CL2355 were a kind gift from Dr. Ashwini Godbole (The University of Trans-Disciplinary Health Sciences and Technology). The transgenic strains expressing human Aβ used in the current study are summarized in Table 1. Wild-type *C. elegans* strain N2 (Bristol) was grown and maintained on a standard NGM seeded with *Escherichia coli* (*E. coli*) strain OP50 as a food source and maintained at 20 °C. The transgenic CL4176 (smg-1 (cc546); myo-3::Aβ_{1-42}::3′-UTR (long)), CL2006 (unc-54::Aβ_{1-42} (wild type); dimer Aβ_{1-42} or Met[34] Cys Aβ_{1-42}), and CL2355 (smg-1 (cc546); snb-1::Aβ_{1-42}::3′-UTR (long) + mtl-2::GFP) strains were grown and maintained on standard NGM seeded with *E. coli* strain OP50 and maintained at 15−16 °C. To obtain age-synchronized worms, the reproductive worms (3 days of age) were transferred to the fresh NGM plates and allowed to lay eggs for 4−6 h. Worms hatched from these eggs were used for the study.

Paralysis Assay. The paralysis assays was performed as illustrated in Figure 3a. Briefly, age-synchronized transgenic CL4176 strains were maintained at 16 °C on NGM plates (60 mm × 10 mm culture plates, ~20 eggs/plate) spotted with an OP50-containing vehicle, that is, 100 mM ammonium bicarbonate buffer (negative control) and treated with or without CRF from different tissues (concentrations: 20, 80, and 200 μg/mL). Vitamin C (0.1 μg/mL) was used as a positive control. Transgenic expression of Aβ in body wall muscles in CL4176 strains can be induced by a temperature upshift from 16 to 23 °C (starting at 36 h after egg laying). Paralysis was scored every 8 h, starting from the 20th h until ~80% of the worms in the negative control experiment were paralyzed. The worm was considered paralyzed if it did not move or only head movement was observed upon prodding with a platinum loop. The paralysis time course was plotted.

Chemotaxis Assay. The chemotaxis assay was performed as illustrated in Figure 4a. Briefly, synchronized transgenic CL2355 strains were maintained at 16 °C on NGM plates (90 mm × 15 mm culture plates). They were spotted with an OP50-containing vehicle (100 mM ammonium bicarbonate buffer) and with or without CRF from different tissues (concentrations: 20, 80, and 200 μg/mL). Vitamin C (0.1 μg/mL) was used as a positive control. For the assay, eggs were maintained at 16 °C for 36 h (~20−60 eggs/plate), and then temperature was upshifted to 23−25 °C to induce expression of neuronal Aβ and the eggs were incubated for another 36 h. Worms were then collected, washed with M9 buffer thrice, and assayed on 90 mm plates containing 1.9% agar, 1 mM $CaCl_2$, 1 mM $MgSO_4$, and 25 mM phosphate saline buffer. Worms were placed at the center, and 2 μL of 0.1% benzaldehyde in 100% ethanol and 1 μL of sodium azide were added at positions A and D as attractant odorants (Figure 4a). 100% ethanol (2 μL) and sodium azide (1 μL) were added at positions B and C as control odorants. Assay plates were incubated at 23−25 °C for 2 h and CI was calculated using the formula: CI = [(number of worms at the attractant position − numbers of worms at the control position)/total number of worms scored)].

Intracellular ROS Assay. ROS levels were analyzed in whole live nematodes using a cell-permeable molecular probe, 2′,7′-dichlorodihydrofluorescein diacetate (H_2DCF−DA). Inside the cell, deacetylated by the intracellular esterases, H_2DCF−DA is retained whereas upon oxidation by ROS, non-fluorescent H_2DCF−DA is converted into highly fluorescent 2′,7′-DCF. The intensity of DCF fluorescence correlates with the levels of intracellular ROS. For detection of ROS production, age-synchronized *C. elegans* (CL4176; 100 eggs per plate) worms were treated with or without CRF from different tissues (concentration: 200 μg/mL) for 36 h at 16 °C followed by a shift to 23 °C for 36 h. The worms were collected and the *E. coli* (OP50) bacteria were removed by washing thrice with phosphate-buffered saline with Tween-20 (PBS-T). A 200 μL volume of the suspension was pipetted into a 96-well plate. Fresh 100 μM H_2DCF−DA solution was prepared in PBS-T from 50 mM stock solution. A volume of 50 μL of 100 μM H_2DCF−DA was then added to each well. Immediately after the addition, the basal fluorescence of each well was measured using a microplate reader at excitation/emission wavelengths of 485 and 538 nm. The plates were read at 37 °C every 30 min for 2.5 h. The time-dependent increase in the fluorescence was linear in this time period. We therefore chose to use the fluorescence intensity at 180 min for the quantification. Vitamin C (0.1 μg/mL) was used as a positive control. To further examine the changes in ROS, the worms from the plate were observed at 10× magnification using a fluorescence microscope (DM5000B, Leica, Germany) with an FITC excitation and emission filter setup. The images were acquired using a CCD digital camera (Hamamatsu ORCA-ER).

Fluorescence Staining of β-Amyloid. Age-synchronized CL2006 strains were plated on NGM plates (60 mm × 10 mm culture plates, ~30 eggs/plate) spotted with OP50 at 20 °C for 72 h. Then, the worms were transferred on NGM plates containing a vehicle and treated with or without CRF from different tissues (concentration: 200 μg/mL) for 48 h. Vitamin C (0.1 μg/mL) was used as a positive control. The worms were then fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) at pH 7.4 at 4 °C for 24 h. The worms were then permeabilized in 1% Triton X-100, 5% fresh β-mercaptoethanol, and 125 mM Tris, at pH 7.4, and incubated at 37 °C for another 24 h. The worms were washed 2−3 times with PBS-T and then stained with 0.125% ThS (Sigma) in 50% ethanol for 2 min and then destained sequentially in ethanol (50, 75, and 90%). The worms were mounted on coverslips covered with 80% glycerol and observed using a fluorescence microscope (DM5000B, Leica, Germany) to detect the presence of amyloid plaques in the head region of individual worms. The images were acquired at 40× magnification using a CCD digital camera (Hamamatsu ORCA-ER).

Statistical Analysis. Statistical analyses were carried out using GraphPad Prism 7 software package (GraphPad Software, La Jolla, CA, USA). Data were tested for normality and homoscedasticity, whenever required, using the D'Agostino−Pearson normality test and Bartlett's test, respectively. Paralysis, chemotaxis, and ROS assay data were subjected to one-way ANOVA followed by Tukey's honestly significant difference procedure when comparing untreated and CRF-treated groups. Differences between the positive control and CRF-treated groups were analyzed for statistical significance using the unpaired two-tailed Student's *t* test. Further details of statistical analyses of all the assays are provided in the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.1c00033.

Expanded MALDI-TOF spectra of HPLC-purified fractions A, B, C, D, and E from crude extracts of different tissues (pods, stems, leaves, flowers and roots) of *C. ternatea* (Figure S1), stacked plot of full MALDI-TOF spectra of CRF from different tissues (pods, stems, leaves, flowers, and roots) of *C. ternatea* (Figure S2), cyclotide transcript sequences determined in the previous study using the transcriptome assembly of *C.*

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

ternatea (Table S1), descriptive statistics of paralysis assay in transgenic *C. elegans* CL4176 (Table S2), descriptive statistics of chemotaxis assay in transgenic *C. elegans* CL2355 (Table S3), and descriptive statistics of intracellular ROS assay in transgenic *C. elegans* CL2006 (Table S4) (PDF)

AUTHOR INFORMATION

Corresponding Author

Radhika Venkatesan − *National Centre for Biological Sciences (TIFR), Bangalore, Karnataka 560065, India; Department of Biological Sciences, Indian Institute of Science Education and Research, Mohanpur, West Bengal 741246, India;* orcid.org/0000-0002-8112-9505; Email: rv@iiserkol.ac.in

Authors

Neha V. Kalmankar − *National Centre for Biological Sciences (TIFR), Bangalore, Karnataka 560065, India; The University of Trans-Disciplinary Health Sciences and Technology (TDU), Bangalore, Karnataka 560064, India;* orcid.org/0000-0003-4834-3566

Hrudya Hari − *National Centre for Biological Sciences (TIFR), Bangalore, Karnataka 560065, India*

Ramanathan Sowdhamini − *National Centre for Biological Sciences (TIFR), Bangalore, Karnataka 560065, India;* orcid.org/0000-0002-6642-2367

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.1c00033

Author Contributions

N.K. and R.V. conceived and designed the experiments, N.K. and H.H. performed the experiments, N.K., R.S., and R.V. conducted data analysis, and N.K., R.V., and R.S. contributed to manuscript preparation.

Notes

All animal experiments performed in the manuscript were conducted in compliance with institutional guidelines.
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Prof. P. Balaram (National Centre for Biological Sciences, Bangalore, India) for his valuable suggestions throughout the project and help with the initial version of the manuscript. We thank Dr. Ashwini Godbole (The University of Trans-Disciplinary Health Sciences and Technology, Bangalore, India) for providing *C. elegans* strains and Dr. Varsha Singh (Indian Institute of Science, Bangalore, India) for her inputs on the manuscript. We also thank the *C. elegans* facility at NCBS (TIFR). N.K. was supported by the Tata Education and Development Trust. Mass spectra were obtained at the Proteomics Facility supported by the Department of Biotechnology at the Indian Institute of Science, Bangalore. We thank NCBS (TIFR) and acknowledge JC Bose fellowship provided to R.S. (SB/S2/JC-071/2015). Funding to R.V. by the DST-SERB (Early Career Award, Ramanujan Fellowship), the Max Planck Society (Partner group program), and the Department of Biotechnology is gratefully acknowledged.

ABBREVIATIONS

MeCN, acetonitrile; CRF, cyclotide-rich fraction; CI, chemotaxis index; DCF, 2′,7′-dichlorofluorescein; H_2DCF-DA, 2′,7′-dichlorodihydrofluorescein diacetate; ThS, thioflavin-S; NGM, nematode growth media; CCA, α-cyano-4-hydroxycinnamic acid; DHB, 2,5-dihydroxy benzoic acid; UTR, untranslated region; $CaCl_2$, calcium chloride; $MgSO_4$, magnesium sulfate; PBS-T, phosphate-buffered saline with Tween-20; FITC, fluorescein isothiocyanate; CCD, charge-coupled device; HSD, honestly significant difference; ANOVA, analysis of variance.

REFERENCES

(1) Hardy, J.; Selkoe, D. J. The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics. *Science* **2002**, *297*, 353−356.

(2) Jokar, S.; Khazaei, S.; Behnammanesh, H.; Shamloo, A.; Erfani, M.; Beiki, D.; Bavi, O. Recent Advances in the Design and Applications of Amyloid-β Peptide Aggregation Inhibitors for Alzheimer's Disease Therapy. *Biophys. Rev.* **2019**, *11*, 901−925.

(3) Andrade, S.; Ramalho, M. J.; Loureiro, J. A.; Pereira, M. d. C. Natural Compounds for Alzheimer's Disease Therapy: A Systematic Review of Preclinical and Clinical Studies. *Int. J. Mol. Sci.* **2019**, *20*, 2313.

(4) Wu, Y.; Wu, Z.; Butko, P.; Christen, Y.; Lambert, M. P.; Klein, W. L.; Link, C. D.; Luo, Y. Amyloid-β-Induced Pathological Behaviors Are Suppressed by Ginkgo Biloba Extract EGB 761 and Ginkgolides in Transgenic *Caenorhabditis elegans*. *J. Neurosci.* **2006**, *26*, 13102−13113.

(5) Li, J.; Cui, X.; Ma, X.; Wang, Z. RBTI Reduced β-Amyloid-Induced Toxicity by Promoting Autophagy-Lysosomal Degradation via DAF-16 in *Caenorhabditis elegans*. *Exp. Gerontol.* **2017**, *89*, 78−86.

(6) Martorell, P.; Bataller, E.; Llopis, S.; Gonzalez, N.; Álvarez, B.; Montón, F.; Ortiz, P.; Ramón, D.; Genovés, S. A Cocoa Peptide Protects *Caenorhabditis elegans* from Oxidative Stress and β-Amyloid Peptide Toxicity. *PLoS One* **2013**, *8*, No. e63283.

(7) Wu, Y.; Luo, Y. Transgenic *C. elegans* as a Model in Alzheimer's Research. *Curr. Alzheimer Res.* **2005**, *2*, 37−45.

(8) Calahorro, F.; Ruiz-Rubio, M. *Caenorhabditis elegans* as an Experimental Tool for the Study of Complex Neurological Diseases: Parkinson's Disease, Alzheimer's Disease and Autism Spectrum Disorder. *Invertebr. Neurosci.* **2011**, *11*, 73−83.

(9) Ewald, C. Y.; Li, C. *Caenorhabditis elegans* as a Model Organism to Study APP Function. *Exp. Brain Res.* **2012**, *217*, 397−411.

(10) Mukherjee, P. K.; Kumar, V.; Kumar, N. S.; Heinrich, M. The Ayurvedic Medicine *Clitoria ternatea*-From Traditional Use to Scientific Assessment. *J. Ethnopharmacol.* **2008**, *120*, 291−301.

(11) Jain, N. N.; Ohal, C. C.; Shroff, S. K.; Bhutada, R. H.; Somani, R. S.; Kasture, V. S.; Kasture, S. B. *Clitoria ternatea* and the CNS. *Pharmacol. Biochem. Behav.* **2003**, *75*, 529−536.

(12) Kavitha, R. Antidiabetic and Enzymatic Antioxidant Potential from Ethanolic Extracts of Leaf and Fruit of *Trichosanthes dioica* and Leaf of *Clitoria ternatea* on Diabetic Rats Induced by Streptozotocin. *Asian J. Pharm. Clin. Res.* **2018**, *11*, 233−239.

(13) Rai, K. S.; Murthy, K. D.; Karanth, K. S.; Nalini, K.; Rao, M. S.; Srinivasan, K. K. *Clitoria ternatea* Root Extract Enhances Acetylcholine Content in Rat Hippocampus. *Fitoterapia* **2002**, *73*, 685−689.

(14) Daisy, P.; Rajathi, M. Hypoglycemic Effects of *Clitoria ternatea* Linn. (Fabaceae) in Alloxan-Induced Diabetes in Rats. *Trop. J. Pharm. Res.* **2009**, *8*, 393−398.

(15) Poth, A. G.; Colgrave, M. L.; Philip, R.; Kerenga, B.; Daly, N. L.; Anderson, M. A.; Craik, D. J. Discovery of Cyclotides in the Fabaceae Plant Family Provides New Insights into the Cyclization, Evolution, and Distribution of Circular Proteins. *ACS Chem. Biol.* **2011**, *6*, 345−355.

