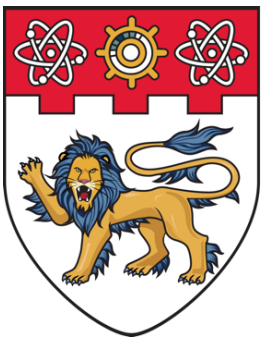


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# Discovery, characterization, and applications of cysteine-rich peptides from medicinal plants

**HUANG JIAYI**

**SCHOOL OF BIOLOGICAL SCIENCES**

**2019**

# Discovery, characterization, and applications of cysteine-rich peptides from medicinal plants

**HUANG JIAYI**

**SCHOOL OF BIOLOGICAL SCIENCES**

A thesis submitted to the Nanyang Technological University in partial fulfillment of the requirement for the degree of Doctor of Philosophy

2019

# Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

....Aug 13th 2019 ..... .....

Date Huang Jiayi

## Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

.....Aug 13th 22019 ..... .....

Date James P. Tam

# Authorship Attribution Statement

This thesis contains material from [3] paper(s) published in the following peer-reviewed journal(s)/ from papers accepted at conferences where I was the first author.

Chapter 3 is published as Huang, J., Wong, K. H., Tay, S. V., How, A. & Tam, J. P., Cysteine-rich peptide fingerprinting as a general method for herbal analysis to differentiate Radix Astragali and Radix Hedysarum. *Frontiers in Plant Science* **10**, 973 (2017). DOI: 10.3389/fpls.2019.00973.

The contributions of the co-authors are as follows:

- Prof James P Tam provided the initial project direction and edited the manuscript drafts.
- I wrote the drafts of the manuscript. The manuscript was revised together with Dr. Wong.
- I performed the sample collection, extraction, MALDI-TOF MS analysis, UPLC analysis, *de novo* sequencing, and multivariate analyses for the RA and RH samples.
- Dr. Wong provided guidance and help in analyzing the data using MATLAB.
- Dr. How conducted the CRP fingerprinting analysis for 100 herbs and herbal products.
- Ms. Tay assisted in the extraction and *de novo* sequencing of the peptides in RA and RH samples.

Chapter 4 is published as Huang, J., Wong, K. H., Tay, S. V., Serra, A., Sze, S. K., & Tam, J. P. Astratides: Insulin-Modulating, Insecticidal, and Antifungal Cysteine-Rich Peptides from *Astragalus membranaceus*. *Journal of natural products* **82**, 194-204 (2019). DOI: 10.1021/acs.jnatprod.8b00521.

The contributions of the co-authors are as follows:

- Prof James P Tam provided the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised together by Dr Wong.

- I prepared the manuscript drafts. The manuscript was revised together by Dr Wong.
- I performed the extraction, purification of the peptides, reduction and alkylation, sequence comparison, biosynthesis analysis, phylogenetic analysis, anti-fungal assays, insecticidal assays, microscopic analysis and insulin-modulating assays at the School of Biological Sciences. I also analyzed the data.
- Dr. Wong provided help in the biosynthesis analysis.
- Ms. Tay helped in the extraction of peptides and anti-fungal assays.
- Dr. Serra assisted in the LC-MS sequencing of the peptides and Assoc Prof Sze provided the instrument and guidance.

Reprinted (adapted) with permission from ([1]). Copyright (2019) American Chemical Society.

Chapter 5 is published as Huang J., Wong K. H., Tam J. P. (2019). Coffeetides: Iron-chelating Cysteine-rich Peptides from Coffee Waste, 2019 American Peptide Symposium, Monterey, USA.

The contributions of the co-authors are as follows:

- Prof James P. Tam provided the initial project direction and edited the abstract and poster drafts.
- I prepared the abstract draft and poster drafts. The drafts were revised together by Dr Wong.
- I performed the extraction, purification of the peptides, reduction and alkylation, sequence comparison, biosynthesis analysis, phylogenetic analysis, ion-binding assays at the School of Biological Sciences. I also analyzed the data.
- Dr. Wong provided help in the biosynthesis analysis.

...Aug 13th 2019..... .....

Date Huang Jiayi

# Acknowledgments

This four-year journey has shaped me into a responsible, determined person with great logical thinking ability. The completion of this thesis could not be possible without the people that motivated, helped, and challenged me along the way.

First and foremost, I would like to express my sincere gratitude to my supervisor Professor James P. Tam for the opportunity to join as a Ph.D. student in his laboratory under his guidance. For the past four years, his constant and patient guidance kept me motivated to face the tough challenges and fulfill all the research works. In addition, his extensive experiences and enthusiasm towards research have inspired me to develop myself beyond my expectations. The learning opportunities he has provided are indeed rich sources for shaping me into a well-prepared Ph.D. student. I would also like to thank my co-supervisor Professor Wang Rong, and my thesis advisory committee, Assoc. Prof Newman Sze Siu Kuan, and Assoc. Prof Liu Chuanfa, for their invaluable guidance and advice throughout my Ph.D.

Next, I would like to express my gratitude to my lab members for their motivation and support during the four years. In particular, I would like to thank my mentor Dr. Wong Ka Ho for his endless guidance, care, and support throughout this arduous journey. His friendliness and expertise allowed me to learn from him substantially and I am grateful for having him as a great friend and working partner. I also wish to thank Professor Yang Daiwen and Dr. Fan Jin Song for helping me with the NMR-titration experiments. I would also like to thank Dr. Hemu Xinya for teaching me the chemical synthesis of peptides when I just began my Ph.D. journey, allowing me to quickly get up to speed in my research. I also thank Dr. Nguyen Kien Truc Giang for providing the research materials for me from Vietnam and guidance on the initial stage of my projects. I also appreciate Ms. Stephanie Victoria Tay for planning, performing experiments and writing manuscripts together with me. I am thankful to have wonderful friends in the lab, Dr. Tan Weiliang, Dr. Kini Shruthi Gopalkrishna, Dr. Geeta Kumari, Clarene Chan, Lim Wei Qin, Tan Fan, Shuan Tan. Our daily lunch breaks, dinners, and outings allow us to have the courage to overcome the difficulties and achieve many milestones together. I am also thankful to have gone through the journey and obtained help from Dr. Janet To, Dr. Bamaprasad Dutta, Dr. Aida S. Maqueda, Dr. Xiao Tianshu, Ms. Yee Lee Choo and many others in the lab. I am immensely grateful to have the many URECA and FYP students who worked with me and learn along together.

To survive the rigorous journey for obtaining a Ph.D. degree, a strong support system is essential. I felt lucky to have met so many wonderful people that I can call as friends during

these four years. In particular, I would like to thanks may best friend in SBS, Guo Xue, who has gone through the same journey as me. We spent times together to do research, travel, chat, hang out, plan futures and shared all the happiness and difficulties with each other. Her friendliness and humor always make me feel happy and without her accompany, I would not be able to survive the long journey. Also, I am pleasant for known Wu Dan, who has been my roommate and a good friend for these four years. Her encouragement and suggestions always cheered me up when I am in difficulties. My time in Singapore has genuinely been a pleasant one thanks to who have spent the precious four years together with me. My Ph.D. would not have been possible without the encouragement from all my friends back home in China. I especially thank Chen Zilin and Zhou Jiali for encouraging me to take up the Ph.D. opportunity and constantly motivating me despite the distance. Also, I would like to thank Yang Ming, who has accompanied me for more than 10 years. Her brilliant mind, enthusiasm towards life and the courage to take up challenges have impressed me a lot and make me be determined to accomplish the Ph.D. study. Her constant support makes me feel optimistic about the future. Besides, I would like to thank Zheng Tong, who has spent one year with me in Singapore and life with her is extremely happy. Although she went back to China afterward, we still kept in touch and had fun together. Her support for me is a solid foundation for me to finish my PhD study. Moreover, I appreciate the help from Ma Hanhan, who possesses an optimistic attitude towards life and always stands by my side despite the distance. Her logical thinking and words always make me think through my future. Also, I am grateful for gaining support from Xiong Sipin, Qiu Yingshan, He Ruixiang, Liang Gengyi, Deng Zifeng, Liu Zitian, Liu Junjie and many other friends through these four years.

My boyfriend Ding Bosheng has been my greatest supporter, listener, motivator throughout this journey. Especially in the last two years, when I felt doubt about myself that I cannot produce papers or graduate on time, he always encouraged me and gave me advice. I thank him for his unwavering faith in me and his patience in dealing with me during the stressful phase of my Ph.D. He is a lover, friend, mentor, life-sharing person to me, and I feel happy whenever we travel, study, talk, and plan for the future. I am glad to have him in this Ph.D. journey and will be forever grateful to him for his unconditional love for me.

There are no words to express how thankful I am to have such great parents who are very supportive of me. Their love, blessings, and encouragement have helped me carry forward my research. My parents provided me guidance, a safe home, and unconditional support so that I can freely pursue my dreams and achieve what I have achieved today. They felt proud of my every tiny success and consoled me through the failures during this journey. Being surrounded

by such unconditional love, I was able to grow up into a better person. Thus, I hope this thesis makes them proud, and I have lived up to their expectations.