(16) Saether, O.; Craik, D. J.; Campbell, I. D.; Sletten, K.; Juul, J.; Norman, D. G. Elucidation of the Primary and Three-Dimensional

Structure of the Uterotonic Polypeptide Kalata B1. *Biochemistry* **1995**, *34*, 4147−4158.
(17) Craik, D. J.; Conibear, A. C. The Chemistry of Cyclotides. *J. Org. Chem.* **2011**, *76*, 4805−4817.
(18) Colgrave, M. L.; Craik, D. J. Thermal, Chemical, and Enzymatic Stability of the Cyclotide Kalata B1: The Importance of the Cyclic Cystine Knot. *Biochemistry* **2004**, *43*, 5965−5975.
(19) Danho, W.; Swistok, J.; Khan, W.; Chu, X. J.; Cheung, A.; Fry, D.; Sun, H.; Kurylko, G.; Rumennik, L.; Cefalu, J.; Cefalu, G.; Nunn, P. Opportunities and Challenges of Developing Peptide Drugs in the Pharmaceutical Industry. *Adv. Exp. Med. Biol.* **2009**, *611*, 467−469.
(20) Craik, D. J.; Fairlie, D. P.; Liras, S.; Price, D. The Future of Peptide-Based Drugs. *Chem. Biol. Drug Des.* **2013**, *81*, 136−147.
(21) Zhang, Z.; Ma, H.; Wang, X.; Zhao, Z.; Zhang, Y.; Zhao, B.; Guo, Y.; Xu, L. A Tetrapeptide from Maize Protects a Transgenic *Caenorhabditis elegans* Aβ1-42 Model from Aβ-Induced Toxicity. *RSC Adv.* **2016**, *6*, 56851−56858.
(22) Zhang, X. G.; Wang, X.; Zhou, T. T.; Wu, X. F.; Peng, Y.; Zhang, W. Q.; Li, S.; Zhao, J. Scorpion Venom Heat-Resistant Peptide Protects Transgenic *Caenorhabditis elegans* from Beta-Amyloid Toxicity. *Front. Pharmacol.* **2016**, *7*, 227.
(23) Xu, J.; Yuan, Y.; Zhang, R.; Song, Y.; Sui, T.; Wang, J.; Wang, C.; Chen, Y.; Guan, S.; Wang, L. A Deuterohemin Peptide Protects a Transgenic *Caenorhabditis elegans* Model of Alzheimer's Disease by Inhibiting Aβ1−42 Aggregation. *Bioorg. Chem.* **2019**, *82*, 332−339.
(24) Manzanares, P.; Martínez, R.; Garrigues, S.; Genovés, S.; Ramón, D.; Marcos, J. F.; Martorell, P. Tryptophan-Containing Dual Neuroprotective Peptides: Prolyl Endopeptidase Inhibition and *Caenorhabditis elegans* Protection from β-Amyloid Peptide Toxicity. *Int. J. Mol. Sci.* **2018**, *19*, 1−18.
(25) Kalmankar, N. V.; Venkatesan, R.; Balaram, P.; Sowdhamini, R. Transcriptomic Profiling of the Medicinal Plant *Clitoria ternatea*: Identification of Potential Genes in Cyclotide Biosynthesis. *Sci. Rep.* **2020**, *10*, No. 12658.
(26) Lublin, A. L.; Link, C. D. Alzheimer's Disease Drug Discovery: In Vivo Screening Using *Caenorhabditis elegans* as a Model for β-Amyloid Peptide-Induced Toxicity. *Drug Discov. Today Technol.* **2013**, *10*, e115−e119.
(27) Oguis, G. K.; Gilding, E. K.; Jackson, M. A.; Craik, D. J. Butterfly Pea (Clitoria Ternatea), a Cyclotide-Bearing Plant with Applications in Agriculture and Medicine. *Front. Plant Sci.* **2019**, *10*, 645.
(28) Serra, A.; Hemu, X.; Nguyen, G. K. T.; Nguyen, N. T. K.; Sze, S. K.; Tam, J. P. A High-Throughput Peptidomic Strategy to Decipher the Molecular Diversity of Cyclic Cysteine-Rich Peptides. *Sci. Rep.* **2016**, *6*, No. 23005.
(29) Gilding, E. K.; Jackson, M. A.; Poth, A. G.; Henriques, S. T.; Prentis, P. J.; Mahatmanto, T.; Craik, D. J. Gene coevolution and regulation lock cyclic plant defence peptides to their targets. *New Phytol.* **2016**, *210*, 717−730.
(30) Nguyen, G. K. T.; Zhang, S.; Nguyen, N. T. K.; Nguyen, P. Q. T.; Chiu, M. S.; Hardjojo, A.; Tam, J. P. Discovery and characterization of novel cyclotides originated from chimeric precursors consisting of albumin-1 chain a and cyclotide domains in the fabaceae family. *J. Biol. Chem.* **2011**, *286*, 24275−24287.
(31) Poth, A. G.; Colgrave, M. L.; Lyons, R. E.; Dalya, N. L.; Craik, D. J. Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 10127−10132.
(32) Nguyen, K. N. T.; Nguyen, G. K. T.; Nguyen, P. Q. T.; Ang, K. H.; Dedon, P. C.; Tam, J. P. Immunostimulating and Gram-negative-specific antibacterial cyclotides from the butterfly pea (Clitoria ternatea). *FEBS J.* **2016**, *283*, 2067−2090.
(33) Gerlach, S. L.; Göransson, U.; Kaas, Q.; Craik, D. J.; Mondal, D.; Gruber, C. W. A Systematic Approach to Document Cyclotide Distribution in Plant Species from Genomic, Transcriptomic, and Peptidomic Analysis. *Biopolymers* **2013**, *100*, 433−437.
(34) Monacelli, F.; Acquarone, E.; Giannotti, C.; Borghi, R.; Nencioni, A. Vitamin C, Aging and Alzheimer's Disease. *Nutrients* **2017**, *9*, 670.
(35) Cheng, F.; Cappai, R.; Ciccotosto, G. D.; Svensson, G.; Multhaup, G.; Fransson, L. Å.; Mani, K. Suppression of Amyloid β A11 Antibody Immunoreactivity by Vitamin C: Possible Role of Heparan Sulfate Oligosaccharides Derived from Glypican-1 by Ascorbate-Induced, Nitric Oxide (NO)-Catalyzed Degradation. *J. Biol. Chem.* **2011**, *286*, 27559−27572.
(36) Kook, S. Y.; Lee, K. M.; Kim, Y.; Cha, M. Y.; Kang, S.; Baik, S. H.; Lee, H.; Park, R.; Mook-Jung, I. High-Dose of Vitamin C Supplementation Reduces Amyloid Plaque Burden and Ameliorates Pathological Changes in the Brain of 5XFAD Mice. *Cell Death Dis.* **2014**, *5*, No. e1083.
(37) Thell, K.; Hellinger, R.; Sahin, E.; Michenthaler, P.; Gold-Binder, M.; Haider, T.; Kuttke, M.; Liutkevičiute, Z.; Göransson, U.; Gründemann, C.; Schabbauer, G.; Gruber, C. W. Oral Activity of a Nature-Derived Cyclic Peptide for the Treatment of Multiple Sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 3960−3965.
(38) Hellinger, R.; Koehbach, J.; Puigpinós, A.; Clark, R. J.; Tarragó, T.; Giralt, E.; Gruber, C. W. Inhibition of Human Prolyl Oligopeptidase Activity by the Cyclotide Psysol 2 Isolated from *Psychotria solitudinum*. *J. Nat. Prod.* **2015**, *78*, 1073−1082.
(39) Keowkase, R.; Shoomarom, N.; Bunargin, W.; Sitthithaworn, W.; Weerapreeyakul, N. Sesamin and Sesamolin Reduce Amyloid-β Toxicity in a Transgenic *Caenorhabditis elegans*. *Biomed. Pharmacother.* **2018**, *107*, 656−664.
(40) Narayani, M.; Babu, R.; Chadha, A.; Srivastava, S. Production of Bioactive Cyclotides: A Comprehensive Overview. *Phytochem. Rev.* **2020**, *19*, 787−825.
(41) Gella, A.; Durany, N. Oxidative Stress in Alzheimer Disease. *Cell Adh. Migr.* **2009**, *3*, 88−93.
(42) Butterfield, D. A. Amyloid β-Peptide (1-42)-Induced Oxidative Stress and Neurotoxicity: Implications for Neurodegeneration in Alzheimer's Disease Brain. A Review. *Free Radic. Res.* **2002**, *36*, 1307−1313.
(43) Drake, J.; Link, C. D.; Butterfield, D. A. Oxidative Stress Precedes Fibrillar Deposition of Alzheimer's Disease Amyloid β-Peptide (1-42) in a Transgenic *Caenorhabditis elegans* Model. *Neurobiol. Aging* **2003**, *24*, 415−420.
(44) Mushtaq, Z.; Iqbal, T.; Ahmed, N.; Jamil, A. Antioxidants from Selected Indigenous Plants Possessing Cyclotides. *Oxid. Commun.* **2017**, *40*, 102−119.
(45) Kumar, D.; Dhobi, M. Screening Antianxiety and Antioxidant Profile of Stems and Leaves of Blue Variety of Clitoria Ternatea L. *Indian J. Pharm. Sci.* **2017**, *79*, 1022−1025.
(46) Kamkaen, N.; Wilkinson, J. M. The Antioxidant Activity of Clitoria Ternatea Flower Petal Extracts and Eye Gel. *Phytother. Res.* **2009**, *23*, 1624−1625.
(47) Matsuura, H. N.; Poth, A. G.; Yendo, A. C. A.; Fett-Neto, A. G.; Craik, D. J. Isolation and Characterization of Cyclotides from Brazilian Psychotria: Significance in Plant Defense and Co-Occurrence with Antioxidant Alkaloids. *J. Nat. Prod.* **2016**, *79*, 3006−3013.
(48) Adewole, K. E.; Attah, A. F.; Sonibare, M. A.; Moody, J. O.; Adebayo, J. O. Identification of Antioxidant Cysteine-Stabilised Peptides of *Morinda lucida* Benth. Leaf. *Indian J. Pharm. Sci.* **2018**, *80*, 99−107.
(49) Li, Y.; Yu, J. Research Progress in Structure-Activity Relationship of Bioactive Peptides. *J. Med. Food* **2015**, *18*, 147−156.

https://doi.org/10.1021/acs.jmedchem.1c00033
J. Med. Chem. 2021, 64, 7422−7433

Manuscript 6

Effects of cyclotide on conformational dynamics and destabilization of β-amyloid fibrils through molecular dynamics simulations

Neha V. Kalmankar, Radhika Venkatesan and Ramanathan Sowdhamini*

5.3 ABSTRACT

Aggregation of β-amyloid (Aβ) peptide is one of the hallmarks of Alzheimer's disease (AD) which results in chronic and progressive neurodegeneration of the brain. A recent study by our group have shown the ability of cyclic disulfide-rich peptides to inhibit the aggregation of Aβ peptides and reduce oxidative stress caused by reactive oxygen species using *in vivo* models of transgenic *Caenorhabditis elegans*. In the present study, through extensive computational docking and multi-ns molecular dynamics (MD) simulation, we evaluated if cyclotides can stably bind to Aβ molecules and/or destabilize the Aβ fibril by preventing conformational changes from α-helical to β-sheet rich structures. We demonstrate that cyclotides bind effectively and stably to different forms of Aβ structures via hydrogen bonding and hydrophobic interactions. A detailed MD simulation analyses further revealed that cyclotides form hydrogen bonds with the toxic amyloid assemblies thereby weakening the inter-strand hydrogen bonds between the Aβ peptide. The current study provides evidence for the first time that cyclotides have anti-amyloid aggregation properties and could be a source of novel pharmacophore scaffold against neurodegenerative diseases.

5.4 OVERVIEW

Alzheimer's disease (AD) is the most common cause of senile dementia and primarily characterized by extracellular plaques of beta-amyloid peptide (Aβ) deposits in the brain. However, the molecular mechanisms involved in cause and progression of AD remains to be fully understood. In all these years, inhibiting Aβ aggregation has become the starting point for therapy against AD as the neurotoxicity is often associated with the pathogenesis of Aβ monomers aggregating into oligomers and fibrils. Peptides are a great source for antioxidant properties which can be used therapeutically for the prevention of age-related diseases. Additionally, peptides advantageous over small molecules due to their non-immunogenicity, better bioavailability, low-toxicity and convenient chemical synthesis. In particular, ribosomal synthesized and post-translational modified peptides (RiPP) contain

modifications such as cyclization, which improves their ADME properties (absorption, distribution, metabolism, and excretion). Such modifications can improve the oral bioavailability and ability to pass the blood-brain-barrier, which still remains as a major challenge of peptide-based drug development.

Few years ago, synthetically created cyclotides, analogous to the *Oldenlandia affinis* plant-derived peptide, could be orally administered in animal models to suppress multiple sclerosis (MS), an autoimmune disease affecting the central nervous system. More recently, cyclotides' anti-neurodegenerative properties were studied in *Psychotria solitudinum*, where they acted as inhibitors of the human prolyl oligopeptidase, a promising target for the treatment of cognitive deficits in several psychiatric and neurodegenerative diseases. Our group has recently demonstrated the *in vivo* anti-Aβ effects of cyclotides using three transgenic *Caenorhabditis elegans* models that exhibits pathological behaviours associated with Aβ. To understand the molecular and atomic level interactions between Aβ and cyclotides, we performed in silico studies through molecular docking analyses and molecular dynamics (MD) simulations. We evaluated if cyclotides can stably bind to Aβ molecules and/or destabilize the Aβ fibril by preventing conformational changes from α-helical to β-sheet rich structures. Overall, our findings from the current study provides novel insights on the potential of cyclotides as Aβ inhibitors, specifically elucidating the molecular mechanisms involved in destabilization of Aβ protofibrils. The results provide evidence for the first time that cyclotides have anti-amyloid aggregation properties and can act as an important source of drug-like molecules.

5.5 MATERIALS & METHODS

5.5.1 Protein - peptide docking

In order to better understand and visualize the mode of interactions between different forms of Aβ fibrillar species and cyclotide, we performed molecular docking experiments using FRODOCK2.0 (Ramírez-Aportela et al., 2016). The five different three-dimensional (3D) NMR solution structures of Aβ used in this study are: a) PDB ID: 1IYT, $A\beta_{1-42}$ monomer (Crescenzi et al., 2002), b) PDB ID: 2BEG, U-shaped pentamer $A\beta_{17-42}$ (Lührs et al., 2005), c) PDB ID: 2MXU, S-shaped model $A\beta_{11-42}$ (Xiao et al., 2015), d) PDB ID: 2NAO, disease relevant $A\beta_{1-42}$ fibrils (Wälti et al., 2016) and e) PDB ID: 2M4J, $A\beta_{1-40}$ isolated from brain of AD patient (Lu et al., 2013). We used the three-dimensional NMR solution structure of Cter-M cyclotide from *C. ternatea* (PDB ID: 2LAM) to dock against each of the five Aβ fibrillar species. For all the generated docked poses for each of the cyclotide-Aβ complexes (5 complexes), PPCheck was used to calculate binding energies

and to predict best native-like docking pose (Sukhwal and Sowdhamini, 2015).

5.5.2 Molecular dynamics simulation

To confirm stable binding of each Aβ-cyclotide complex, MD simulation without constraints was performed using the Desmond package of Maestro Schrodinger (Bowers et al., 2006). The TIP4P solvent model was used with an orthorhombic box shape and buffer distance of 10 Å. The OPLS3e force field was used for building of the system. The system was minimized using steepest descent for 2000 steps until a gradient threshold of 50 kcal mol−1 Å−1 was reached. The system was neutralized by adding Na^+ ions and 0.15 M salt (NaCl). The output from the system builder was used for MD simulations. The simulation time was set at 00 ns in the NPT ensemble class. The RESPA integrator was used with a time step of 2.0 fs (Humphreys et al., 1994). The temperature and pressure were set at 300K and 1 bar, using the Nose-Hoover chain (Martyna et al., 1992) and the Martyna-Tobias-Klein method (Martyna et al., 1994) respectively.

5.5.3 Conformational analysis of protein - peptide complex

For stability and conformational analyses, the entire range of simulation time was considered. RMSD, RMSF, radius of gyration and number of hydrogen bonds was calculated for MD runs of all the five Aβ (protein) and cyclotide (peptide) complexes. A set of control MD runs were performed on each of the unbound Aβ structure. Hydrogen bonds were identified using the Simulation Event Analysis module implemented in the Desmond package. Protein RMSF shows the fluctuation of Aβ residues and peptide RMSF shows fluctuation of cyclotide residues. Protein-peptide interactions were also monitored throughout the simulation time.

5.6 RESULTS

C. ternatea has a history of being used in traditional medicine with anecdotal evidences coming from several countries such as India, Philippines, Cuba, China, etc for its use ranging from expediting childbirth to being used as a brain tonic for memory improvement. Shankhpushpi or Aparajita, as it is locally known in India, is used in Ayurvedic medicine as a memory and cognition enhancer, potent antioxidant, anxiolytic agent, etc. It is also the only known cyclotide-producing species from the Fabaceae family. However, up till now there have been no reports of pharmacological activities of cyclotides derived from this plant as potential drug leads to treat neurodegenerative disorders such as Alzheimer's disease, Parkinson's disease, Huntington's disease, etc. In order to obtain a structural understanding of how the presence of cyclotides would reduce the aggregation of Aβ

peptides, we modelled the interactions and performed MD simulations.

5.6.1 Molecular docking of cyclotide and Aβ structures

A general schema for docking and MD simulation of the multi-protein complexes employed in the present study is described in **Figure 5.1**. Two most common isoforms of Aβ are $A\beta_{1-40}$ and $A\beta_{1-42}$ depending on the position of cleavage on the amyloid precursor protein (APP). Several polymorphs of amyloid structures exist and plenty of high-resolution structures are available for $A\beta_{1-40}$ fragments. However, high resolution three-dimensional structures of the disease relevant form of Aβ are relatively few. Hence, we used the available NMR derived structure of $A\beta_{1–42}$ hexamer (PDB ID: 2NAO) as a representative amyloid structure to understand the possible interaction of Aβ fibril and cyclotide. However, there is high complexity and diversity in the amyloid folds and in order to evaluate the ability of cyclotide to bind to different polymorphs, we used a variety of Aβ quaternary structures. These include: PDB ID 1IYT ($A\beta_{1-42}$ monomer), PDB ID 2BEG, (U-shaped pentamer $A\beta_{17-42}$), PDB ID 2MXU (S-shaped model $A\beta_{11-42}$) and PDB ID 2M4J ($A\beta_{1-40}$ isolated from brain of AD patient) (**Figure 5.1A**) (Crescenzi et al., 2002; Lu et al., 2013; Lührs et al., 2005; Wälti et al., 2016; Xiao et al., 2015). Of these NMR models, one is a monomeric form of $A\beta_{1-42}$ (PDB ID: 1IYT), two are fragments of the fibrillary form (PDB IDs: 2BEG and 2MXU) and one is $A\beta_{1-40}$ truncated model (PDB ID: 2M4J). Only one NMR structure of chemically synthesized Cter-M cyclotide from *C. ternatea* (PDB ID 2LAM) is known till date and we have used it as a representative cyclotide structure for all the docking and simulation studies (**Figure 5.1B**) (Poth et al., 2011). To evaluate the molecular interactions and binding affinities, we carried out rigid-body docking studies using FRODOCK2.0 between cyclotide and different forms of Aβ peptide (**Figure 5.1C**) (Ramírez-Aportela et al., 2016). FRODOCK output comprising of top 10 poses for each of protein-peptide complex were evaluated using our in-house program PPCheck to predicting the best docking pose based on the total stabilizing energy and normalized energy per residue (Sukhwal and Sowdhamini, 2015). PPCheck quantifies the strength of protein-protein interaction using pseudo-energies as van der Waals, electrostatic and hydrogen bonds. For 1IYT-2LAM, 2BEG-2LAM, 2MXU-2LAM, 2NAO-2LAM and 2M4J-2LAM complexes, pose-2, 3, 7, 10 and 4 were the best-docked poses, respectively (Table S5.1). From the PPCheck analyses, the pose with lowest normalized energy per residue were selected to perform MD simulations (**Figure 5.1D**).

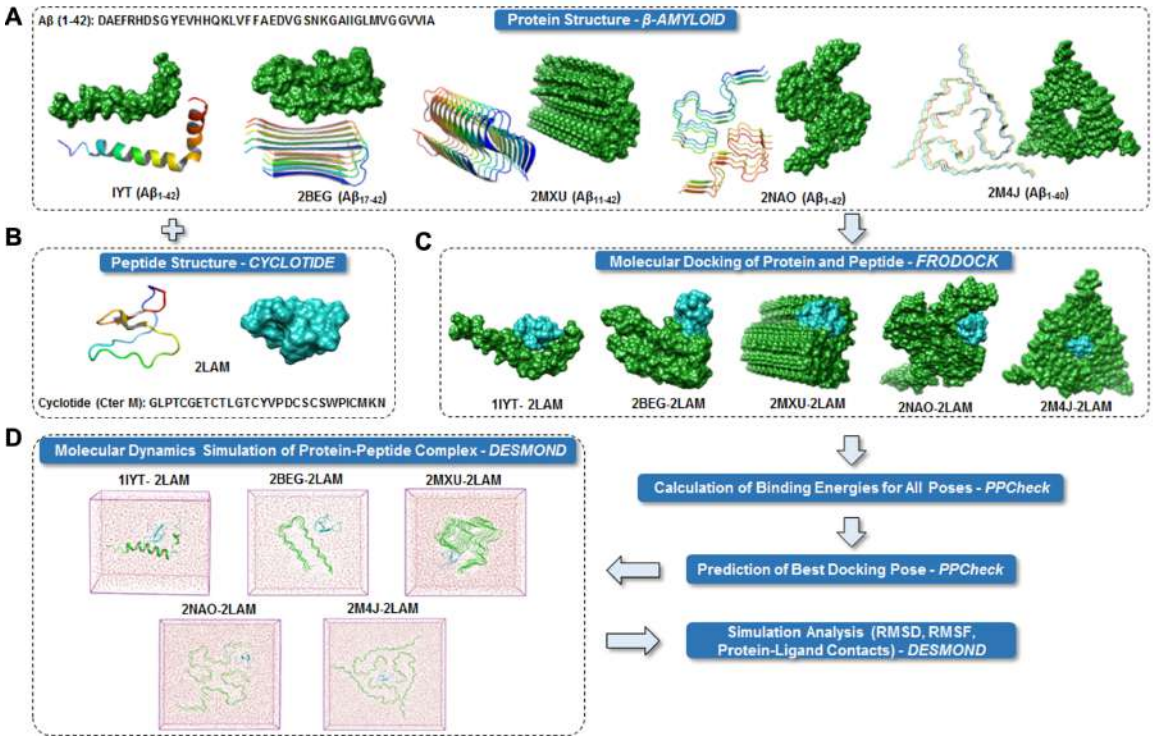


Figure 5.1: Schematic representation of the research protocol used in the present study. (A) Cartoon and surface representations of the five different Aβ structures. (B) Cartoon and surface representation of cyclotide Cter M structure. (C) Surface representations of the best binding mode of the five docked complexes of Aβ and cyclotides. Color scheme used for surface representation: green - Aβ and cyan - cyclotide. (D) The protein complexes are minimized and placed in a box of water. The systems are then set up for MD simulation.

5.6.2 Molecular dynamics simulation analyses of cyclotide - Aβ complexes

To assess the stability of the docked poses for each protein-peptide complex, molecular dynamics simulations were carried out using the Desmond package of Maestro Schrodinger (Bowers et al., 2006). The $A\beta_{1–42}$ fragment is the dominant Aβ species compared to other polymorphs in the amyloid plaques of AD. Therefore, we used the solid-state NMR derived hexameric structure (PDB ID: 2NAO) of $A\beta_{1–42}$ fibril as a representative for modeling long range MD simulations. We performed three independent 300 ns MD simulations for cyclotide (PDB ID: 2LAM) in complex with the $A\beta_{1–42}$ fibril (PDB ID: 2NAO). The averaged backbone root mean square deviations (RMSD), root mean square fluctuation (RMSF) and radius of gyration (Rg) of the $A\beta_{1-42}$ with and without cyclotides were calculated as a function of time. Between the bound and unbound forms of Aβ, there were greater fluctuations seen in the cyclotide-bound Aβ structure compared to unbound Aβ, based on backbone RMSD, RMSF and radius of gyration parameters (Figure S5.1A). The stability of the cyclotide Cter-M interaction with the disease relevant $A\beta_{1-42}$ fibril is highlighted in **Figure 5.2**. It is evident that upon binding to Cter-M, $A\beta_{1-42}$ fibril undergoes

large conformational changes and the secondary structures, especially extended β-sheet conformations, begin to disrupt soon after 50 ns (Supplementary Information Movie 5.1). This is in contrast to the unbound $A\beta_{1-42}$ fibril without the cyclotide molecules wherein extended β-sheet conformations are relatively more stable throughout the simulation (**Figure 5.2A & B**, Supplementary Information Movie 5.1 and 5.2).

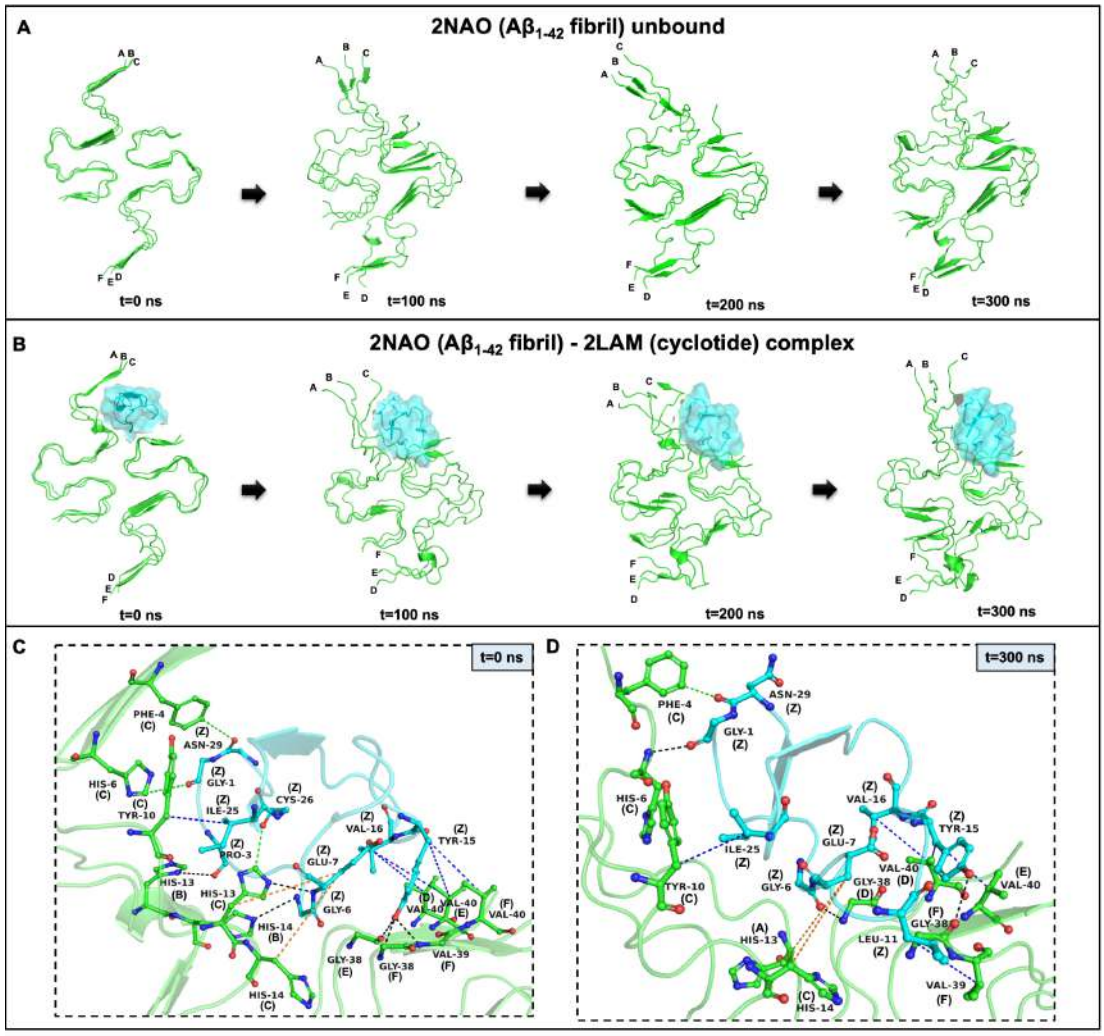


Figure 5.2: Representative example of MD simulations between the $A\beta_{1-42}$ fibril and cyclotide Cter-M. (A) Snapshots of MD simulation of unbound $A\beta_{1-42}$ fibril (PDB ID: 2NAO; green colored cartoon representation) at different time points along the simulation period. (B) Snapshots of MD simulation of $A\beta_{1-42}$ fibril (PDB ID: 2NAO; green colored cartoon representation) and Cter-M (PDB ID: 2LAM; cyan surface representation) complex at different time points along the simulation period. The cyclizing residues (Gly 1 and Asn 29) are shown as sticks to indicate the N- and C-termini. Capital letters, which denote the polypeptide chains of 2NAO, are placed at the 1st residue of each chain. (C, D) Molecular interactions between the $A\beta_{1-42}$ fibril (PDB ID: 2NAO; chains A-F; green cartoon representation) and cyclotide Cter-M (PDB ID: 2LAM; chain Z; cyan cartoon representation) at the beginning (0th ns snapshot) and end (300th ns snapshot) of the simulation period, respectively. Color scheme for interactions used: classical hydrogen

bonds in black, aromatic hydrogen bonds in green, hydrophobic interactions in blue, electrostatic interactions in orange, salt-bridges in pink and π-π interactions in red.

The timeline analysis of the secondary structural variations during the 300 ns MD simulations shows that extended β-sheet conformations (in yellow) were stable throughout the trajectory in the unbound $A\beta_{1-42}$ fibril. However, in the cyclotide-$A\beta_{1-42}$ bound trajectory, the binding pocket region undergoes significant structural changes soon after 50 ns and loses the extended β-sheet conformations, especially around residues 2-7 (chains A-C) and 38-41 (chains D-F) of $A\beta_{1-42}$ structure (Figure S5.2). Hydrogen bond analysis of the residues in the binding pocket over the 300 ns trajectory also substantiate this observation. Throughout the trajectory, the number of hydrogen bonds between the β-strands of $A\beta_{1-42}$ structure bound to cyclotide is reduced in comparison to the unbound form (Figure S5.1B). PPCheck and Maestro was utilized for identifying all the possible protein-protein interactions including classical hydrogen bonds, aromatic hydrogen bonds, electrostatic and hydrophobic interactions, salt bridges and π-π interactions. **Figure 5.2C & D** summarizes the molecular interactions between the $A\beta_{1-42}$ fibril and Cter-M at the beginning (0^{th} ns snapshot) and end (300^{th} ns snapshot) of the simulation period, respectively. At t=0 ns, the $A\beta_{1-42}$ fibril and Cter-M complex is mainly stabilized by nine hydrogen bonds, five hydrophobic interactions and two electrostatic interactions (**Figure 5.2C**; Table S5.2). The interaction interface involves polar and aromatic residues of Aβ *i.e.*, His13, His14 from chain B, Phe4, His6, Tyr10, His13 and His14 from chain C and aliphatic residues such as Val40 from chain D, Gly38 and Val40 from chain E, and Gly38, Val39 and Val40 from chain F. The residues of Cter-M involved in the interaction are Gly1, Pro3, Gly6, Gly7, Tyr15, Val16, Ile25, Cys26, and Asn29. At t=300 ns, the $A\beta_{1-42}$ fibril and Cter-M complex is mainly stabilized by four hydrogen bonds, four hydrophobic interactions and two electrostatic interactions (**Figure 5.2D**; Table S5.2). The interaction interface at the end of 300ns simulations involves similar residues of Aβ as mentioned earlier *i.e.*, His13 from chain A, Phe4, His6, Tyr10, His14 from chain C, and aliphatic residue such as Gly38 from chain D and F, Val39 from chain F and Val40 from chain D and E. Cter-M residues stabilizing the $A\beta_{1-42}$ are also comparable comprising of Gly1, Gly6, Gly7, Leu11, Tyr15, Val16, Ile25 and Asn29. The total stabilizing energetics at t=0 ns and 300 ns frames for the cyclotide - $A\beta_{1-42}$ fibril complex is detailed in **Table 5.1**. The intermolecular hydrogen bonds were also monitored throughout the simulation period as these are relative measures of binding affinity. An average number of 4 hydrogen bonds are present in the cyclotide - $A\beta_{1-42}$ fibril complex (Figure S5.1).