# Table of Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| *A. membranaceus* | *Astragalus membranaceus* (Fisch.) Bje |
| *A. alternate* | *Alternaria alternata* |
| ACN | Acetonitrile |
| AMP | Antimicrobial peptide |
| BLAST | Basic Local Alignment Search Tool |
| Boc | Ter-butoxycarbonyl |
| *C. arabica* | *Coffea arabica* |
| *C. canephora* | *Coffea canephora* |
| *C.liberica* | *Coffea liberica* |
| *C.rosmosa* | *Coffea rosmosa* |
| *C. lunata* | *Curvularia lunata* |
| CART | Classification and regression tree |
| cDNA | Complementary deoxyribonucleic acid |
| CKAI | Cystine-knot α-amylase inhibitors |
| COSY | Correlation spectroscopy |
| CRP | Cysteine-rich peptide |
| DCM | Dichloromethane |
| DIC | *N,N*-diisopropylcarbodiimide |
| DIEA | *N,N*-diisopropylethylamine |
| DMEM | Dulbecco's modified Eagle's medium |
| DMF | *N, N*-dimethylformaide |
| DMSO | Dimethylsulfoxide |
| DTT | Dithiothreitol |
| ER | Endoplasmic reticulum |

| | |
|---|---|
| EtOH | Ethanol |
| EST | Expressed sequence tag |
| FA | Formic acid |
| Fmoc | 9-fluroenylmethyloxycarbonyl |
| *F. oxysporum* | *Fusarium oxysporum* |
| HATU | (1-[Bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid -hexafluorophosphate) |
| HBTU | *N,N,N′,N′*-Tetramethyl-O-(1H-benzotriazol-1-yl)uronium hexafluorophosphate |
| HCA | Hierarchical clustering |
| HCl | Hydrochloric acid |
| HPLC | High performance liquid chromatography |
| HOBt | *N*-hydroxybenzotriazole |
| IAA | Iodoacetamide |
| $IC_{50}$ | Half-maximal inhibitory concentration |
| kDa | Kilo Dalton |
| KNN | K-nearest neighbors |
| MALDI | Matrix Assisted Laser Desorption/Ionization |
| MeOH | Methanol |
| MS | Mass spectrometry |
| MS/MS | Tandem Mass Spectrometry |
| NaCl | Sodium chloride |
| NEM | N-ethylmaleimide |
| NMR | Nuclear magnetic resonance |
| NOESY | Nuclear Overhauser effect spectroscopy |
| PBS | Phosphate buffered saline |

| | |
|---|---|
| PCA | Principal component analysis |
| PCR | Polymerase chain reaction |
| PLS-DA | Partial least square-discriminant analysis |
| RA | Radix Astragali |
| RH | Radix Hedysarum |
| *R. solani* | *Rhizoctonia solani* |
| RNA | Ribonucleic acid |
| RP | Reversed phase |
| SCX | Strong Cation exchange |
| SIMCA | Soft independent modelling of class analogy |
| SVM-DA | Support vector machine-discriminant analysis |
| SPPS | Solid phase peptide synthesis |
| tBLASTn | translated nucleotide BLAST |
| TCEP | Tris(2-carboxyethyl)phosphine |
| TCM | Traditional Chinese Medicine |
| TFA | Trifluoroacetic acid |
| TIS | Triisopropylsilane |
| TOF | Time of Flight |
| UPLC | Ultra-performance liquid chromatography |

# Abstract

Medicinal plants showed great importance in managing and treating human diseases. Currently, plant-derived small-molecule metabolites with molecular weight (M.W.) <1 kDa are major active components that account for approximately 46% of all the clinically approved drugs. Another major family of pharmaceuticals that have been extensively studied and used clinically is proteins with M.W. >10 kDa. However, few peptidyl products which occupy the chemical space between metabolites and proteins are clinically approved as drugs. Typically, peptides are susceptible to harsh conditions and have poor bioavailability. Disulfide-constrained cysteine-rich peptides (CRPs) are a family of molecules with highly compact structures, in a particular range of M.W. from 2 – 6 kDa. These disulfide bridges confer them the stability against thermal, acidic and enzymatic degradation. Currently, these naturally-occurring disulfide-constrained peptides are highly under-explored in medicinal plants.

Another vital issue for herbal medicines is the misidentification of plant species and the presence of adulterants. The traditional authentication method using chromatographic fingerprinting is precise, sensitive, and reproducible. However, laborious sample preparation, relatively long analytical run-times, and the large volume of organic solvents consumption in HPLC hinders its application as a high-throughput screening technique. Hence, the development of a rapid and accurate quality control method is urgently needed.

My thesis aims to discover and characterize CRPs from medicinal plants and to apply them in the authentication of herbal medicines. A general and rapid method, which employs CRPs as unique chemical markers for the authentication of herbal medicines, was described in this thesis. This CRP fingerprinting method produces consistent results for herbal authentication regardless of the morphology, chemical composition, and origins of the plant species. The differentiation of two similar species, Radix Astragali and Radix Hedysarum, was used as an example to validate the method. Coupling with multivariate analyses, the study showed that CRP fingerprinting is fast, and the classification accuracy is comparable to that of the conventional authentication method using UPLC. To further understand the usefulness and functions of CRPs, clusters of CRPs discovered from medicinal plants and important crops were studied. They include CRPs from *Coffea canephora* and *Coffea liberica*, which are plants with medicinal values used to produce the second-largest commodity, coffee drinks. Proteomic and transcriptomic analyses showed that coffeetides identified from the *Coffea* species are non-chitin-binding hevein-like peptides. Another group of CRPs was identified from the roots of *Astragalus membranaceus* (Fisch.) Bje, which is a traditional Chinese medicine for improving

overall vitality and treating diabetes. Two different types of CRPs were identified in this plant and designated as $\alpha$- and $\beta$-astratides. NMR spectroscopy, bioinformatics analysis, and functional bioassays were used to determine the structure, evolutionary relationship, and functions of all these CRPs. Taken together, my thesis expanded the existing library of CRPs and explored their potential for drug design and their usefulness as fingerprints for the authentication of herbal medicine.

# Chapter 1 Introduction

## 1.1. Underexplored bioactive peptides in natural products for drug discovery

Natural products have made a significant contribution to the development of new drugs. Over the past 20 years, more than 30% of the US Food and Drug Administration (FDA)-approved therapeutics are natural product-derived drugs. Additionally, approximately half of the small-molecule therapeutics were originated from natural products, or natural product-derived mimetics [2]. For example, the use of herbal medicine dates back from prehistoric times and has been mainstream in the healthcare approach due to the presence of their bioactive components. These medicinal plants are well-studied natural products that have been shown to possess high medicinal values and nowadays still considered as a reliable source for developing drug leads [3]. Currently, the widely utilized drugs developed from the bioactive compounds can be classified into two groups, small-molecule therapeutics which occupy the chemical space <500 Da and larger biologics that are >10,000 Da. Small-molecule drugs have been extensively identified via screening based on ligands, mechanism, or receptors [4]. These molecules have a wide range of bioactivities such as anti-cancer, anti-virus, and anti-bacterial activities and are widely favored by pharmaceutical companies due to their oral bioavailability, metabolic stability and low cost for production [5].

In the late 20$^{th}$ century, new therapeutics that lie on the opposite end of the size spectrum of small molecules began to emerge. This change may be due to the development of recombinant protein expression system, better molecular biology, and protein purification techniques. These protein therapeutics showed great potency and high selectivity that may lead to less off-target side effects, which is the limitation of small-molecule drugs. For example, insulin, growth factors, and engineered antibiotics are proteinaceous molecules and are termed as 'biologics.' However, these protein-based drugs which have large molecular size are usually not suitable for oral administration and thus require intranasal delivery or injection [4, 6].

With the advances in genome sequencing at the beginning of the 21$^{st}$ century, it has been predicted that because of the occurrence of numerous new drug targets, there will be a significant advance in the drug development field. It was true that a large amount of gene expression data has been generated, but not all have been translated into validated drug targets [4]. It is speculated that protein-protein interactions are involved in the new targets obtained from sequencing, featuring interaction sites with shallow grooves spanning across large surface areas [4]. These targets are not tractable for small molecules and inaccessible for proteins that are not membrane permeable but are suitable for peptides [7].

Until now, studies have reported approximately 7000 naturally occurring peptides. They are reported to possess multiple functional roles such as hormones, anti-infective molecules, growth factors, and ion channel ligands [8, 9]. Generally, peptides can act as efficacious signaling molecules which can trigger intracellular activity by binding to specific cell surface receptors. Compared with small-molecule drugs, peptide therapeutics showed higher safety, higher selectivity, a broader range of targets, and biological diversity. Meanwhile, when compared to protein-based biologics, peptide therapeutics are less complex and cost less in production [6].

For the past decades, peptides have gained increasing applications in medicine and biotechnology. Currently, >60 US FDA-approved peptide therapeutics have been sold in the market, and nearly 140 peptide therapeutics have been used clinically [6]. In addition, the peptide therapeutics showed high commercial values. From the year 2011 to 2018, there was a US$10 billion increase in the global peptide drug market [10]. Generally, metabolic diseases and oncology are the primary disease fields that have been addressed by peptide therapeutics.

Peptides are suitable drug candidates for site-specific modification based on their potency, specificity, and safety. With the emerging peptide techniques, the future development of peptide drugs will become promising. However, the low bioavailability, cell permeability, and low metabolic stability remain major concerns and hinder their wide application as drugs.

### 1.2. Cysteine-rich peptides in plants

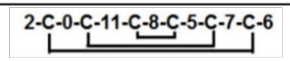
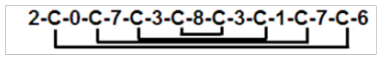
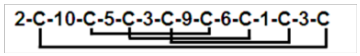
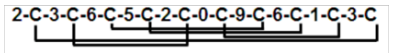
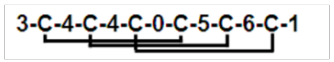
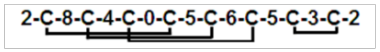
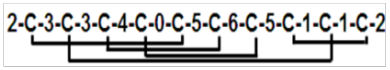
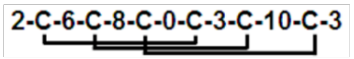
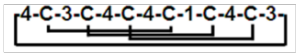
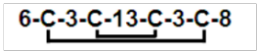
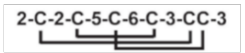
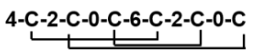
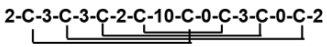
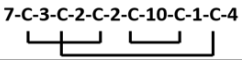
Plants have developed multiple defense mechanisms to counter harsh natural conditions, which include drought, dryness, cold, wounding, heavy metals, air pollution, and attacks by pathogens [11]. In response to the infections by the pathogens, plants express a set of genes for the resistance. Generally, the liberation of secondary metabolites such as tannins, pathogenesis-related proteins, and enzyme inhibitors will contribute to their host-defense mechanism [11-14], which usually contains interaction with the receptors on the cell surface. Thus, the pathogen's membrane will be disrupted and lead to their death [15]. Recent reviews have shown that more than 17 plant families have various defense-related functions which include antifungal, antibacterial, antiviral and protease inhibitory activities, part of which are associated with the peptides present in the plants [12].