Table 5.1. The total stabilizing energy at t=0 ns and t=300 ns for the disease relevant $A\beta_{1-42}$ fibril (PDB ID: 2NAO) interacting with cyclotide Cter-M (PDB ID: 2LAM).

Total Stabilizing Energy	**0th ns frame**	**300th ns frame**
Hydrogen Bond Energy (kJ/mol)	-20.12	-25.54
Electrostatic Energy (kJ/mol)	-4.65	-2.59
Van der Waals Energy (kJ/mol)	-208.13	-184.12
Total Stabilizing Energy (kJ/mol)	-232.91	-212.25
Number of interface residues	75	75
Normalized Energy per residue (kJ/mol)	-3.11	-2.83
No. of Hydrophobic Interactions	5	5
No. of van der Waals Pairs	7768	7687
No. of Salt Bridges	0	0
No. of Potential Favourable Electrostatic Interactions	2	2

Cyclotide has the potential to stably bind and disrupt the ordered structures of several forms of amyloid fibrils and the interactions remain stable throughout (Figure S5.3). Due to the complexity of different structures of Aβ and to reduce the burden on computational resources, we performed shorter 100ns simulations on four other cyclotide-Aβ complexes. As evident from the plot of RMSD variation, all systems experienced some degree of fluctuations at first, but gradually tended to converge after 60 ns implying that the simulations reached equilibrium. The intermolecular hydrogen bonds in all the four protein-peptide complexes were also monitored throughout the simulation period. The average number of hydrogen bonds between complexes 1IYT-2LAM, 2BEG-2LAM, 2MXU-2LAM and 2M4J-2LAM were 3, 1, 3 and 3, respectively (Figure S5.4). The total stabilizing energy at t=0 ns and t=100 ns frames for the four cyclotide - Aβ complexes is detailed in Table S5.3. Molecular interactions between the other four pairs of Aβ and cyclotides are illustrated in Figure S5.5 and Tables S5.4-S5.7.

5.7 DISCUSSION & CONCLUSIONS

In silico tools such as molecular docking and MD simulations allow one to know the affinity and interaction between the Aβ structures and its inhibitor scaffolds. Aβ is an extracellular protein and in its monomeric form in the membrane adopts an α-helical structure and conformationally transition into β-sheet rich structures in the process of aggregation (Xu et al., 2005). In this work, we evaluated the mode and stability of inter-molecular interactions between cyclotide and diverse models of Aβ. The results of a comprehensive structural analyses presented here shows that the cyclotide conformation and interactions with Aβ fibrils can be reinforced by hydrogen bonding, hydrophobic and

long-range electrostatic interactions between key residues of Aβ peptide and cyclotides. We show a representative complex of disease-relevant $A\beta_{1-42}$ fibril (PDB ID: 2NAO) and cyclotide Cter-M (PDB ID: 2LAM) to highlight the strong interactions between the two molecules. Studies have also shown that there aggregation of Aβ is different in varying pH conditions. Aβ monomers form Aβ fibrils and oligomers at neutral pH conditions (pH 6-8) (Kobayashi et al., 2015). Therefore, we have used water microenvironment at neutral pH conditions for the MD simulations as this has been the standard procedure for modeling interactions between inhibitors and various Aβ oligomers. It is clearly evident that the protein-peptide pairs display several non-covalent interactions throughout the simulation period. The persistence of high number of intermolecular hydrogen bonds and strong non-covalent interactions throughout the trajectory highlights the stable binding of cyclotide in the exposed pockets of amyloid fibril. Moreover, cyclotide disrupts the inter-chain hydrogen bonds and salt bridges in Aβ which are crucial for the fibril structure and shape.

In conclusion, the results of our extensive molecular docking and MD simulations efforts have enabled us to predict the mode of interaction between cyclotides and different physiological forms of Aβ structures. We have described how cyclotides bind tightly to the Aβ chains via strong hydrogen bonds, hydrophobic and π-π interactions and thereby inhibit Aβ aggregation process. Due to their immense potential in peptide therapeutics, cyclotides can be regarded as a new class of cyclic peptide Aβ inhibitors.

5.8 REFERENCES OF CHAPTER 5 (MANUSCRIPT 6)

Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., et al. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. in *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing, SC'06* doi:10.1145/1188455.1188544.

Crescenzi, O., Tomaselli, S., Guerrini, R., Salvadori, S., D'Ursi, A. M., Temussi, P. A., et al. (2002). Solution structure of the Alzheimer amyloid β-peptide (1-42) in an apolar microenvironment. *Eur. J. Biochem.* 269, 5642–5648. doi:10.1046/j.1432-1033.2002.03271.x.

Humphreys, D. D., Friesner, R. A., and Berne, B. J. (1994). A multiple-time-step Molecular Dynamics algorithm for macromolecules. *J. Phys. Chem.* doi:10.1021/j100078a035.

Kobayashi, S., Tanaka, Y., Kiyono, M., Chino, M., Chikuma, T., Hoshi, K., et al. (2015). Dependence pH and proposed mechanism for aggregation of Alzheimer's disease-related amyloid-β(1-42) protein. *J. Mol. Struct.* doi:10.1016/j.molstruc.2015.03.023.

Lu, J. X., Qiang, W., Yau, W. M., Schwieters, C. D., Meredith, S. C., and Tycko, R. (2013). XMolecular structure of β-amyloid fibrils in alzheimer's disease brain tissue. *Cell* 154, 1257. doi:10.1016/j.cell.2013.08.035.

Lührs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Döbeli, H., et al. (2005). 3D structure of Alzheimer's amyloid-β(1-42) fibrils. *Proc. Natl. Acad. Sci. U. S. A.* 102, 17342–17347. doi:10.1073/pnas.0506723102.

Martyna, G. J., Klein, M. L., and Tuckerman, M. (1992). Nosé-Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* 97, 2635–2643. doi:10.1063/1.463940.

Martyna, G. J., Tobias, D. J., and Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* doi:10.1063/1.467468.

Poth, A. G., Colgrave, M. L., Lyons, R. E., Dalya, N. L., and Craik, D. J. (2011). Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10127–10132. doi:10.1073/pnas.1103660108.

Ramírez-Aportela, E., López-Blanco, J. R., and Chacón, P. (2016). FRODOCK 2.0: Fast protein-protein docking server. *Bioinformatics*. doi:10.1093/bioinformatics/btw141.

Sukhwal, A., and Sowdhamini, R. (2015). PPcheck: A webserver for the quantitative analysis of protein-protein interfaces and prediction of residue hotspots. *Bioinform. Biol. Insights*. doi:10.4137/BBI.S25928.

Wälti, M. A., Ravotti, F., Arai, H., Glabe, C. G., Wall, J. S., Böckmann, A., et al. (2016). Atomic-resolution structure of a disease-relevant Aβ(1-42) amyloid fibril. *Proc. Natl. Acad. Sci. U. S. A.* 113, E4976–E4984. doi:10.1073/pnas.1600749113.

Xiao, Y., Ma, B., McElheny, D., Parthasarathy, S., Long, F., Hoshi, M., et al. (2015). Aβ(1-42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nat. Struct. Mol. Biol.* 22, 499–505. doi:10.1038/nsmb.2991.

Xu, Y., Shen, J., Luo, X., Zhu, W., Chen, K., Ma, J., et al. (2005). Conformational transition of amyloid β-peptide. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5403–5407. doi:10.1073/pnas.0501218102.

5.9 SUPPLEMENTARY INFORMATION OF CHAPTER 5 (MANUSCRIPT 6)

Table S5.1: Results of PPCheck webserver for predicting normalized energy per residue for FRODOCK docked models. Best docking pose for each protein-peptide complex is highlighted in green.

a) 1IYT-2LAM

Decoy-ID	Hydrogen Bond Energy (kJ/mol)	Electrostatic Energy (kJ/mol)	van der Waals Energy (kJ/mol)	Total Energy (kJ/mol)	Number of Interface Residues	Normalized Energy per Residue (kJ/mol)
pose2	0.00	-2.58	-99.83	-102.41	45	-2.28
pose4	0.00	0.00	-101.76	-101.76	49	-2.08
pose6	0.00	-2.20	-76.59	-78.79	45	-1.75
pose5	0.00	1.06	-22.65	-21.58	51	-0.42
pose8	0.00	0.00	-20.65	-20.65	52	-0.40
pose3	0.00	0.00	7.13	7.13	50	0.14
pose10	0.00	0.00	36.74	36.74	51	0.72
pose1	-8.40	0.00	86.88	78.48	46	1.71
pose7	0.00	0.00	261.67	261.67	49	5.34
pose9	0.00	0.00	365.36	365.36	51	7.16

b) 2BEG-2LAM

Decoy-ID	Hydrogen Bond Energy (kJ/mol)	Electrostatic Energy (kJ/mol)	van der Waals Energy (kJ/mol)	Total Energy (kJ/mol)	Number of Interface Residues	Normalized Energy per Residue (kJ/mol)
pose3	0.00	0.00	-91.52	-91.52	49	-1.87
pose4	0.00	0.00	-79.78	-79.78	53	-1.51
pose10	0.00	0.00	-12.82	-12.82	51	-0.25
pose9	0.00	0.00	74.80	74.80	61	1.23
pose7	-2.88	0.00	77.83	74.94	42	1.78
pose5	0.00	0.00	114.80	114.80	60	1.91
pose8	0.00	0.00	174.19	174.19	43	4.05
pose1	0.00	0.00	387.97	387.97	47	8.25
pose2	-19.84	0.00	543.28	523.44	43	12.17
pose6	0.00	0.00	819.29	819.29	67	12.23

c) 2MXU-2LAM

Decoy-ID	Hydrogen Bond Energy (kJ/mol)	Electrostatic Energy (kJ/mol)	van der Waals Energy (kJ/mol)	Total Energy (kJ/mol)	Number of Interface Residues	Normalized Energy per Residue (kJ/mol)
pose7	0.00	0.00	-70.81	-70.81	84	-0.84
pose6	-4.13	0.00	-17.44	-21.57	78	-0.28
pose1	-16.55	0.00	139.07	122.51	89	1.38
pose9	0.00	0.00	158.42	158.42	74	2.14
pose8	0.00	0.00	170.22	170.22	76	2.24
pose4	0.00	0.00	302.98	302.98	74	4.09
pose3	-7.25	0.00	362.80	355.55	84	4.23
pose10	0.00	0.00	463.90	463.90	81	5.73
pose5	0.00	0.00	616.12	616.12	94	6.55
pose2	-12.92	0.00	1034.95	1022.03	98	10.43

d) 2NAO-2LAM

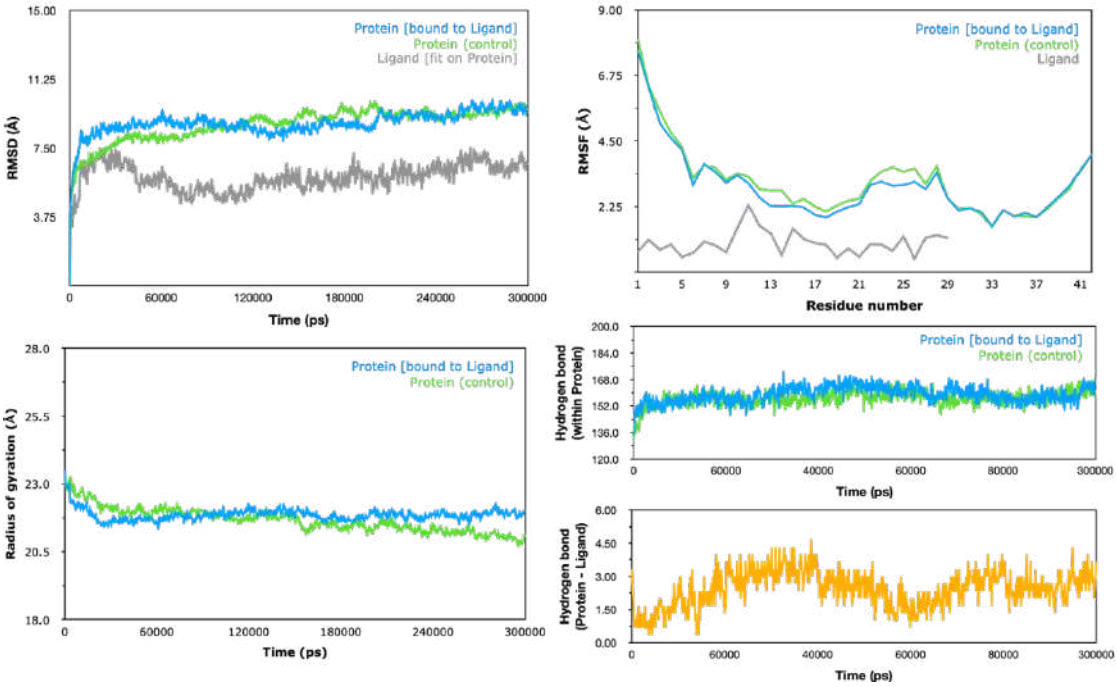
Decoy-ID	Hydrogen Bond Energy (kJ/mol)	Electrostatic Energy (kJ/mol)	van der Waals Energy (kJ/mol)	Total Energy (kJ/mol)	Number of Interface Residues	Normalized Energy per Residue (kJ/mol)
pose10	-4.17	0.00	-95.50	-99.67	65.00	-1.53
pose5	-4.78	0.00	-15.15	-19.94	72.00	-0.28
pose8	-14.51	4.94	146.56	136.99	73.00	1.88
pose7	0.00	0.00	135.39	135.39	68.00	1.99
pose1	0.00	0.00	144.70	144.70	63.00	2.30
pose6	-4.27	0.00	254.73	250.46	70.00	3.58
pose2	-18.43	12.34	337.50	331.40	84.00	3.95
pose4	0.00	0.00	324.05	324.05	76.00	4.26
pose3	-7.91	0.00	382.54	374.63	71.00	5.28
pose9	-8.57	0.00	1144.61	1136.04	82.00	13.85

e) 2M4J-2LAM

Decoy-ID	Hydrogen Bond Energy (kJ/mol)	Electrostatic Energy (kJ/mol)	van der Waals Energy (kJ/mol)	Total Energy (kJ/mol)	Number of Interface Residues	Normalized Energy per Residue (kJ/mol)
pose4	0.00	0.00	-67.47	-67.47	95.00	-0.71
pose9	-9.08	0.00	-10.26	-19.33	83.00	-0.23
pose2	-15.64	0.00	14.70	-0.94	92.00	-0.01
pose5	-6.80	0.00	61.42	54.62	94.00	0.58
pose7	-17.90	0.00	129.68	111.79	94.00	1.19
pose8	-36.02	0.00	152.62	116.59	87.00	1.34
pose1	-16.73	0.00	208.95	192.22	97.00	1.98
pose3	0.00	0.00	254.71	254.71	92.00	2.77
pose6	-89.56	0.00	659.28	569.71	95.00	6.00
pose10	0.00	0.00	987.48	987.48	96.00	10.29

Figure S5.1: Plots of triplicate 300 ns MD simulations for cyclotide (PDB ID: 2LAM) and $A\beta_{1-42}$ disease relevant $A\beta_{1-42}$ fibrils (PDB ID: 2NAO) complex, averaged over triplicate simulations. (A) RMSD (top left panel), RMSF (top right panel), Radius of gyration (bottom left panel), number of hydrogen bonds (bottom right panels) and (B) Number of hydrogen bonds in β-sheets at binding site (2NAO; chains A-F).

A



B

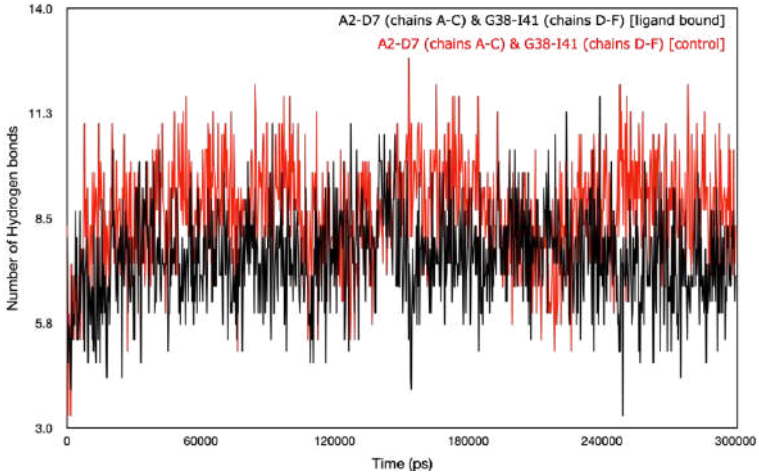
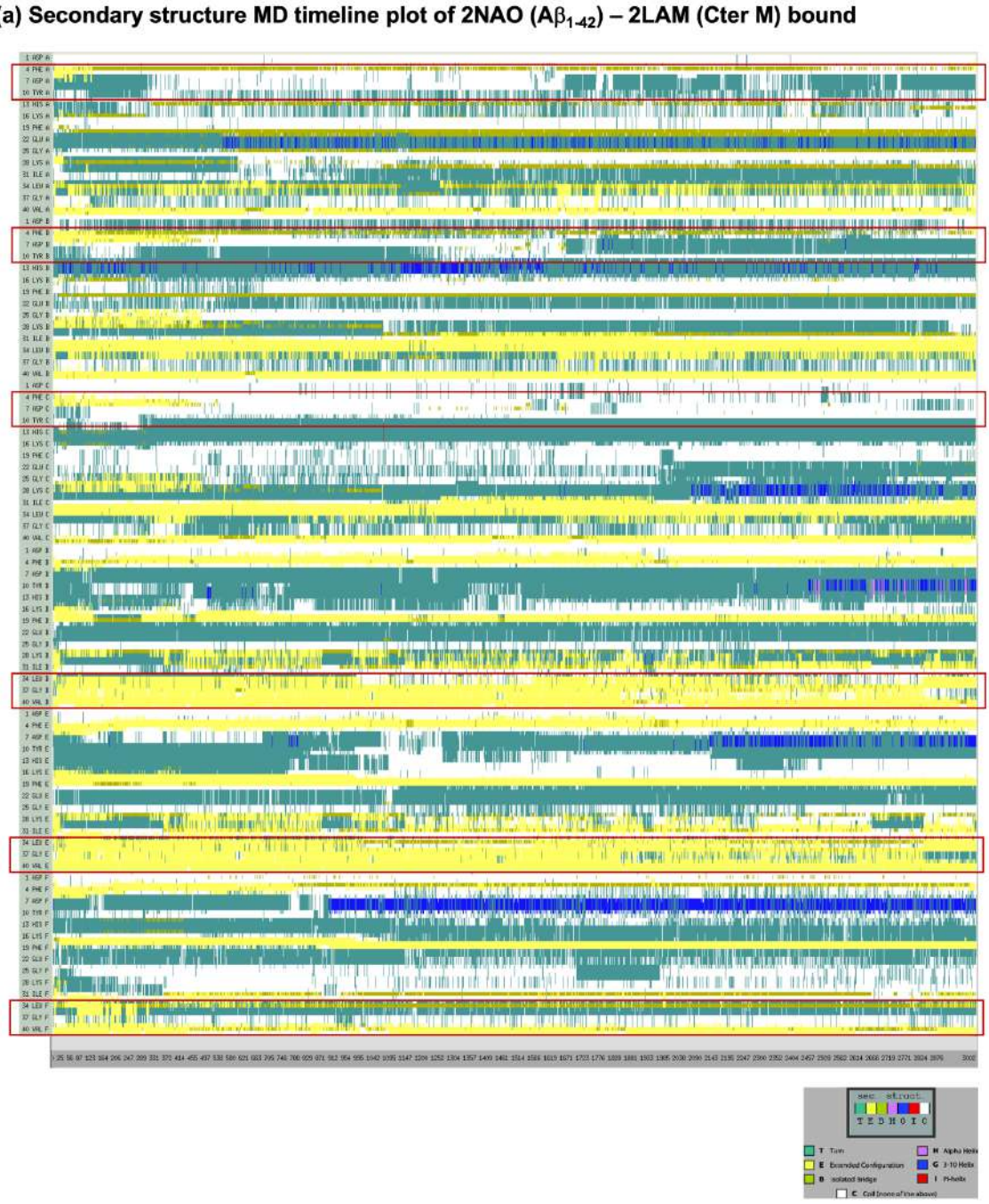


Figure S5.2: Secondary structure timeline analysis during 300 ns MD simulation of (A) $A\beta_{1-42}$ fibril (PDB ID: 2NAO) bound to cyclotide Cter-M (B) unbound $A\beta_{1-42}$ fibril (PDB ID: 2NAO). Colour code explanation of the secondary structures are shown in the small panel below. T denotes turn (aqua); E represents the β-sheet (yellow); B represents isolated bridges (dark yellow); H denotes α-helix (pink); G indicates the 3_{10} helix (blue); I denotes π-helix (red) and C indicates random coils (white). These structural analyses were computed using VMD timeline plugin.



(b) Secondary structure MD timeline plot of 2NAO ($A\beta_{1-42}$) unbound

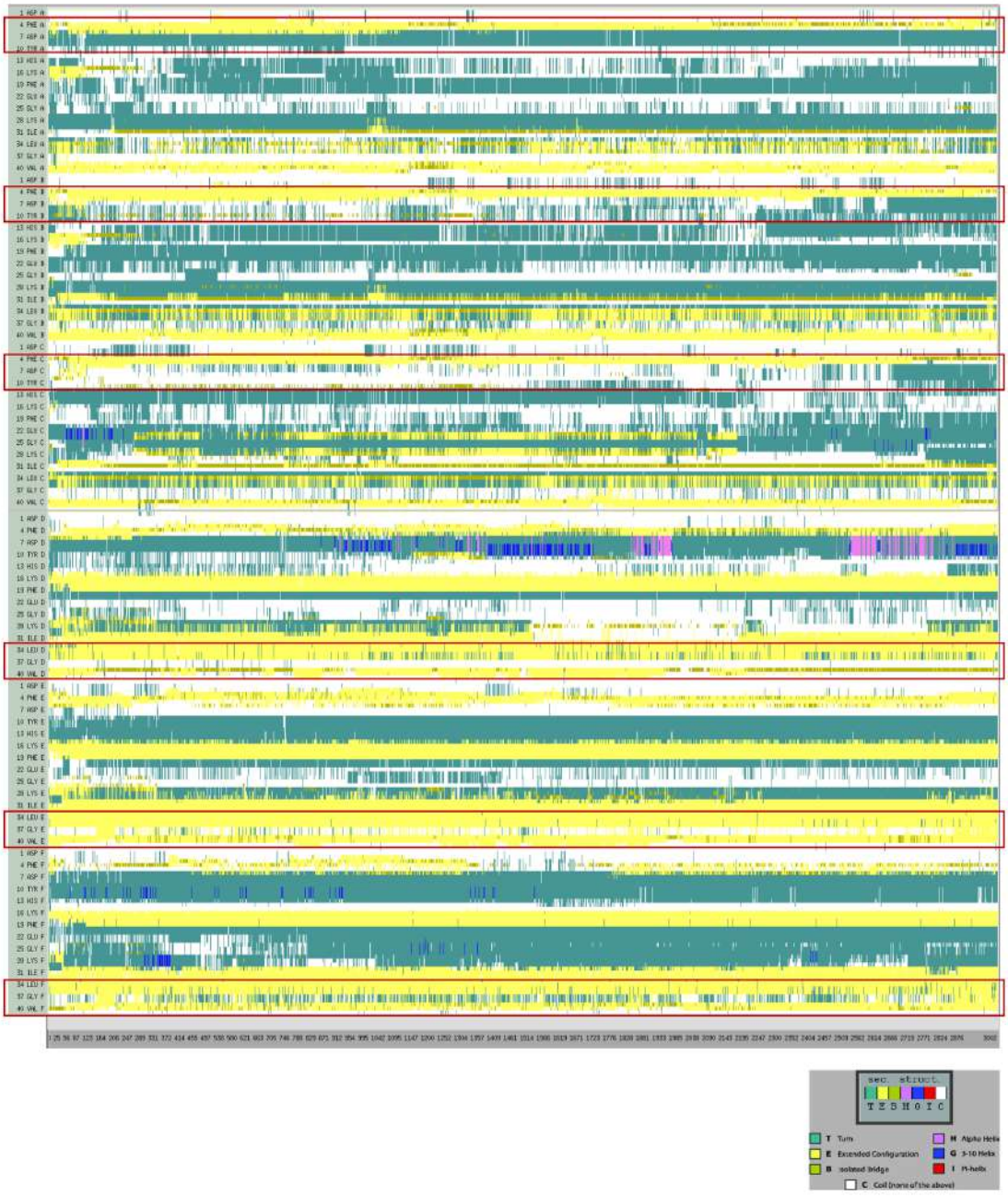


Table S5.2. Analysis of protein-protein interactions at t=0 ns (top) and t=300ns (bottom) frames between cyclotide Cter-M (PDB ID: 2LAM) and $A\beta_{1-42}$ fibril (PDB ID: 2NAO).

Interactions	[a]Maestro	[a]PPCheck	2NAO				2LAM				[b]Type of Bond	Distance Å
			Residue-1				Residue-2					
			Res No.	Res Name	Chain	Atom Name	Res No.	Res Name	Chain	Atom Name		
t=0 ns												
Hydrogen Bond	✓	-	13	His	C	NE2	7	Glu	Z	N	SB	3.2
	✓	-	13	His	B	ND1	3	Pro	Z	O	SB	3.7
	✓	-	14	His	B	NE2	6	Glu	Z	N	SB	3.5
	✓	✓	38	Gly	E	O	15	Tyr	Z	OH	BS	2.7
	-	✓	39	Val	F	N	15	Tyr	Z	OH	BS	2.8
	-	✓	38	Gly	F	N	15	Tyr	Z	OH	BS	2.9
Aromatic Hydrogen Bond	✓	-	4	Phe	C	CZ	29	Asn	Z	O	SB	3.8
	✓	-	6	His	C	CE1	1	Gly	Z	O	SB	4.4
	✓	-	13	His	C	CE1	26	Cys	Z	O	SB	3.2
Hydrophobic Interactions	-	✓	10	Tyr	C	CB	25	Ile	Z	CB	SS	5.0
	-	✓	40	Val	D	CB	16	Val	Z	CB	SS	6.9
	-	✓	40	Val	E	CB	15	Tyr	Z	CB	SS	6.5
	-	✓	40	Val	E	CB	16	Val	Z	CB	SS	6.6
	-	✓	40	Val	F	CB	15	Tyr	Z	CB	SS	5.3
Electrostatic Interactions	-	✓	13	His	C	CB	7	Glu	Z	CB	SS	7.9
	-	✓	14	His	C	CB	7	Glu	Z	CB	SS	7.6
t=300 ns												
Hydrogen Bond	✓	✓	6	His	C	N	1	Gly	Z	O	BB	3.0
	✓	✓	38	Gly	D	N	6	Gly	Z	O	BB	2.7
	✓	-	38	Gly	F	O	15	Tyr	Z	OH	SB	3.3
Aromatic Hydrogen Bond	✓	-	4	Phe	C	CE2	1	Gly	Z	O	SB	3.4
Hydrophobic Interactions	-	✓	10	Tyr	C	CB	25	Ile	Z	CB	SS	5.6
	-	✓	40	Val	D	CB	15	Tyr	Z	CB	SS	5.0
	-	✓	40	Val	D	CB	16	Val	Z	CB	SS	5.5
	-	✓	40	Val	E	CB	15	Tyr	Z	CB	SS	4.8
	-	✓	39	Val	F	CB	11	Leu	Z	CB	SS	7.0
Electrostatic Interactions	-	✓	13	His	A	CB	7	Glu	Z	CB	SS	9.3

[a] Protein-protein interaction identified from Maestro and/or PPCheck is highlighted with ✓.

[b] "SS" represents sidechain-sidechain interaction, "SB" represents sidechain-backbone interaction, and "BB" represents backbone-backbone mode of interaction between the two interacting amino acids.

Figure S5.3: Snapshots of MD simulation of cyclotide (PDB ID: 2LAM; cyan surface representation) and (A) $A\beta_{1-42}$ monomer (PDB ID: 1IYT, green cartoon representation), (B) $A\beta_{17-42}$ U-shaped pentamer (PDB ID: 2BEG, green cartoon representation), (C) $A\beta_{11-42}$ S-shaped model (PDB ID: 2MXU, green cartoon representation), (D) $A\beta_{1-40}$ isolated from brain of AD patient (PDB ID: 2M4J, green cartoon representation) complexes at different time points along the simulation period.