Antimicrobial peptides (AMPs) are the unique peptides that play a host defense role against pathogens and insects from multiple biological sources. They are widely present in different molecular forms such as linear peptides, polycyclic peptides, circular peptides, and cyclotides [16, 17]. In plants, many of them are cysteine-rich, forming disulfide bonds, and hence

contribute to the solid structure and resistance against harsh conditions [12]. These cysteine-rich peptides (CRPs) are found to be involved in plant physiology, including cell signaling, reproduction, and defense. However, genomic data mining and sequence analysis based on cysteine motifs showed that CRPs are under-explored [18]. In plants, CRPs may account for 3% gene products, and the expression of these CRPs is induced and often tissue-specific [12]. In this work, we are interested in plant CRPs ranging from 2 to 6 kDa.

Evolutionarily, plant CRPs are hypervariable in their sequences, being encased in a particular scaffold, which enables the classification of CRP families that describes their molecular diversity. Table 1.1 shows the major plant CRP families such as thionins, defensins, hevein-like peptides, and knottin-type peptides.

**Table 2.1.1. Plant CRP families.** The consensus sequence and cysteine connectivity of each family are summarized from PhytAMP database [19].

| CRP family | Disulfide number | Representative member | | | Structural motif |
|---|---|---|---|---|---|
| | | Name | No. of AA | Disulfide motif | |
| 6C-Thionin | 3 | Crambin | 46 |   | Gamma (Γ) fold β1-α1-α2-β2-coil motif |
| 8C-Thionin | 4 | β-Purothionin | 45 |   | |
| 8C-Defensin | 4 | NaD1 | 47 |   | CSαβ motif β1-coil-α-β2-β3 |
| 10C-Defensin | 5 | PhD1 | 47 |   | |
| 6C-Hevein | 3 | Ac-AMP1 | 29 |   | Gly & Cys rich Central β strands & (short helical) side coils |
| 8C-Hevein | 4 | Hevein | 43 |   | |
| 10C-Hevein | 5 | EAFP1 | 41 |   | |
| Knottin | 3 | PAFP-S | 38 |   | Cystine knot Short β strand & coil |
| Cyclic Knottin | 3 | Kalata B1 | 29 |   | |
| α-Hairpinin | 2 | Ec-AMP1 | 37 |   | α1-turn-α2 |
| Jasmintide | 3 | jS1 | 27 |   | |
| *β*-gingkotide | 3 | gB1 | 20 |   | |
| Lybatide | 4 | Lyba2 | 33 |   | Cystine-stabilized αT |
| Potentide | 3 | pA3 | 35 |   | |

### 1.3. Classification of CRPs based on the cysteine framework

CRPs contain a high content of disulfide bonds, which is an important feature for maintaining their metabolic stability. The cysteine framework of CRPs refers to two parts, the spacing, and arrangement of cysteine residues in the sequences and the disulfide connectivity. The different combination of these two factors may generate an enormous amount of sequence variability and structural diversity. CRPs ranging from 2 to 6 kDa with six to ten cysteine residues may generate possible 176 combinations on disulfide connectivities. However, previous studies showed that the major types of disulfide connectivity could be categorized into two different clusters: the type I cystine-knot connectivity and the type II symmetric disulfide connectivity (Figure 1.1).

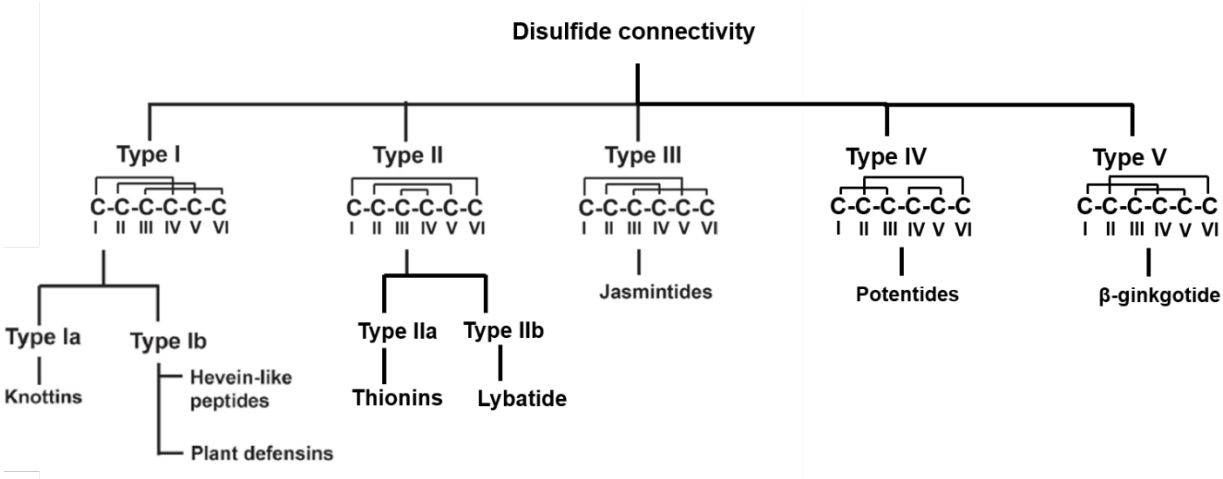



**Figure 2.1.1. Classification of plant CRPs based on disulfide connectivity.** Type I shows the cystine-knot connectivity, and Type II showed the knottin-type connectivity. The figure was modified from Tan W. L. (2018). (Doctoral dissertation, Nanyang Technological University, Singapore). Retrieved from https://repository.ntu.edu.sg/handle/10356/73452)

With disulfide connectivity of Cys I to Cys IV, Cys II to Cys V and Cys III to Cys VI, cysteine-knot disulfide connectivity is the most common type of disulfide linkages observed in plant CRPs. Type Ia represents the CRPs containing a cystine-knot structure, which is formed by three disulfide bonds, such as knottins and 6C-hevein-like peptides [20]. This structural motif is characterized by a disulfide bond which interconnects the backbone, penetrating the ring formed by the other two disulfide bonds (Figure 1.2). Usually, the formation of the additional one or two disulfide bonds does not disturb the knot-topology. Hence the class of CRPs with

additional disulfide bonds such as defensins and 8C-hevein-like peptides [21] are termed as type Ib.

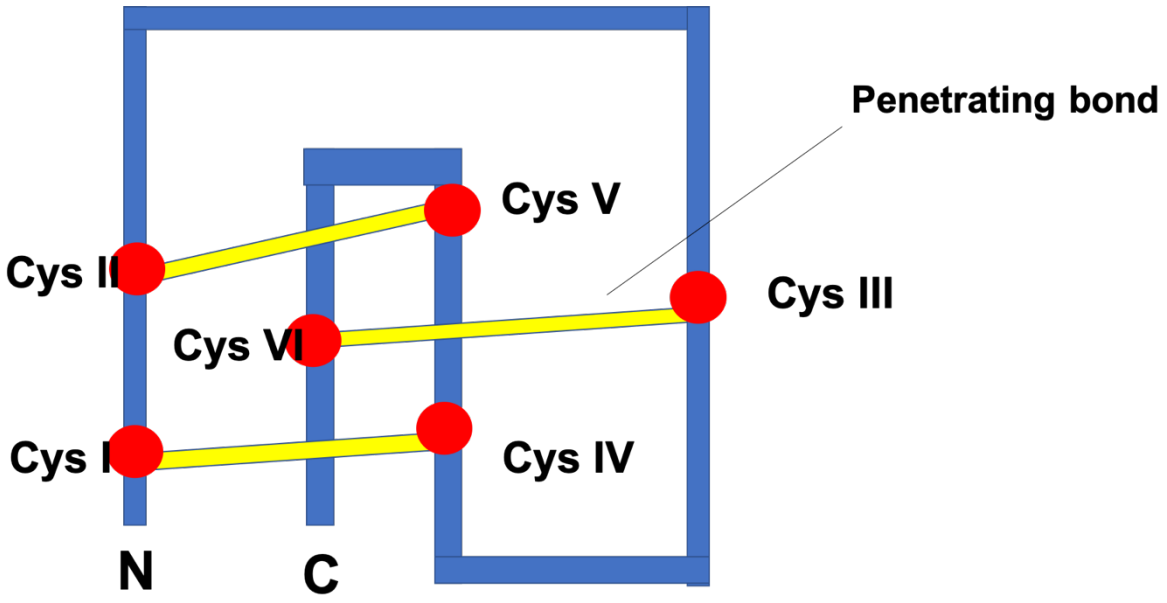



**Figure 1.2. Schematic representation of cystine-knot connectivity.** Cysteines are shown as small red circles with Cys number I to VI from N- to C-terminus. Yellow lines are used to indicate disulfide bonds, and the blue line is used to represent the backbone.

In contrast, the type II disulfide connectivity is in a symmetric arrangement and displayed as Cys I-Cys VI, Cys II-Cys V, and Cys III-Cys IV. This cystine arrangement can often be observed in plant CRPs such as 6C-thionins and α-hairpinis [22]. In 8C-thionins, the additional disulfide bond follows the same symmetric pattern of connectivity. This symmetric type of connectivity was termed as type IIa. However, our laboratory has recently discovered another disulfide connectivity pattern of 8C-CRP, which is modified based on the symmetric connectivity. The lybatides, adopting symmetric disulfide connectivity as CysII–VIII, CysIII–VII and CysIV–V, however, contain an additional disulfide bond of Cys I-VI, were grouped as type IIb [21].

Additionally, our laboratory identified other three new disulfide patterns that have not been reported for plant CRPs, which are jasmintides from *Jasminum sambac*, potentides from *Potentilla anserine* and *β*-ginkgotide from *Ginkgo biloba* [23-25]. Jasmintides display a CysI-CysV, CysII-CysIV and CysIII-VI disulfide connectivity which is different from the type I and type II disulfide connectivity and is thus termed type III. Differently, potentides contain a

disulfide linkage of Cys I- Cys III, Cys II-Cys VI, and Cys IV-Cys V, which is different from all the previous types of connectivity and thus termed type IV. Similarly, the different disulfide linkage of Cys I– Cys IV, Cys II– Cys VI and Cys III– Cys V, which found in *β*-ginkgotides was termed as type V.