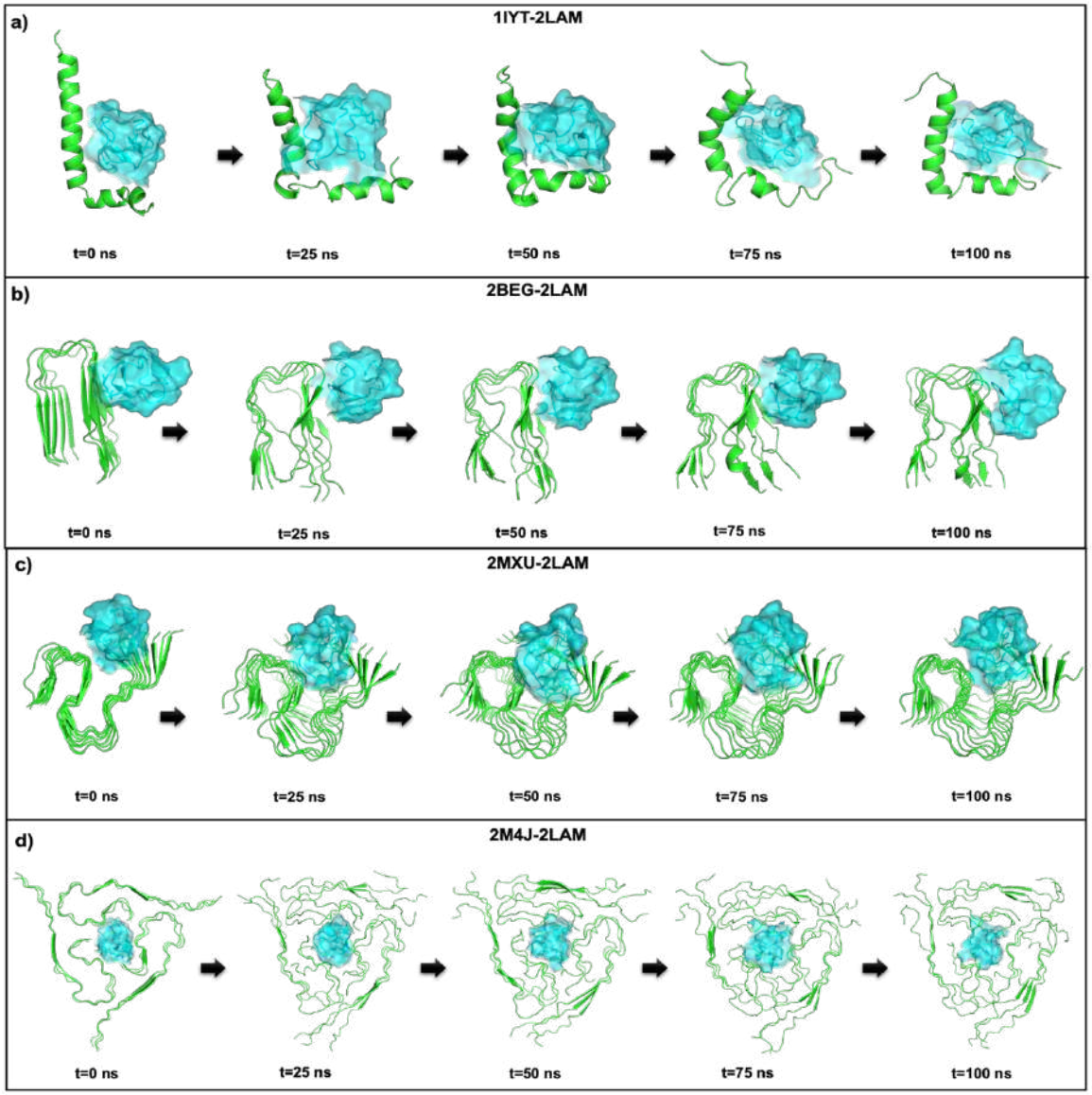
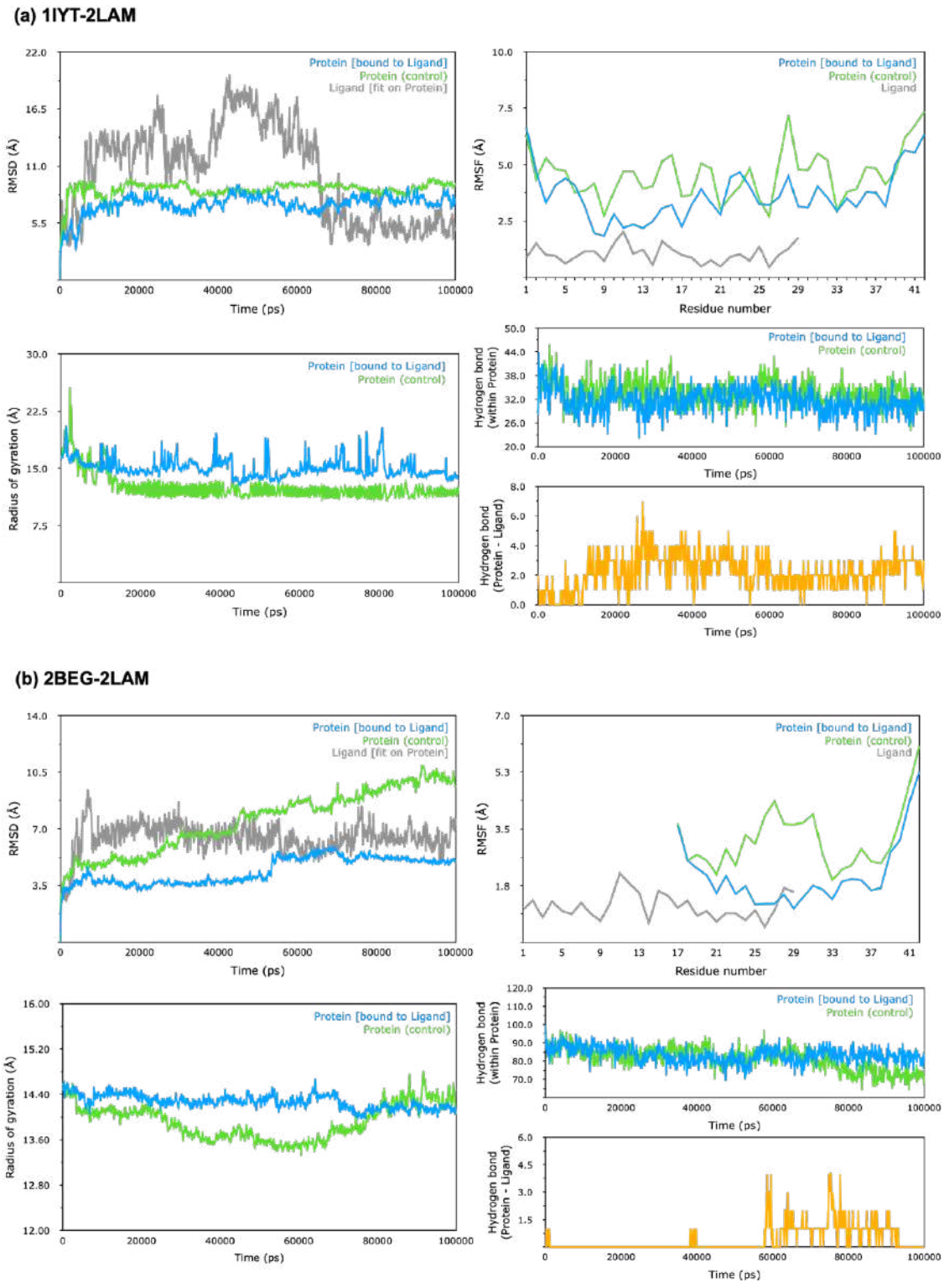
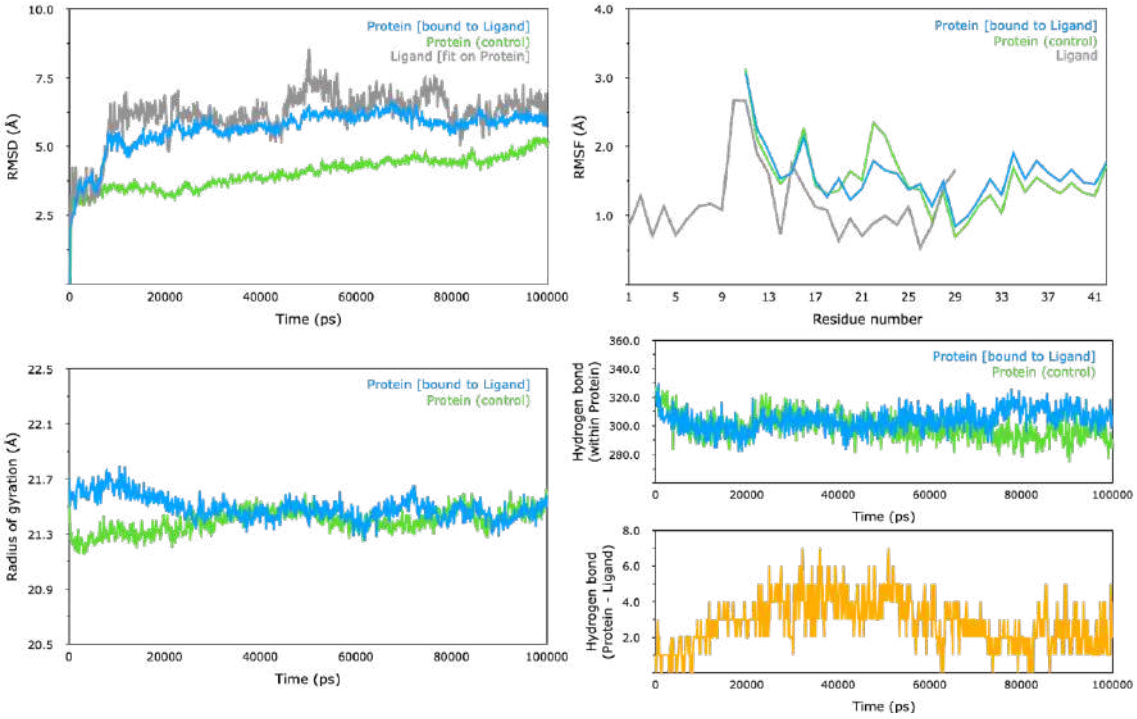


Figure S5.4: Plots of 100ns simulations showing RMSD (top left panel), RMSF (top right panel), Radius of gyration (bottom left panel) and number of hydrogen bonds (bottom right panel) for complexes between cyclotide (PDB ID: 2LAM) and (A) $A\beta_{1-42}$ monomer (PDB ID: 1IYT), (B) $A\beta_{17-42}$ U-shaped pentamer (PDB ID: 2BEG), (C) $A\beta_{11-42}$ S-shaped model (PDB ID: 2MXU) and (D) $A\beta_{1-40}$ isolated from brain of AD patient (PDB ID: 2M4J).



(c) 2MXU-2LAM



(d) 2M4J-2LAM

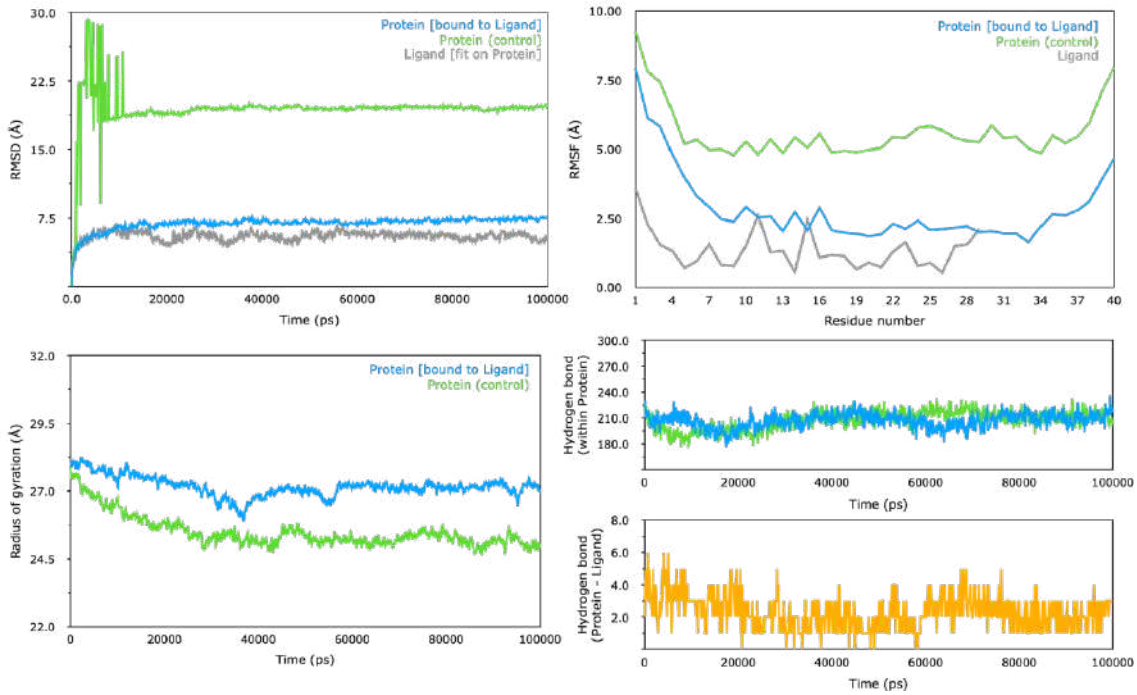
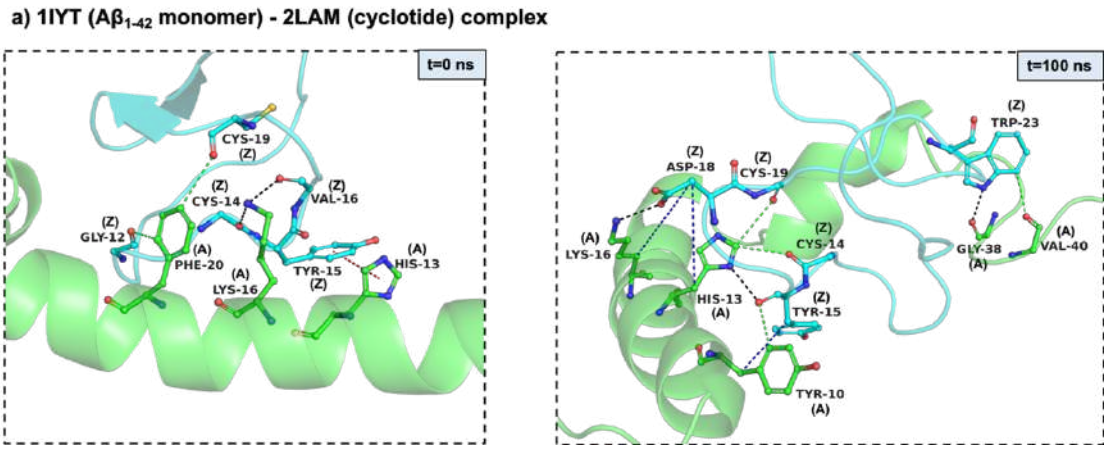


Table S5.3: The total stabilizing energy at t=0 ns and t=100 ns for four other cyclotide - Aβ complexes.

Total Stabilizing Energy	1IYT-2LAM		2BEG-2LAM		2MXU-2LAM		2M4J-2LAM	
	t=0 ns	t=100 ns	t=0 ns	t=100 ns	t=0 ns	t=100 ns	t=0 ns	t=100 ns
Hydrogen Bond Energy (kJ/mol)	0.00	0.00	0.00	0.00	-8.83	-21.18	-11.17	-20.33
Electrostatic Energy (kJ/mol)	0.00	-21.84	0.00	0.00	0.00	0.00	0.00	0.00
Van der Waals Energy (kJ/mol)	-97.70	-176.85	-123.14	-129.41	-180.95	-208.29	-249.88	-271.29
Total Stabilizing Energy (kJ/mol)	-97.70	-198.69	-123.14	-129.41	-189.77	-229.47	-261.05	-291.62
Number of interface residues	43	53	52	52	87	91	101	97
Normalized Energy per residue (kJ/mol)	-2.27	-3.75	-2.37	-2.49	-2.18	-2.52	-2.58	-3.01
No. of Hydrophobic Interactions	0	1	2	2	1	4	3	4
No. of van der Waals Pairs	3791	6116	5534	5736	8970	8974	11085	10841
No. of Salt Bridges	0	1	0	0	0	0	0	0
No. of Potential Favourable Electrostatic Interactions	0	2	0	0	0	0	0	0

Figure S5.5: Molecular interactions between cyclotide (PDB ID: 2LAM; chain Z; cyan cartoon representation) and (A) $A\beta_{1-42}$ monomer (PDB ID: 1IYT; chain A; green cartoon representation), (B) $A\beta_{17-42}$ U-shaped pentamer (PDB ID: 2BEG; chain A-E; green cartoon representation), (C) $A\beta_{11-42}$ S-shaped model (PDB ID: 2MXU; chain A-L; green cartoon representation), (D) $A\beta_{1-40}$ isolated from brain of AD patient (PDB ID: 2M4J; chain A-I; green cartoon representation) at the beginning (0^{th} ns snapshot) and end (100^{th} ns snapshot) of the simulation period. Colour scheme for interactions used: classical hydrogen bonds in black, aromatic hydrogen bonds in green, hydrophobic interactions in blue, electrostatic interactions in orange, salt-bridges in pink and π-π interactions in red.