Cysteine spacing in CRPs is another important feature for their classification, which determines the numbers and length of intercysteinyl loops. These loops represent the peptide backbone segments that are divided by the successive cysteine residues. For example, 6C-CRPs usually possess a cysteine spacing of C-C-CC-C-C, which is well represented by the 6C-hevein-like peptides (6C-HLPs) and cystine-knot α-amylase inhibitors (CKAIs). Another typical cysteine spacing of 6C-CRPs is C-C-C-C-C-C, which is well represented by carboxypeptidase inhibitors [23]. Cyclotides, although containing the same cysteine spacing as carboxypeptidase inhibitors, their head-to-tail cyclization makes them 6-loop CRPs. In addition, C-CC-C-CC, which is a unique type of cysteine motif with only three intercysteinyl loops, is found in 6C-CRPs and can be represented by *β*-ginkgotides. In a great number of CRP families, the presence of a -CC- motif is commonly observed. For example, both CKAIs and HLPs contain a -CC- motif in the middle position of the sequences [26, 27]. However, there are also CRPs containing a –CC- motif at the N-terminus, such as thionins [25].

Together, the different combination of disulfide connectivity and cysteine spacing may give rise to multiple structural folds of CRPs. When compared to the conotoxins, which contain a total of 27 different frameworks [28], plant CRPs showed much less reported cysteine framework, and therefore, they are highly underexplored (Table 1.2).

**Table 2.1.2. Conotoxin cysteine framework and plant CRP equivalent.**

| Framework | Cysteine pattern | #Cys | Connectivity | Plant CRPs with similar framework | Ref |
|---|---|---|---|---|---|
| 1 | CC-C-C | 4 | I-III, II-IV | | Gray,W.R. et al. (1981) |
| 2 | CCC-C-C-C | 6 | | | Ramilo,C. et al. (1992) |
| 3 | CC-C-C-CC | 6 | | | Sato,S. et al. (1983) |
| 4 | CC-C-C-C-C | 6 | I-V, II-III,IV-VI | 6C-Thionins | Fainzilber,M. et al. (1995) |
| 5 | CC-CC | 4 | I-III, II-IV | | Walker,C.S. et al. (1999) |
| 6/7 | C-C-CC-C-C | 6 | I-IV, II-V, III-VI | 6C-HLPs, CKAIs | Olivera,B.M. et al. (1984) |
| 8 | C-C-C-C-C-C-C-C-C-C | 10 | | | England,L.J. et al. (1998) |
| 9 | C-C-C-C-C-C | 6 | I-IV, II-V, III-VI | Cyclotides | Lirazan,M.B. et al. (2000) |
| 10 | CC-C.[PO]C | 4 | I-IV, II-III | | Balaji,R.A. et al. (2000) |
| 11 | C-C-CC-CC-C-C | 8 | I-IV, II-VI, III-VII, V-VIII | | Jimenez,E.C. et al. (2003) |
| 12 | C-C-C-C-CC-C-C | 8 | | | Brown,M.A. et al. (2005) |
| 13 | C-C-C-CC-C-C-C | 8 | | | Aguilar,M.B. et al. (2005) |
| 14 | C-C-C-C | 4 | I-III, II-IV | A-Hairpinins, 8C-HLPs | Moller,C. et al. (2005) |
| 15 | C-C-CC-C-C-C-C | 8 | | | Peng,C. et al. (2008) |
| 16 | C-C-CC | 4 | | | Pi,C. et al. (2006) |
| 17 | C-C-CC-C-CC-C | 8 | | | Yuan,D.D. et al. (2008) |
| 18 | C-C-CC-CC | 6 | | | Chen,J.S. et al. (1999) |
| 19 | C-C-C-CCC-C-C-C-C | 10 | | | Chen,P. et al. (2008) |
| 20 | C-CC-C-CC-C-C-C-C | 10 | | | Loughnan,M.L. et al. (2009) |
| 21 | CC-C-C-C-CC-C-C-C | 10 | | | Möller,C. and Marí,F. (2011) |
| 22 | C-C-C-C-C-C-C-C | 8 | | 8C-Defensins | Elliger,C.A. et al. (2011) |
| 23 | C-C-C-CC-C | 6 | | | Ye,M. et al. (2012) |
| 24 | C-CC-C | 4 | | | Luo,S. et al. (2013) |
| 25 | C-C-C-C-CC | 6 | | Jasmintides | Aguilar,M.B. et al. (2013) |
| 26 | C-C-C-C-CC-CC | 8 | | Lybatides | Bernáldez,J. et al. (2013) |
| 27 | C-CC-C-C-C | 6 | | | Kancherla,A.K. et al. (2015) |
| NR | CC-C-C-C-C-C-C | 8 | I-VIII, II-VII, III-VI, IV-V | 8C-Thionins | Mak AS and Jones BL (1976) |
| NR | C-C-C-C-CC-C-C-C-C | 10 | I-X, II-V, III-VII, IV-VIII, VI-IX | 10C-Defensins | Lay FT et al. (2003) |
| NR | C-C-C-CC-C-C-C-C-C | 10 | I-V, II-IX, III-VI, IV-VII, VIII-IX | EAFP, WAMP | Andreev Y A et al. (2012) |
| NR | C-C-CC-C-CC-C-C-C | 10 | I-IV, II-V, III-VI, VII-X, VIII-IX | Ee-CBP | Van den Bergh KP et al. (2004) |

The table was modified from Tan W. L. (2018). (Doctoral dissertation, Nanyang Technological University, Singapore). Retrieved from https://repository.ntu.edu.sg/handle/10356/73452)

## 1.4. Major CRP families

### 1.4.1. Thionins

Thionins are prototypic and cationic plant CRPs containing 45-48 amino acids with three to four disulfide bonds. Initially, this cluster of peptides was classified as plant toxins based on their toxicity against insects, fungi, bacteria, and plants [12]. The first discovered thionin was the antimicrobial peptide α-purothionin, which was discovered from the wheat endosperm [29, 30]. Upon its discovery, the following thionins were labeled with Greek alphabet in the order of their discovery. However, γ-thionins were later found to be a member of plant defensins due to the similar structures. Therefore, the classification of thionins is simplified as 8C-thionins,

referring to $\alpha/\beta$-thionins with eight Cys residues and 6C-thionins, which contain six Cys residues (Figure 1.3).

Thionins are broadly expressed in various parts of plants in monocots and dicots [31]. Their expression in plants was shown to be associated with the infection and invasion by microbes.[32]. Compared to other CRPs, thionins have relatively conserved sequences and a gamma (Γ) structural fold that is formed by the β1-α1-α2-β2-coil motif. The symmetric disulfide connectivity is conserved in 8C-thionins while the disulfide bond between Cys II-VII is absent in 6C-thionins, in which the end-to-tail disulfide bond makes their structure pseudocyclic. Specifically, the first pair of disulfide bond Cys I- Cys VIII links β1 strand to the C-terminal coil while the second pair of disulfide bond Cys II- Cys VII connects the β1 and β2 strands. The third and fourth pairs, Cys II- Cys VI and Cys IV- Cys V, usually stabilize the two $\alpha$-helices. In all, except for minor differences in the C-terminal coil region, this Γ fold is highly conserved in 6C- and 8C-thionins [33].

The toxicity of thionins against bacteria, fungi, plant, and animal cells was likely attributed to the membrane interaction with their hydrophobic or acidic residues [31]. Structure analysis showed that the highly conserved Lys1 and Tyr13 are the relevant residues responsible for the toxicity in thionins. The toxic effects were proposed to be a result of the cell membrane lysis, but it is still not been confirmed [34].

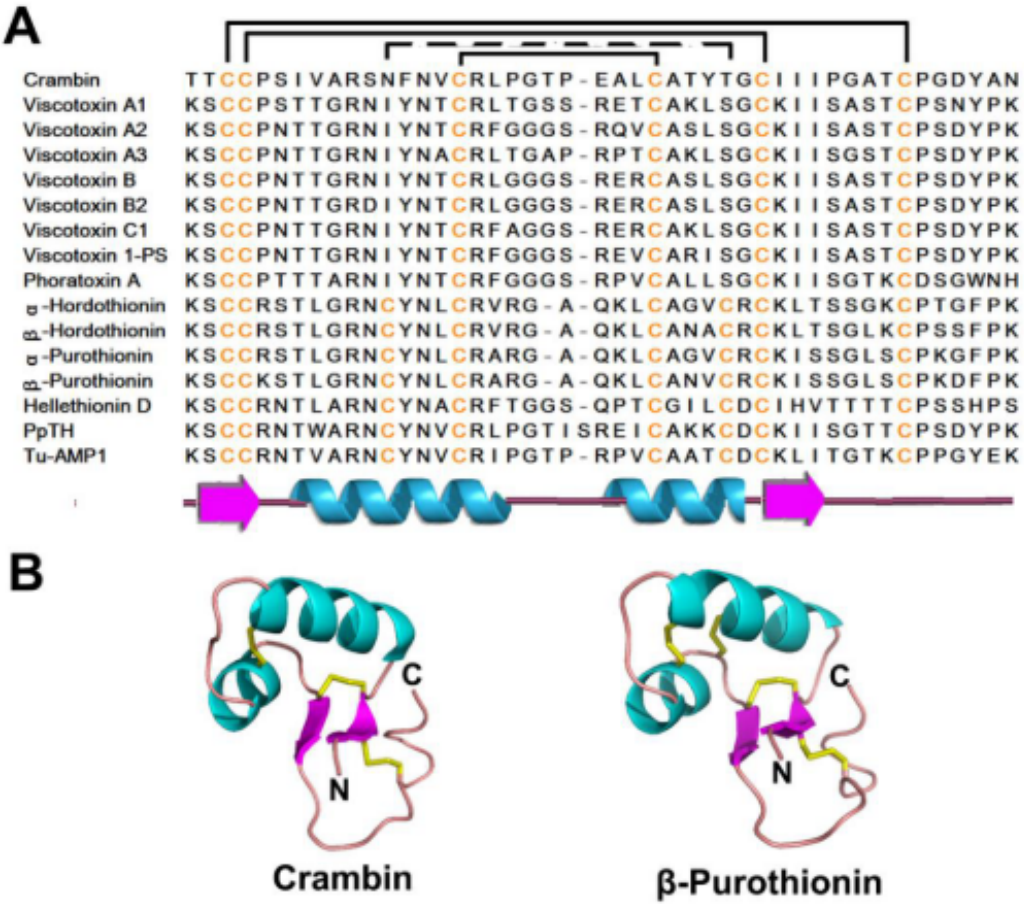



**Figure 1.3. (A) Sequence alignment and (B) structures of representative thionins.** The α-helix is shown in cyan, β-strand is represented in magenta, random coil is displayed in pink and disulfide bonds are displayed in yellow. The figure is adapted from reference [12].