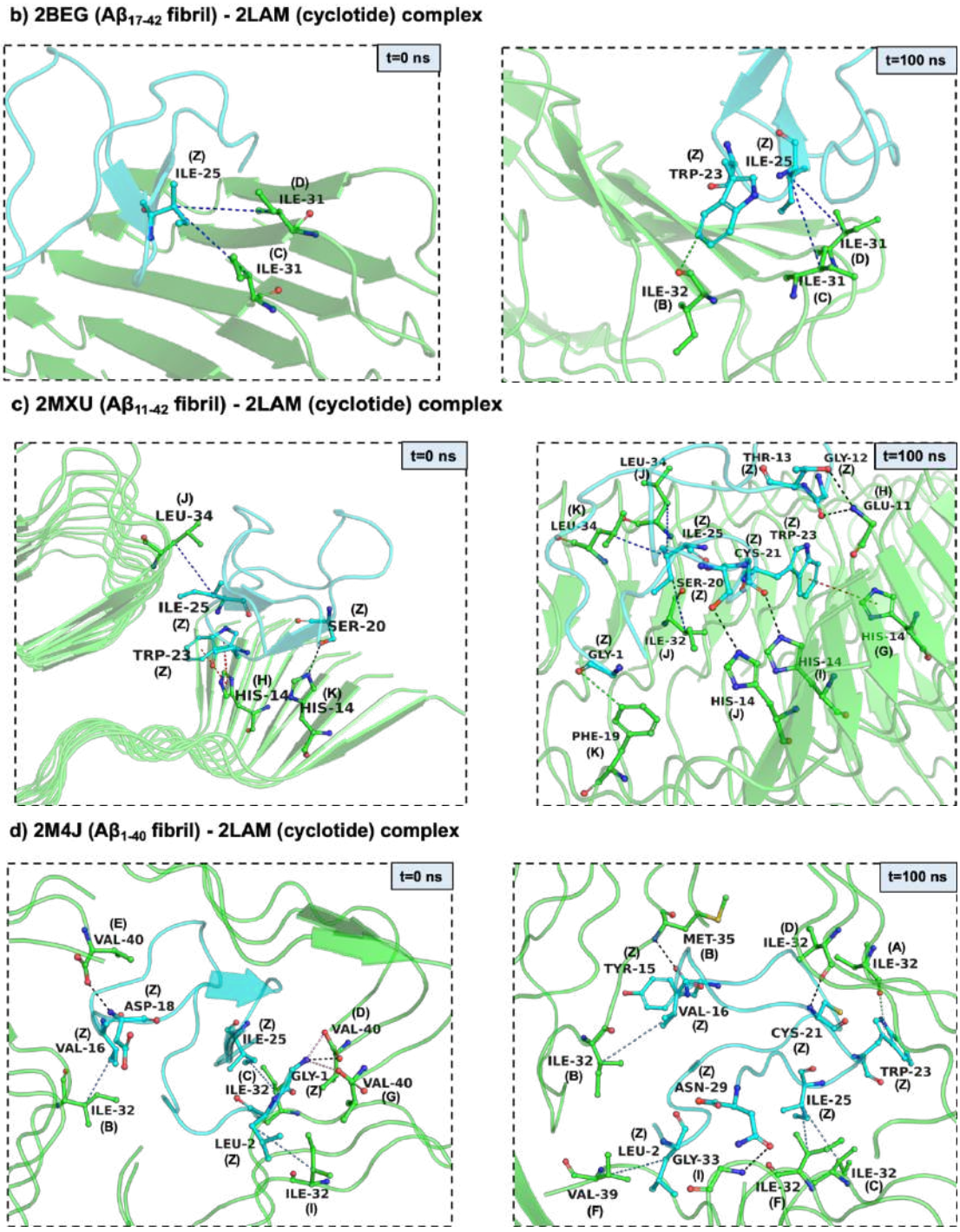


Table S5.4: Analysis of protein-protein interactions at t=0 ns (top) and t=100ns (bottom) frames between cyclotide (PDB ID: 2LAM) and Aβ1-42 monomer (PDB ID: 1IYT).

Interactions	[a]Maestro	[a]PPCheck	1IYT				2LAM				[b]Type of Bond	Distance Å
			Residue-1				Residue-2					
			Res No.	Res Name	Chain	Atom Name	Res No.	Res Name	Chain	Atom Name		
t=0 ns												
Hydrogen Bond	✓	-	16	LYS	A	NZ	14	CYS	Z	O	SB	3.3
	✓	-	16	LYS	A	NZ	16	VAL	Z	O	SB	2.8
	✓	-	20	PHE	A	CD2	12	GLY	Z	O	SB	3.6

Aromatic Hydrogen Bond	✓	-	20	PHE	A	CZ	19	CYS	Z	O	SB	3.5
π-π Interactions	✓	-	13	HIS	H	ring	15	TYR	Z	ring	SS	5.1
t=100 ns												
Hydrogen Bond	✓	-	16	LYS	A	NZ	18	ASP	Z	OD2	SB	2.9
	✓	-	13	HIS	A	ND1	15	TYR	Z	O	SB	3.0
	✓	-	38	GLY	A	O	23	TRP	Z	NE1	BS	2.8
Aromatic Hydrogen Bond	✓	-	13	HIS	A	CE1	14	CYS	Z	O	SB	3.4
	✓	-	13	HIS	A	CE1	19	CYS	Z	O	SB	3.6
	✓	-	10	TYR	A	CD1	15	TYR	Z	O	SB	3.6
	✓	-	40	VAL	A	O	23	TRP	Z	CZ2	BS	3.2
Hydrophobic Interactions	-	✓	10	TYR	A	CB	15	TYR	Z	CB	SS	5.2
Electrostatic Interactions	-	✓	13	HIS	A	CB	18	ASP	Z	CB	SS	6.6
	-	✓	16	LYS	A	CB	18	ASP	Z	CB	SS	8.3

[a] Protein-protein interaction identified from Maestro and/or PPCheck is highlighted with as ✓.

[b] "SS" represents sidechain-sidechain interaction, "SB" represents sidechain-backbone interaction, and "BB" represents backbone-backbone mode of interaction between the two interacting amino acids.

Table S5.5: PPCheck results for protein-protein interactions at t=0 ns (top) and t=100ns (bottom) frames between cyclotide (PDB ID: 2LAM) and $A\beta_{17-42}$ U-shaped pentamer (PDB ID: 2BEG).

Interactions	**[a]Maestro**	**[a]PPCheck**	**2BEG**				**2LAM**				**[b]Type of Bond**	**Distance Å**
			Residue-1				**Residue-2**					
			Res No.	**Res Name**	**Chain**	**Atom Name**	**Res No.**	**Res Name**	**Chain**	**Atom Name**		
t=0 ns												
Hydrophobic Interactions	-	✓	31	ILE	C	CB	25	ILE	Z	CB	SS	6.4
	-	✓	31	ILE	D	CB	25	ILE	Z	CB	SS	6.5
t=100 ns												
Aromatic Hydrogen Bond	✓	-	32	ILE	B	O	23	TRP	Z	CZ3	BS	3.3
Hydrophobic Interactions	-	✓	31	ILE	C	CB	25	ILE	Z	CB	SS	6.6
	-	✓	31	ILE	D	CB	25	ILE	Z	CB	SS	5.8

[a] Protein-protein interaction identified from Maestro and/or PPCheck is highlighted with as ✓.

[b] "SS" represents sidechain-sidechain interaction, "SB" represents sidechain-backbone interaction, and "BB" represents backbone-backbone mode of interaction between the two interacting amino acids.

Table S5.6: PPCheck results for protein-protein interactions at t=0 ns (top) and t=100ns (bottom) frames between cyclotide (PDB ID: 2LAM) and $A\beta_{11-42}$ S-shaped model (PDB ID: 2MXU).

Interactions	**[a]Maestro**	**[a]PPCheck**	**2MXU**				**2LAM**				**[b]Type of Bond**	**Distance Å**
			Residue-1				**Residue-2**					
			Res No.	**Res Name**	**Chain**	**Atom Name**	**Res No.**	**Res Name**	**Chain**	**Atom Name**		
t=0 ns												
Hydrogen Bond	✓	✓	14	HIS	K	NE2	20	SER	Z	OG	SS	3.1

Hydrophobic Interactions	-	✓	34	LEU	J	CB	25	ILE	Z	CB	SS	5.4
π-π Interactions	✓	-	14	HIS	H	ring	23	TRP	Z	ring 1	SS	4.9
	✓	-	14	HIS	H	ring	23	TRP	Z	ring 2	SS	4.5
t=100 ns												
Hydrogen Bond	✓	-	11	GLU	H	N	12	GLY	Z	O	BB	2.8
	-	✓	11	GLU	H	N	13	THR	Z	OG1	BS	2.9
	✓	✓	14	HIS	I	NE2	21	CYS	Z	O	SB	3.1
	-	✓	14	HIS	J	NE2	20	SER	Z	OG	SS	3.2
Aromatic Hydrogen Bond	✓	-	19	PHE	K	CE1	1	GLY	Z	O	SB	3.9
Hydrophobic Interactions	-	✓	32	ILE	J	CB	25	ILE	Z	CB	SS	6.8
	-	✓	34	LEU	J	CB	25	ILE	Z	CB	SS	6.4
	-	✓	34	LEU	K	CB	25	ILE	Z	CB	SS	5.2
π-π Interactions	✓	-	14	HIS	G	ring	23	TRP	Z	ring 2	SS	5.2

[a] Protein-protein interaction identified from Maestro and/or PPCheck is highlighted with as ✓.

[b] "SS" represents sidechain-sidechain interaction, "SB" represents sidechain-backbone interaction, and "BB" represents backbone-backbone mode of interaction between the two interacting amino acids.

Table S5.7: PPCheck results for protein-protein interactions at t=0 ns (top) and t=100ns (bottom) frames between cyclotide (PDB ID: 2LAM) and $A\beta_{1-40}$ isolated from brain of AD patient (PDB ID: 2M4J).

Interactions	**[a]Maestro**	**[a]PPCheck**	**2M4J**				**2LAM**				**[b]Type of Bond**	**Distance Å**
			Residue-1				**Residue-2**					
			Res No.	**Res Name**	**Chain**	**Atom Name**	**Res No.**	**Res Name**	**Chain**	**Atom Name**		
t=0 ns												
Hydrogen Bond	✓	✓	40	VAL	E	OXT	18	ASP	Z	N	SB	2.7
	✓	-	40	VAL	D	O	1	GLY	Z	N	BB	2.7
Hydrophobic Interactions	-	✓	32	ILE	B	CB	16	VAL	Z	CB	SS	5.5
	-	✓	32	ILE	C	CB	25	ILE	Z	CB	SS	6.2
	-	✓	32	ILE	I	CB	2	LEU	Z	CB	SS	6.3
Salt Bridges	✓	-	40	VAL	D	OXT	1	GLY	Z	N	SB	3.2
	✓	-	40	VAL	G	OXT	1	GLY	Z	N	SB	4.3
t=100 ns												
Hydrogen Bond	✓	✓	35	MET	B	N	15	TYR	Z	O	BB	2.8
	✓	✓	32	ILE	D	O	21	CYS	Z	N	BB	2.8
	✓	-	33	GLY	I	N	29	ASN	Z	OD1	SB	3.4
Aromatic Hydrogen Bond	✓	-	32	ILE	A	O	23	TRP	Z	ring	SB	4.2
Hydrophobic Interactions	-	✓	32	ILE	B	CB	16	VAL	Z	CB	SS	6.4
	-	✓	32	ILE	C	CB	25	ILE	Z	CB	SS	5.3
	-	✓	32	ILE	F	CB	25	ILE	Z	CB	SS	5.3
	-	✓	39	VAL	F	CB	2	LEU	Z	CB	SS	5.9

[a] Protein-protein interaction identified from Maestro and/or PPCheck is highlighted with as ✓.

[b] "SS" represents sidechain-sidechain interaction, "SB" represents sidechain-backbone interaction, and "BB" represents backbone-backbone mode of interaction between the two interacting amino acids.

List of Publications

Included in this thesis

1. **Manuscript 1 (Included as part of Chapter 2)**
   Kalmankar N. V., Venkatesan R., Balaram P. and Sowdhamini R. (2020) Transcriptomic profiling of the medicinal plant *Clitoria ternatea*: Identification of potential genes in cyclotide biosynthesis. *Scientific Reports*, 10: 12658. doi: 10.1038/s41598-020-69452-7
2. **Manuscript 2 (Included as part of Chapter 5)**
   Kalmankar N. V., Hari H., Sowdhamini R and Venkatesan R. (2021) Disulfide-rich, cyclic peptides from *Clitoria ternatea* protect against β-amyloid toxicity and oxidative stress in transgenic *Caenorhabditis elegans*. *Journal of Medicinal Chemistry*, 64: 7422-7433. doi: 10.1021/acs.jmedchem.1c00033. [Cover article, 2021]
3. **Manuscript 3 (Included as part of Chapter 3)**
   Kalmankar N. V.*, Balaram P.* and Venkatesan R.* (2021) Mass spectrometric analysis of cyclotides from *Clitoria ternatea*: Xxx-Pro bond fragmentation as convenient diagnostic of Pro residue positioning. *Chemistry - an Asian Journal*. 16: 2920. doi: 10.1002/asia.202100585. [Cover article, 2021] [*corresponding authors].
4. **Manuscript 4 (Included as part of Chapter 4)**
   Kalmankar N. V., Pavalam M., Indrakumar S. and Sowdhamini R. (2021) DSDBASE 2.0: Updated version of DiSulphide DataBASE. (*Under review in Database*)
5. **Manuscript 5 (Included as part of Chapter 3)**
   Kalmankar N. V., Sowdhamini R., Venkatesan R. and Balaram P. (2021) Proteome-wide discovery of macrocyclic cysteine knot peptides: Identification of cyclo-oligomerized cyclotides from butterfly pea. (*In preparation*)
6. **Manuscript 6 (Included as part of Chapter 5)**
   Kalmankar N. V., Venkatesan R. and Sowdhamini R. (2021) Effects of cyclotide on conformational dynamics and destabilization of β-amyloid fibrils through molecular dynamics simulations. (*In preparation for submission to the Journal of Chemical Information and Modeling*).

Other publications

1. **Kalmankar N. V**., Ramakrishnan C. and Balaram P. (2014) Sparsely populated residue conformations in protein structures: Revisiting "experimental" Ramachandran maps. *Proteins*, 82: 1101–1112. doi: 10.1002/prot.24384.
2. Vaigundan D., **Kalmankar N. V**., Krishnappa J., Gowda N. Y., Kutty A. V. M. and Krishnaswamy P. R. (2014) A novel mutation in the transglutaminase-1 gene in an autosomal recessive congenital ichthyosis patient. *BioMed Research International*, 2014: 706827. doi: 10.1155/2014/706827.
3. Toniolo C., Crisma M., Formaggio F., Alemán C., Ramakrishnan C., **Kalmankar N. V**. and Balaram P. (2017) Intramolecular backbone…backbone hydrogen bonds in polypeptide conformations. The other way around: ε-turn. *Biopolymers*, 108: e22911. doi:10.1002/bip.22911.
4. Crisma M., Formaggio F., Alemán C., Torras J., Ramakrishnan C., **Kalmankar N. V**., Balaram P. and Toniolo C. (2017) The fully-extended conformation in peptides and proteins. *Peptide Science*, 110: e23100. doi: 10.1002/bip.23100.
5. **Kalmankar N. V**., Vijaysarathy M., Ramakrishnan C. and Balaram P. (2021) The pivotal role of conformationally unique Asparagine residues in the structure of Leucine-rich repeat (LRR) proteins (*In preparation*).
6. **Kalmankar N. V.**, Vijaysarathy M., Banerjee M., Joshi N. V. and Balaram P. (2021) Mapping Sites of amino acid conservation on the three-dimensional structures of enzymes: Identification of putative functional sites in hypothetical proteins (*In preparation*).