### 1.4.2. Plant defensins

Plant defensins are widely distributed and likely to be the most abundant and best-known plant CRPs with membranolytic defense functions. They usually contain four to five disulfide bonds with 45-54 amino acids in length and were grouped under the thionin family based on limited sequence identity since their first discovery. However, they were later found to possess different structural motifs compared with thionins [31]. In 1995, this cluster of peptides was reclassified as plant defensins due to their similar sequences, structures, and functions as defensins isolated from insects and mammals [35]. It is reported that plant defensins are ubiquitously expressed in more than 100 members of plants such as wheat, radish, potato, sorghum, and soybean [36]. They have been mainly identified in seeds and roots [37] from different plant species but also in tissues including tubers [38], leaves [38], pods [39] and flowers [40].

Structurally, plant defensins are generally characterized by a tertiary structure that comprises of an α-helix being against by a triple-stranded antiparallel β-sheet. Based on the structural differences, defensins can be separated into two superfamilies, the *cis*-defensins, and *trans*-defensins. As illustrated by Figure 1.4, the CXC motif in *cis*-defensins causes disulfide bonds to link to the same cysteine-stabilized α-helix. In contrast, the –CC- motif in *trans*-defensins constrains the disulfide bonds to orient in the opposite direction and link to different secondary structure elements [41].

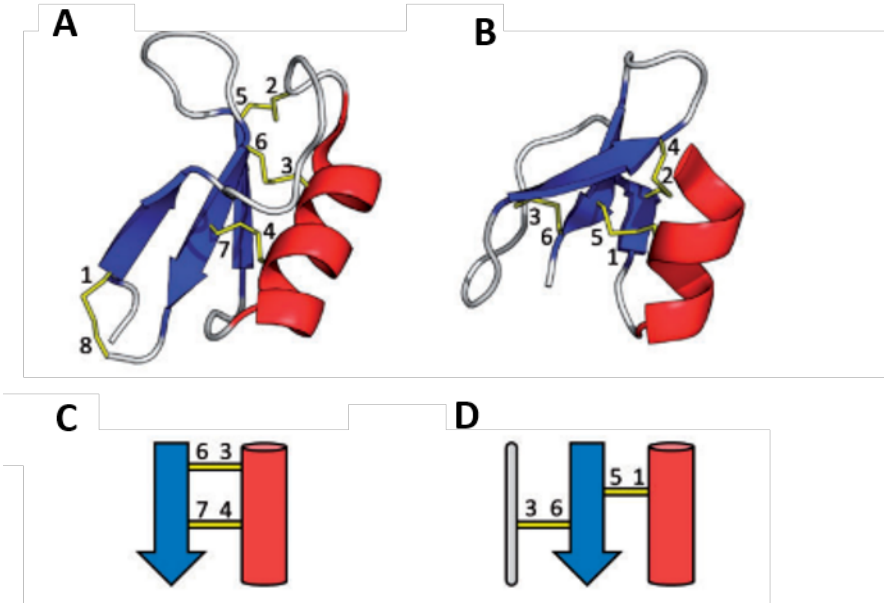



**Figure 1.4. Disulfide connectivity of two superfamilies of defensins.** (A) The *cis*-defensins are exemplified by NaD1, which is an 8C-defensin (PDB: 1MR4). (B) The *trans*-defensins are exemplified by the human β-defensin HBD1 (PDB: IJV). (C) The conserved *cis*-defensin disulfide connectivity which oriented disulfides from the C-terminal β-strand to bond to the same β-helix in the same direction. (D) The conserved *trans*-defensin disulfide connectivity that oriented disulfides from the C-terminal β-strand to link different secondary structure elements in the opposite direction. The figure is adapted from reference [41].

A Cys-stabilized αβ (CSαβ) motif, well represented by a β1-coil-α-β2-β3 pattern, is a key structural characteristic of plant defensins and first reported from charybdotoxin, a $K^+$ channel blocker [42]. Their secondary structures show that an α-helix is parallel to three antiparallel β-strands (Figure 1.5). Similar to thionins, defensins contain a pseudocyclic structure, which is formed by the backbone cyclization between N- and C-terminus. However, defensins are different from other cysteine-stabilized helical peptides in which the α-helix is stabilized by the C-terminal β-sheet instead of other secondary structural elements of the peptide. The great

numbers of disulfide bonds and the conservation of CSαβ motif enable plant defensin to remain stable in harsh conditions [36].

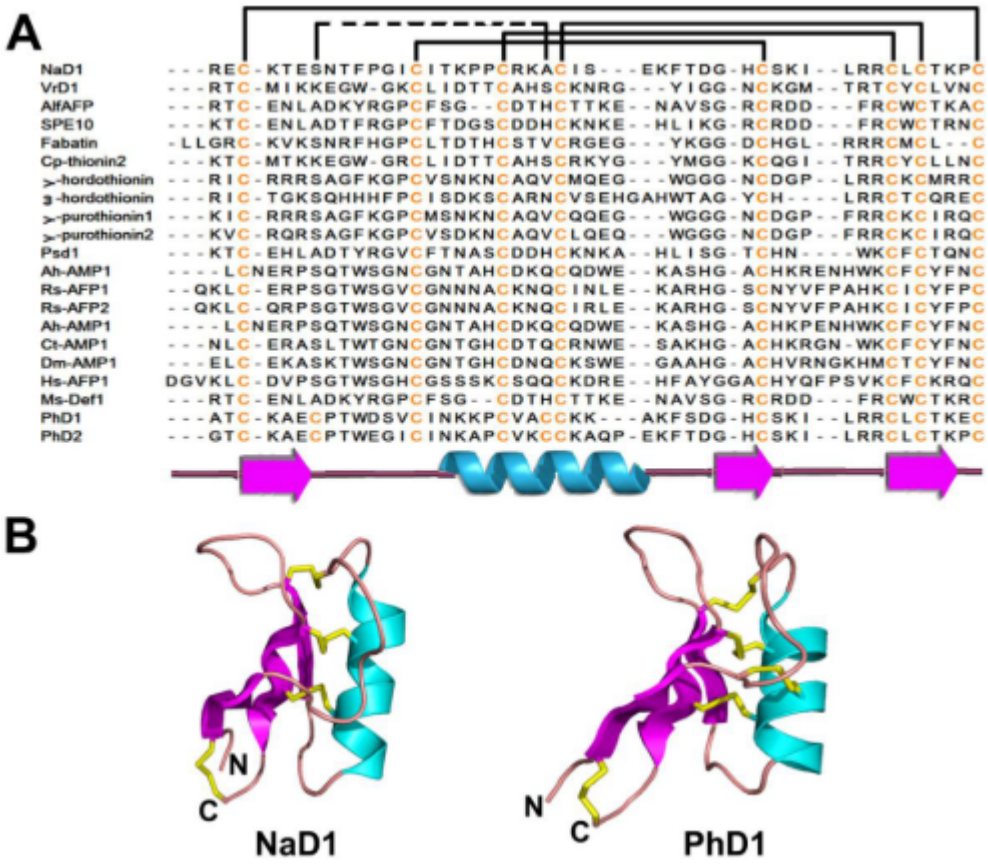



**Figure 1.5. (A) Sequences and (B) structures of examples of plant defensins.** The α-helix is shown in cyan, β-strand is represented in magenta, random coil is displayed in pink and disulfide bonds are displayed in yellow. The figure is adapted from reference [12].

Plant defensins usually contain multiple biological functions such as antimicrobial, antifungal, enzyme inhibition activities [43, 44]. Their roles can be illustrated by the following plant defensins identified from different plant sources. For example, the plant defensin NaD1 isolated from the flower of *Nicotiana alata*, showed an inhibitory effect on plant pathogens *Botrytis cinerea* and *Fusarium oxysporus* [45]. VrD1, the plant defensin from *Vigna radiate,* also displayed antimicrobial and insect-resistant activities [46]. ω-Hordothionin, which was identified from barley, showed inhibition effect in both eukaryotic and prokaryotic cell-free systems [47]. AlfAFP from *Medicago sativa* seeds and MS-Def1 from *M. sativa* seeds showed strong antifungal activity against fungal pathogens [43, 48]. Defensins S1α1, S1α2, and S1α3 purified from the *Sorghum bicolor* seeds displayed an inhibitory effect against α-amylase while two plant defensins from *Cassia fistula* showed inhibitory activity against trypsin [49].

It has been postulated that the positive charge and amphipathic characteristic of plant defensin are associated with their antimicrobial activities. This feature allows the binding of plant defensins and microbial membranes through specific binding sites. Therefore, positive ions, such as $Ca^{2+}$ and $K^{+}$ reflux will be triggered [50]. Some other mechanisms are also proposed. For example, in NaD1, its action is likely attributed to the permeabilization of the fungal hyphae instead of causing membrane permeabilization via canonical mechanism and thus leads to the inducing of ROS oxidative stress [51].

### 1.4.3. Hevein-like peptides

Hevein is a 10 kDa anionic non-glycosylated protein that is rich in cysteine, aspartic acid, and serine residues, from the latex of rubber tree *Hevea brasiliensis* [52]. Sequence analysis showed that it displays high similarity to the chitin-binding proteins like chitinases from tobacco, chitin-binding lectins from wheat and *Urtica dioica agglutinin* (UDA), displaying strong anti-fungal activity *in vitro* [53, 54]. Hevein possesses potent antifungal activity which is attributed to the presence of a chitin-binding domain, consisting of Ser19, Trp21, Trp23, and Tyr30. These residues can bind to a polymer that is present in fungal cell walls, namely N, N'-diacetylchitobiose $(GlcNAc)_2$ [55]. The binding motif distinguishes it from other peptides such as thionins, defensins and cyclotides, which possess antimicrobial activities as well.

'Hevein-like peptides (HLPs)' is a term which was first introduced when two eight-cysteine hevein homologs were identified and named Pn-AMPs from *Pharbitis nil* in 1998 [56]. Until now, only about 40 HLPs were identified from 14 different plant species [57]. Generally, HLPs are Cys- and Gly-rich peptides with 29-45 residues with three to five disulfide bonds, which protect plants by showing inhibitory effect against chitin-containing fungi and fungal pathogens. Based on cysteine numbers, HLPs are classified into 6C-, 8C- and 10C- HLPs (Figure 1.6). At the N-terminus of 8C-HLPs, a cystine-knot core arranged by three disulfide bonds is conserved. At their C-terminus, an extender part was shown to form the fourth disulfide bond comprising >12 residues [12]. The cystine-knot core is conserved in all HLPs. However, the extender at C-terminal is absent in 6C-HLPs, which were considered as the truncated version of 8C-HLPs. Until now, 13 6C-HLPs from plant Amaranthaceae family were isolated and identified [20]. In contrast, the fifth disulfide bond in 10C-HLPs either locates within the C-terminal extender or acts as a linker to the extender with the cystine-knot core [57]. It can be observed that the additional disulfide bond in 10C-HLPs is not as conserved as 6C- and 8C-HLPs, resulting in diverse structures and functions.

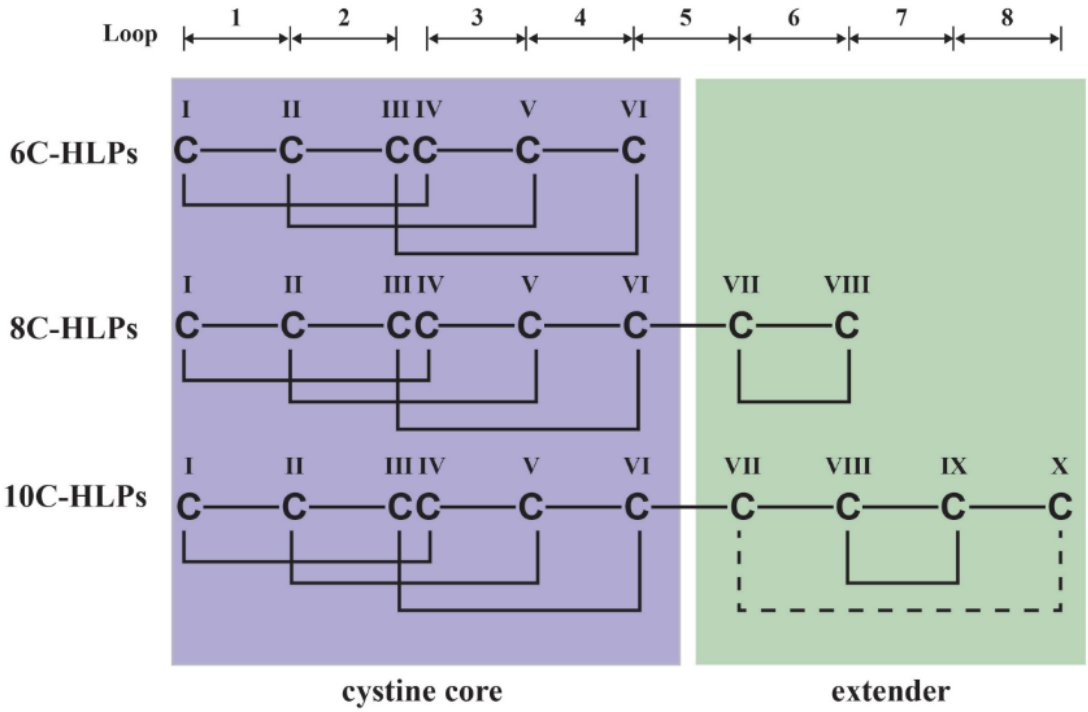



**Figure 1.6. Graphic illustration of HLPs.** The intercysteinyl loops are labeled as 1-8. HLPs are categorized into 6C-, 8C- and 10C-HLPs, based on the numbers of cysteine residues. 8C-HLPs are the prototypic member of HLPs whereas 6C-HLPs are the truncated version with the absence of an extended disulfide bond. 10C-HLPs have an additional disulfide bond. The figure was adapted from reference [57].

At the loop 3 and 4 of all the HLPs, there is a highly conserved chitin-binding domain. Glycine, cysteine and aromatic residues are the key amino acids that characterize the chitin-binding domain and are displayed in a pattern as S-X-(F/W/Y)-X-(F/W/Y)-C-G-$X_4$-Y, being stabilized by multiple disulfide bonds [58]. It is hypothesized that the anti-fungal activity of chitinases is attributed to the binding interaction between the chitin-rich fungal wall and chitin-binding domain. However, 6C-, 8C- and 10C-HLPs contain a single chitin-binding domain that is not fused to a catalytic domain with stronger anti-fungal effect than their protein counterparts.

Recently our laboratory has reported a group of CRPs from ginseng species, namely ginsentides [59]. They are 31-33 amino acid in length with eight cysteine residues. The same –CC- motif at the third and fourth position as 8C-HLPs was observed in all ginsentides. However, they lack a chitin-binding domain. The transcriptomic data mining has revealed more than 50 ginsentide-like peptides, which contain the same cysteine motif of CXnCXnCCXnCXnCXnCXnC, are expressed in more than 30 plant species. Without the

presence of the chitin-binding domain, this group of CRPs were classified into a new subfamily of HLPs and termed as non-chitin-binding HLPs [59].

Structural analysis revealed the 3D structures of HLPs and provided a basis for analyzing their carbohydrate-binding activity. Generally, a secondary structural motif of coil-β1-β2-coil-β3 was observed in HLPs with a few exemptions based on the presence of short turns in the two long coils and β3 strand [60] (Figure 1.7). A short helical segment is often found in the C-terminal segment of HLPs, whereas two anti-parallel β-strands with the core are stabilized by multiple disulfide bonds [61].

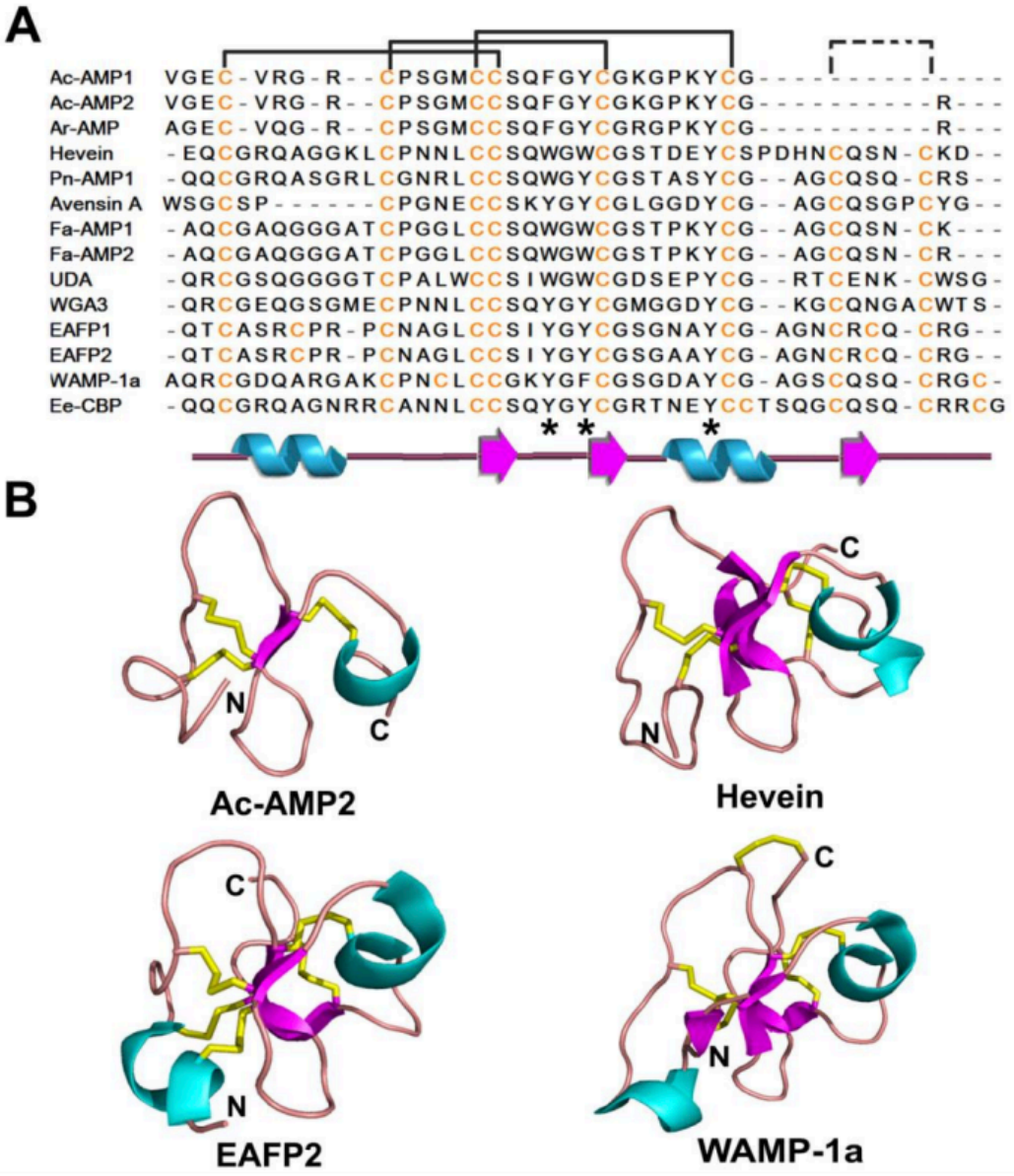



**Figure 1.7. (A) Sequences and (B) structures of HLPs.** The α-helix is shown in cyan, β-strand is represented in magenta, the random coil is displayed in pink and disulfide bonds are

highlighted in yellow. *Residues involved in the chitin-binding domain. The figure is adapted from reference [12].