Conference/Workshop Attendance

1. Hands-on training on "Data Analysis of Transcriptomics & Proteomics" at Shodhaka Life Sciences Pvt. Ltd., Bangalore (22-27 September 2016).
2. **Kalmankar N. V.**, Balaram P., Sowdhamini R., Venkatesan R. Sequencing and Structure-Activity Relationships of Cyclotides from the Medicinal Plant, *Clitoria ternatea*. DST- SERB School on Chemical Ecology at NCBS, Bangalore (3-16 July 2017). Poster presentation (Best Poster Award).
3. SASSIE training course at Indian Institute of Science, Bangalore (21-22 September 2017).
4. **Kalmankar N. V.**, Sowdhamini R., Balaram P., Venkatesan R. Sequencing and Structure-Activity Relationships and Biochemical Studies of Cyclotides from the

Medicinal Plant, *Clitoria ternatea*. Structure Across Scales meeting, at NCBS, Bangalore (7-8 October 2017). Poster presentation.

5. **Kalmankar N. V.**, Sowdhamini R., Balaram P., Venkatesan R. Structure-Activity Relationships of Cyclotides from the Medicinal Plant, *Clitoria ternatea*. Indo-German Workshop on Plant-Insect Interactions Across Gradients at NCBS, Bangalore (16-17 November 2017). Oral presentation.
6. **Kalmankar N. V.**, Sowdhamini R., Balaram P., Venkatesan R. *Isolation,* Characterization and Conformational Analysis of Cyclotides, a Class of Macrocyclic Disulfide Bonded Plant Peptides. 4th International Conference on Circular Proteins and Peptides, at Kawasaki, Japan (28-30 November 2018). 'Hotspot' talk and poster presentation. Awarded the travel grant.
7. **Kalmankar N. V.**, Sowdhamini R., Balaram P., Venkatesan R. Isolation, Characterization and Conformational Analysis of Cyclotides, a Class of Macrocyclic Disulfide Bonded Plant Peptides. 10th International Peptide Symposium (10th IPS) in conjunction with the 55th Japanese Peptide Symposium (55th JPS) at Kyoto, Japan (3-7 December 2018). Poster presentation. Awarded the travel grant.

Synopsis

Natural products that include small molecules, peptides and macromolecules are valuable sources for applications in agriculture and medicine. Of particular interest among these versatile compounds are peptides from plants, cone-snails, snakes, spiders, scorpions, fungi, and bacteria. A largely underexplored class of bioactive peptides are the ribosomally synthesized and post-translationally modified peptides (RiPPs), also known as ribosomal natural products. Such peptides have been evolutionarily optimized over millions of years for the purposes of regulating specific enzymes and pathways, ion channels and receptors. Therefore, nature provides us with excellent combinatorial peptide libraries that could be mined for variety of applications, but particularly for drug discovery and development. Cyclotides are one such unique class of gene-encoded, ribosomally synthesized macrocyclic peptides (26-37 residues) produced in several plant species as a form of their defense strategy (Jennings et al., 2001). In addition to the head-to-tail cyclic backbone, the three disulfide bonded framework (Cys I-IV, II-V, III-VI) together form a cyclic cystine-knot (CCK) arrangement (Wang et al., 2009a, 2009b). This circularised knotted arrangement makes cyclotides remarkably stable against enzymatic, chemical and thermal degradation (Daly et al., 2009). Structurally, cyclotides can be divided into two main subfamilies i.e. Möbius and Bracelet. The main difference between the two subfamilies is that Möbius cyclotides contain a *cis*-proline residue in loop 5 region, whereas this is absent in the Bracelet subtype (Rosengren et al., 2003).

So far cyclotides have been discovered from several plant families, including violet (Violaceae), coffee (Rubiaceae), cucurbit (Cucurbitaceae), pea (Fabaceae), potato (Solanaceae), and grass (Poaceae). However, *Clitoria ternatea* (butterfly pea, also known as "Shankhapushpi" in India) is the only species from the Fabaceae plant family currently reported to produce a large assortment of cyclotides across all its tissues (Poth et al., 2011). *C. ternatea,* an Indian medicinal plant, has been shown to produce more than 100 cyclotides (Oguis et al., 2019). In this thesis, I worked towards purifying and sequencing novel cyclotides from various tissues of *C. ternatea* using chromatographic, mass-spectrometric and transcriptomic techniques. Furthermore, I updated the disulfide bond database - DSDBASE2.0 to include the latest PDB entries and incorporated multiple tools in the webserver for structure prediction of disulfide-rich peptides. I used an in-house algorithm - RANMOD to model the cyclic conformations of cyclotide sequences. At last, I investigated the therapeutic potential of cyclotides, specifically as agents against neurodegenerative disorders. **Chapter 1** (Introduction) of this thesis reviews the current

understanding on cyclotides, their diversity in plant families, their structural characteristics and provides an overview of their biosynthesis and future applications.

Chapter 2 discusses the transcriptome analysis of *C. ternatea* plant, specifically on the mining efforts of cyclotide precursor genes produced in four plant tissues - pods, stems, leaves and flowers. Previous RNA-Seq studies have underscored the genetic origin of cyclotides from the Fabaceae plants to be embedded within the albumin-1 genes, highlighting an unusual architecture compared to the counterpart plant families like Violaceae or Rubiaceae (Poth et al., 2011). However, the mechanism of cyclotide production and the recruitment of enzymes involved in proper folding and processing remains to be understood. In this chapter, the *de novo* transcriptome assembly of *C. ternatea* and identification of 71 precursor genes of cyclotides are reported (Kalmankar et al., 2020). Out of 71 unique cyclotide precursor genes obtained, 51 sequences display unique cyclotide domains, of which 26 are novel cyclotide sequences. Differential expression analysis revealed that numerous cyclotide genes were highly expressed in one plant tissue, but were minimally expressed/absent in the other tissues. Additionally, we have also proposed a model of cyclotide biosynthesis, wherein we followed the presence and expression of oxidative folding enzymes responsible for proper folding and processing of cyclotides in the cell. The genes corresponding to the 'cyclizing' enzyme- asparaginyl endopeptidase (AEP), and oxidative folding enzymes such as protein-disulphide isomerases (PDI), ER oxidoreductin-1 (ERO1) and peptidylprolyl cis-trans isomerases (PPIases) in this plant, and their levels of expression across different tissues were reported.

Chapter 3 describes the complete suite of cyclotides obtained in the Indian variety of *C. ternatea* plant, characterized by MALDI-TOF, high-resolution LC-MS and MS/MS techniques. In the first part of the chapter, tissue specific proteomic variation of cyclotides from pods, stems, leaves, flowers and roots is detailed. An assortment of cyclotides were detected and variations were observed across the different tissues tested. Few mass signals observed in one type of tissue were not observed in the other, the relative expression levels were different. Extensive LC-MS analysis also revealed that cyclotide adopt a few novel structural arrangements. Notable variations in MS/MS product ion distributions were observed in cyclotides belonging to the two structural subfamilies based on the number and positions of proline residues. For instance, Cter M (Möbius subfamily) displayed distinct b-/y- ion characteristics in the MS/MS spectra compared to cliotide T1 (Bracelet subfamily). Distinct fragmentation patterns determined by Xxx-Pro bond fragmentation of prototypical cyclotides was used as a diagnostic to rapidly identify and sequence novel cyclotides ctr pep 30 and ctr pep 43 from this plant. In the second part of this chapter, we

describe the extensive LC-MS analysis of *C. ternatea* extracts, which revealed the presence of few novel cyclo-oligomeric structural arrangements adopted by few cyclotide masses.

In **Chapter 4**, we introduce the update of DiSulphide DataBASE - DSDBASE2.0 (http://caps.ncbs.res.in/dsdbase2). DSDBASE is a database on disulfide bonds in proteins that provides information on native disulfides and those which are stereochemically possible between pairs of residues in a protein (Vinayagam et al., 2004). The database has been updated to include 153,944 PDB entries, 216,096 native and 20,153,850 modelled disulfide bond segments. It also provides a resource to user-friendly search for multiple disulfide bond containing loops, along with annotation of their function using Gene Ontologies (GO terms) and subcellular localization of the query. Additionally, it is now possible to obtain three-dimensional models of disulfide-rich small proteins using an independent algorithm - RANMOD that generates and examines random, but allowed backbone conformations to the query polypeptide. A curated database of disulphide cross-links of all known protein structures from the PDB would be a useful resource for future research related to disulfide-rich proteins. In the Protein Data Bank (PDB), a total of 47 structures of cyclotides have been reported, including a single crystal structure of a Möbius cyclotide and several NMR derived models of the Bracelet conformation. Due to the relative paucity of high-resolution structures, it it important to be able to successfully predict the three-dimensional conformation of cyclotides without relying too much on known cyclotide structures. Therefore, we used RANMOD, a random conformation generation algorithm (Sowdhamini et al., 1993), to model cyclic conformation of a prototypical cyclotide sequence and reported the model it in this chapter.

Chapter 5 discusses the therapeutic potential of cyclotides extracted from *C. ternatea* plant. *C. ternatea* is widely used in traditional Ayurvedic medicine as an antioxidant and enhancer of cognitive functions (Jain et al., 2003; Rai et al., 2002). Cyclotides' anti-neurodegenerative properties were recently described wherein cyclotides extracted from *Psychotria solitudinum* inhibited the human prolyl oligopeptidase protein, a known target for the treatment of cognitive deficits in several psychiatric and neurodegenerative diseases. Similarly, in Alzheimer's disease (AD), neurotoxicity is often associated with the aggregation of Aβ peptide monomers into fibrils in the pathogenesis and therefore, identification of inhibitors of Aβ is critical for developing therapeutics against AD (Xu et al., 2005). In the first part of this chapter, the anti-Aβ *in vivo* activity of cyclotides using transgenic *Caenorhabditis elegans* (CL4176, CL2355 and CL2006 strains) that exhibit pathological behaviors associated with Aβ is described (Kalmankar et al., 2021). Cyclotide-rich fractions (CRF) from different plant tissues were observed to delay Aβ-induced

paralysis in transgenic CL4176 strain expressing human muscle-specific $A\beta_{1-42}$ gene. Additionally, CRF significantly improved Aβ-induced defects in chemotaxis behavior in transgenic strain CL2355, expressing the human $A\beta_{1-42}$ in neuronal cells. Furthermore, Aβ deposits in the transgenic strain CL2006 treated with CRF were significantly reduced. The reactive oxygen species (ROS) assay showed that this protection is likely mediated reduction of intracellular oxidative stress. In the second part of the chapter, we discuss the usage of molecular docking analyses and molecular dynamics (MD) simulations to establish the molecular and atomic level interactions between Aβ protofibril structures and cyclotide. We describe how cyclotides stably bind to Aβ molecules and destabilize the Aβ fibril by inducing conformational changes in the β-sheet rich structural regions. These results show that cyclotides from *C. ternatea* are novel leads for inhibition of Aβ aggregation and potent antioxidants, and could be a source of a new pharmacophore against neurodegenerative diseases.

In conclusion (**Chapter 6**), the findings reported this thesis provide insights into the diversity of cyclotides across multiple tissues from butterfly pea plant and highlights their future applications as therapeutic lead molecules. The principal studies detailed in this thesis broadens the knowledge of cyclotide assortment in the medicinally important plant using combined approaches of transcriptomics and proteomics. Furthermore, as there are few high-resolution structures of cyclotides known till date, we made use of the RANMOD algorithm in conjunction with MODIP disulfide modelling algorithm to generate linear precursors with three disulfide bridges, and performed an energy minimization routine to derive cyclic models of specific cyclotide sequences. Lastly, the concluding findings in this thesis provides information on the development of such disulfide-rich cyclic peptides as templates in drug design in target-based drug discovery, especially for neurodegenerative disorders such as the Alzheimer's disease presented as case study.

REFERENCES OF SYNOPSIS

Daly, N. L., Rosengren, K. J., and Craik, D. J. (2009). Discovery, structure and biological activities of cyclotides. *Adv. Drug Deliv. Rev.* 61, 918–930. doi:10.1016/j.addr.2009.05.003.

Jain, N. N., Ohal, C. C., Shroff, S. K., Bhutada, R. H., Somani, R. S., Kasture, V. S., et al. (2003). Clitoria ternatea and the CNS. *Pharmacol. Biochem. Behav.* 75, 529–536.

Jennings, C., West, J., Waine, C., Craik, D., and Anderson, M. (2001). Biosynthesis and insecticidal properties of plant cyclotides: The cyclic knotted proteins from Oldenlandia affinis. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10614–10619. doi:10.1073/pnas.191366898.

Kalmankar, N. V., Hari, H., Sowdhamini, R., and Venkatesan, R. (2021). Disulfide-Rich Cyclic Peptides from Clitoria ternatea Protect against β-Amyloid Toxicity and Oxidative Stress in Transgenic Caenorhabditis elegans. *J. Med. Chem.* 64, 7422–7433.

Kalmankar, N. V., Venkatesan, R., Balaram, P., and Sowdhamini, R. (2020). Transcriptomic profiling of the medicinal plant Clitoria ternatea: identification of potential genes in cyclotide biosynthesis. *Sci. Rep.* 10. doi:10.1038/s41598-020-69452-7.

Oguis, G. K., Gilding, E. K., Jackson, M. A., and Craik, D. J. (2019). Butterfly pea (Clitoria ternatea), a cyclotide-bearing plant with applications in agriculture and medicine. *Front. Plant Sci.* 10. doi:10.3389/fpls.2019.00645.

Poth, A. G., Colgrave, M. L., Lyons, R. E., Dalya, N. L., and Craik, D. J. (2011). Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. U. S. A.* 108, 10127–10132. doi:10.1073/pnas.1103660108.

Rai, K. S., Murthy, K. D., Karanth, K. S., Nalini, K., Rao, M. S., and Srinivasan, K. K. (2002). Clitoria ternatea root extract enhances acetylcholine content in rat hippocampus. *Fitoterapia* 73, 685–689. doi:10.1016/S0367-326X(02)00249-6.

Rosengren, K. J., Daly, N. L., Plan, M. R., Waine, C., and Craik, D. J. (2003). Twists, knots, and rings in proteins: Structural definition of the cyclotide framework. *J. Biol. Chem.* 278, 8606–8616. doi:10.1074/jbc.M211147200.

Sowdhamini, R., Ramakrishnan, C., and Balaram, P. (1993). Modelling multiple disulphide loop containing polypeptides by random conformation generation. The test cases of α-conotoxin gi and edothelin I. *Protein Eng. Des. Sel.* 6, 873–882. doi:10.1093/protein/6.8.873.

Vinayagam, A., Pugalenthi, G., Rajesh, R., and Sowdhamini, R. (2004). DSDBASE: A consortium of native and modelled disulphide bonds in proteins. *Nucleic Acids Res.* doi:10.1093/nar/gkh026.

Wang, C. K., Colgrave, M. L., Ireland, D. C., Kaas, Q., and Craik, D. J. (2009a). Despite a conserved cystine knot motif, different cyclotides have different membrane binding modes. *Biophys. J.* 97, 1471–1481. doi:10.1016/j.bpj.2009.06.032.

Wang, C. K., Shu-Hong, H., Martin, J. L., Sjögren, T., Hajdu, J., Bohlin, L., et al. (2009b). Combined x-ray and NMR analysis of the stability of the cyclotide cystine knot fold that underpins its insecticidal activity and potential use as a drug scaffold. *J. Biol. Chem.* 284, 10672–10683. doi:10.1074/jbc.M900021200.

Xu, Y., Shen, J., Luo, X., Zhu, W., Chen, K., Ma, J., et al. (2005). Conformational transition of amyloid β-peptide. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5403–5407.