6C-HLPs like Ac-AMP1 and Ac-AMP2 from the *Amaranthus caudatus* seeds exhibit antimicrobial and antifungal activities, which is antagonized by cations [12]. Generally, the structures of 6C-HLPs consist of a β-sheet together with two antiparallel β-stands which linked to an N-terminal coil region, and a disulfide bond that links the C-terminal helical turn to the first β-strand. This structural scaffold exposes chitin-binding domain to the surface. The previous study showed that the absence of the C-terminus of hevein in 6C-HLPs reduces the binding affinity of chitooligosaccharide ligands by 20% and also reduces the binding conformation stability [60].

Unlike 6C-HLPs, the 8C-HLPs contain an additional fourth disulfide bond, which causes the formation of a central β-sheet consisting of three antiparallel β-strands [12]. 8C-HLPs like Pn-AMPs are the first hevein-like peptides reported to possess potent antifungal activity against both chitin-containing and non-chitin-containing fungi, but lose their activity under acidic or reducing conditions [56]. They have been cloned into transgenic plants to endow these plants anti-fungal activities [62]. Additionally, 8C-HLPs were shown to possess potent antibacterial activities, such as Fa-AMPs identified from the seeds of *Fagopyrum esculetum* [63].

10C-HLPs contain an additional fifth disulfide bond, and its position varies by peptides. Two 10C-HLPs EAFP1 and EAFP2 purified from *Eucommia ulmoides Oliv* bark show inhibitory effect against fungi with or without chitin and can be antagonized by cations [64]. Structural analyses reveal a distinct feature that except for containing a chitin-binding domain, their fifth disulfide bond links the N-terminal coil with the third β-strand [65]. Another different structural motif of 10C-HLPs can be observed from WAMP-1a, which is isolated from wheat. Its fifth disulfide bond has linked the C-terminus to the central part of the structure [66]. In contrast, the fifth disulfide bond of the 10C-HLP Ee-CBP is located at its C-terminus [67].

The antifungal activity of hevein was first investigated due to its sequence similarity with UDA, a protein that exhibits an inhibitory effect against fungi growth. [68]. Similar to hevein, HLPs are thought to be associated with plant defense mechanism due to their strong inhibitory effect against the chitin-containing fungi. Although the mechanism of antifungal activities of HLPs still remains unclear, the hypothesis has proposed that these functions largely depend on their highly conserved chitin-binding domain. The interaction between hydrophobic C-H groups of carbohydrates and the π-interaction of aromatic acids in hevein and HLPs play a significant

role in the chitin-binding activity, as shown in Ac-AMP mutants, hevein and hevein32 at key integrating positions [69, 70]. It is proposed that the small-size and the chitin-binding characteristics enable HLPs to penetrate the cell wall and inhibit nascent chitin chains and therefore inhibit the hyphal growth [68]. Additionally, the chitin-binding property may inhibit the hyphae growth by interfering with the chitin-synthesis and chitin-hydrolysis procedures. However, these hypotheses could not explain for the inhibitory effect against non-chitin-containing fungi of other HLPs. An alternative mechanism was illustrated by Pn-AMPs, which revealed that HLPs could penetrate fungal cell walls and result in the burst of the fungal membrane, and therefore causes the leakage of cytoplasmic material [56]. This ability could be attributed to their highly positive charges. Thus the highly cationic HLPs could possess a wide range of anti-fungal activities. In addition to the chitin-binding function, HLPs were shown to display anti-fungal activity via the proteolytic inhibitory effect. For example, WAMPs containing an additional Ser36 can inhibit the fungal metalloprotease Fv-cmp, which can cleave the chitinase at the Gly-Cys site of the chitin-binding domain.

The antibacterial property of HLPs against Gram-positive bacteria is likely through the binding of the chitin-binding domain with the peptidoglycan layers, which account for approximately 90% of their dry weight of Gram-positive bacteria. Peptidoglycan comprises repeated units of N-acetyglucosamine and N-acetylmuraminic acid linked by β-(1-4)-glycosidic bonds, which may assist in the binding of HLPs with the bacterial cell surface and hence results in membrane permeabilization [71]. The aggregation of the peptide on the cell surface could link to the loss of integrity of cell membranes and cause cell death. Together, HLPs showed a wide range of defense mechanism in plant host defense system.

### 1.4.4. Knottin-type peptides

Plant knottins are a superfamily of CRPs with approximately 30-40 amino acids in length. The well-studied plant knottins are α-amylase inhibitors, trypsin inhibitors, carboxypeptidase inhibitors, and cyclotides. Knottins usually contain six cysteine residues with three disulfide bonds arranged in a cystine-knot, a disulfide arrangement can also be found in plant defensins and HLPs. However, their cysteine spacing and disulfide core structure vary from each other [12].

Historically, the knottin-type peptides were collectively named as cystine-knot inhibitor peptides, knottins, since their discovery as protease inhibitors that share the same cystine-knot motif in their structure. The use of knottins enables the differentiation of cystine-knot CRPs from the CRP growth factors identified from animals in terms of their different penetrating disulfide bonds [72]. The potato carboxypeptidase inhibitor (PCI) was the first proteolytic

knottin scaffold reported in 1982 [73]. Previously, the knottins were thought to be the largest CRP superfamily even when compared with defensins in terms of sequence diversity and molecular forms [12].

Generally, knottins are mainly present in two forms, linear and cyclic, based on whether a backbone (head-to-tail) cyclization was observed. Linear knottins are widely expressed in plants, fungi, insects, and spiders. The presence of identical cystine-knot peptides or cysteine-knot peptides with related structures found in various forms of life illustrates the parallel evolution of protein structures [12]. Differently, cyclotides and their acyclic variants are found to be only expressed in plants from families of Rubiaceae, Violaceae, Cucurbitaceae, Fabaceae, Solanaceae and Poaceae [74-78]. Although cyclic knottins such as squash trypsin inhibitors are often grouped as cyclotides, they share little sequence identity.

A typical group of linear knottins that have been well studied is cystine-knot α-amylase inhibitors (CKAIs). They are plant-derived peptides that can inhibit α-amylase enzymes and be first discovered from *Amaranthus hypocondriacus* [79]. Importantly, out of seven known families of proteinaceous α-amylase inhibitors, CKAIs are the group with the smallest molecular size. By containing the feature of pro-rich with at least one of them are in a cis-configuration, CKAIs can be differentiated from other knottins. Recent studies have shown that CKAIs are distributed in plant species from Amaranthaceae family, such as *Wrightia religiosa*, *Allamanda cathartica,* and *Alstonia scholaris* [79, 80].

The common structural motif of knottins was characterized by a cystine-knot and triple-stranded β-sheet, in which a long loop connects the first and second β-strand [81]. Usually, the disulfide bond CysIII-VI penetrates a ring formed by the other two disulfide bonds. Other knottins such as squash trypsin inhibitors and potato carboxypeptidase inhibitor showed that the Cys-stabilized β-sheet motif is maintained by the highly conserved knottin structure formed by the two disulfide bonds (CysII-V and CysIII-VI) [82, 83]. Although the cystine-knot structure is commonly found in plant CRPs at the primary structure level, it shows great variance to form secondary and tertiary structures. Additionally, the knottin-type peptides show great diversity in their sequences, intercysteinyl loops, and the linear or cyclic nature within the highly conserved CRP scaffold (Figure 1.8). Due to the tolerance for sequence diversity and highly bioactive functions, the knottin scaffold shows great potential as a template for engineering peptidyl therapeutics [84].

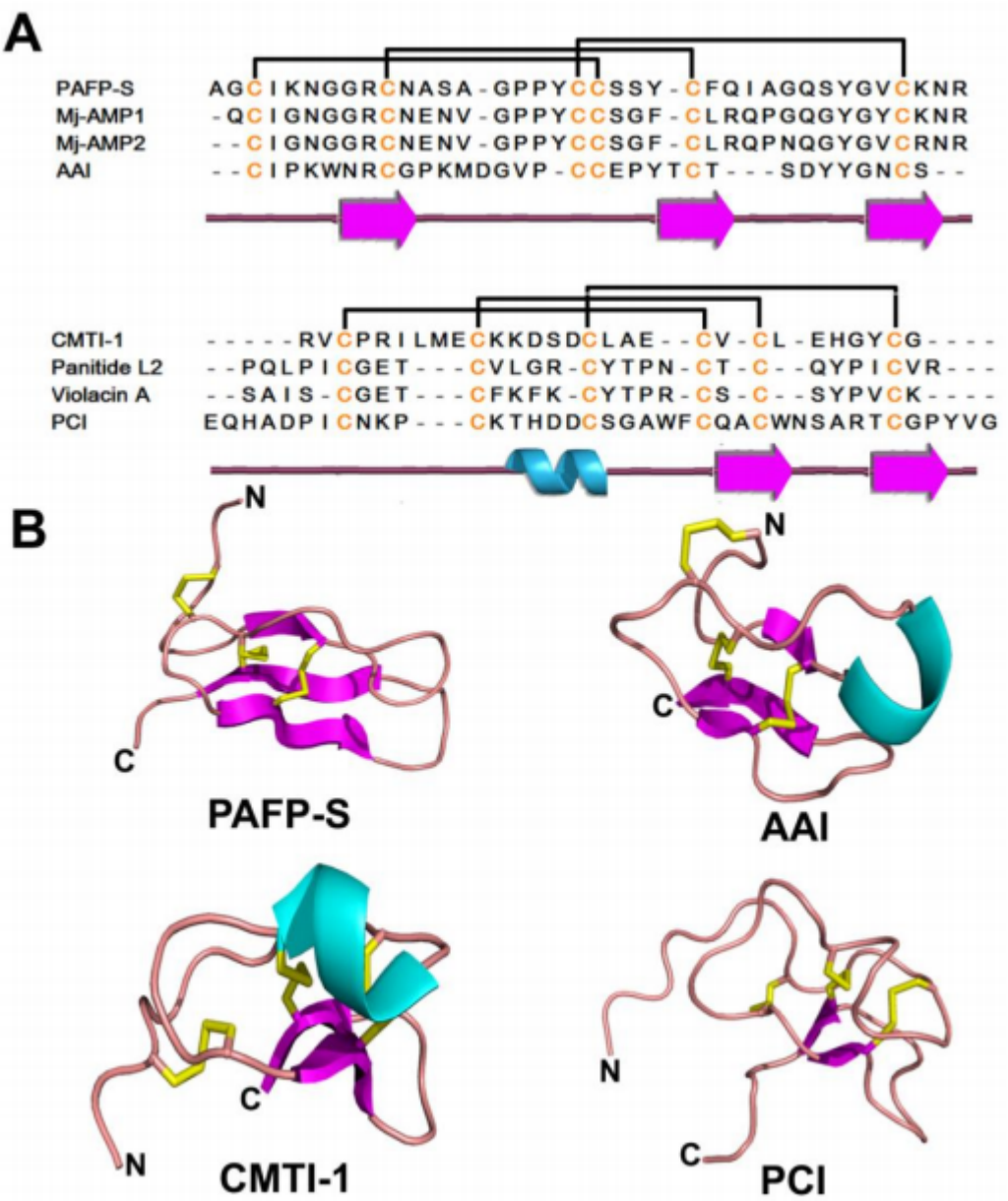



**Figure 1.8. (A) Sequences and (B) structures of representative linear knottin-type peptides.** The α-helix is shown in cyan, β-strand is represented in magenta, the random coil is displayed in pink. Yellow lines are used to indicate the disulfide bonds. The figure is adapted from reference [12].

Cyclotides, which is another subfamily of well-studied knottin-type peptides, show great metabolic stability against heat, acid, and enzymatic degradation [85]. Studies on cyclotides such as kalata B1 and violacin A showed that the disulfide bonds in the knottin structure are important for the enzymatic stability while the cyclized backbone assists in the exopeptidase resistance [86]. In addition, the simulation study reveals that the cyclization increases thermal

stability by inhibiting peptide unfolding. An alternative way for knottin-type peptides to remain stable against exopeptidase without a cyclic backbone structure is through a pseudocyclic structure as illustrated by CKAIs [26]. For example, wrightides are CKAIs isolated from *W. religiosa*, whose termini was able to loop back to the peptide chain via disulfide bonds and hence formed a pseudocyclic structure. Similar structural characteristics can be observed in CKAIs isolated from plant species *Allamanda. cathartica* and *Amaranthus hypocondriacus* [87].

Generally, the knottin-type peptides are the smallest plant CRPs but with most diverse biological activities, which include hormone-like properties, enzyme inhibitory effect, insecticidal, antimicrobial, and anti-HIV functions [81]. For example, a knottin-type peptide PAFP-S showed antifungal activity, whereas the other two knottin-type peptides Mj-AMP1 and Mj-AMP2 display potent antimicrobial activity against 13 tested fungal pathogens and two tested Gram-positive bacteria [88]. In addition, Ps from *Psacothea hilaris* is also an antimicrobial knottin-type peptide with an inhibitory concentration within the micromolar range [89].

Several other knottin-type peptides were shown to possess inhibitory effects against α-amylase or protease and were involved in the plant defense system by conferring resistance to insects, pests, and pathogens. CKAIs isolated from *A. cathartica* and *W. religiosa* are the smallest peptides that inhibit the α-amylase activity [26, 79]. Wr-AI1 and Wr-AI2 isolated from *W. religiosa* also showed an inhibitory effect against the α-amylase activity of *Tenebrio molitor* but no effect against fungal or mammalian α-amylases. Allotide Ac4 was also reported to possess a similar α-amylase inhibitory effect. However, the interaction between allotide and *Tenebrio molitor* α-amylase is different from other CKAIs, which is due to the variation in the N-terminal sequences and high content of *cis*-proline. The specificity of CKAIs targeting insect α-amylase allows them to protect the plant system without interfering mammalian digestive system and thus suggests the potential of CKAIs to act as a target for further development of transgenetic plants [12].

Other types of knottin-type peptides are protease inhibitors that possess trypsin inhibitory and carboxypeptidase inhibitory effects. The first squash trypsin inhibitor was reported from the squash seeds from the Cucurbitaceae family [90]. Naturally, these trypsin inhibitors are in linear forms except for MCoTI-I and MCoTI-II, which contain a cyclic backbone [91]. However, they share low sequence identity with cyclotides. Heitz reported that upon binding to trypsin, the free and complex cyclization structure and loops of MCoTI-II would be converted into a single, well-defined structure [92]. Compared to cyclotides such as kalata B1

or circulin A, MCoTI-I and MCoTI-II contain similar motifs of disulfide-stabilized β-sheet but differ in amino acid composition on loops 3 and 6 [92]. Additionally, they do not possess the amphipathic surface as cyclotides, and these differences may explain why MCoTIs have no antibacterial activity as cyclotides. Carboxypeptidase inhibitors are first reported from potatoes and tomatoes [93, 94]. The potato carboxypeptidase inhibitor (PCI) which can bind to the active site of carboxypeptidase A in a stopper-like manner reinforced by the secondary binding sites, has long loops that are different from the typical β-strand structure observed in other knottins [95].

The superfamily of cyclotides possesses various bioactivities related to plant host defense, as shown by their inhibitory effects against insects, nematodes, and mollusks [96]. In mammals, they show potential pharmacological functions such as anti-HIV, anti-tumor, and neurotensin activities [97]. In addition, the linear variants of cyclotides, namely acyclotides are identified from plants such as *Panicum laxum*, and *Viola odorata*, possessing inhibitory effect against *Escherichia coli* and cytotoxicity to Hela cells [86]. The first cyclotide was identified in 1973 from *Oldenlandia affinis* and possessed uterotonic activity [98]. In the following years, cyclotides are mainly found in plant Rubiaceae, Violaceae, and Solanaceae families with highly conserved structures but variable sequences and thus are divided into type classes: Möbius and bracelet [99]. The two categories do not show significant differences from each other in general scaffold structure but with the main difference that Möbius type of cyclotides contain one cis-proline in loop 5 and a twist in the cyclic backbone, while this feature is absent in bracelet type (Figure 1.9). Kalata B1, circulin A and ctclopsychotride were first reported to possess antimicrobial activity. Subsequently, more cyclotides with antimicrobial activities were discovered. For example, kalata B1 and B7 show antibiotic effects while cycloviolacin O2 displayed resistance to Gram-negative bacteria, and Hedyotide B1 showed inhibitory effects against both Gram-positive and Gram-negative bacteria [100]. Cyclotides cT1 and cT4 are antimicrobial peptides with cytotoxicity to HeLa cells [101].

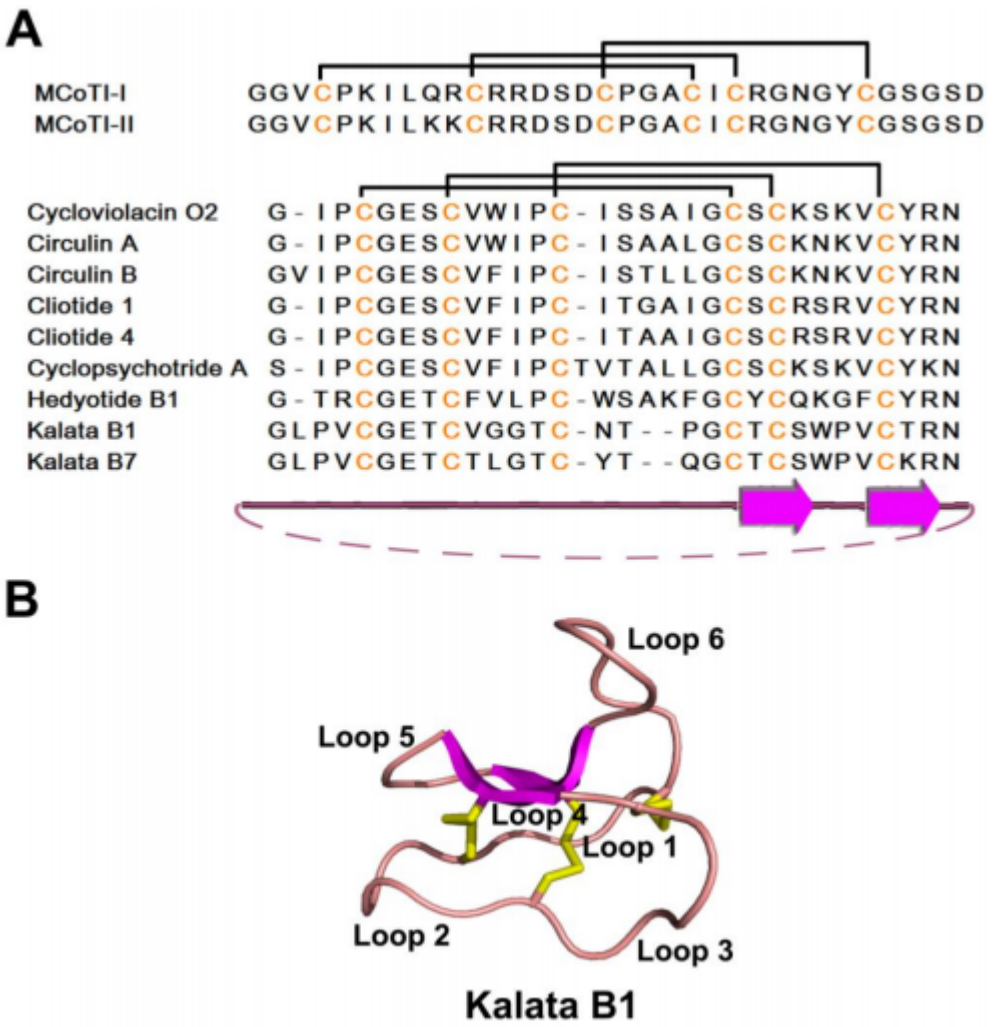



**Figure 1.9. (A) Sequences and (B) structures of representative cyclic knottin-type peptides.** The α-helix is shown in cyan, β-strand is represented in magenta, and random coil is displayed in pink. Yellow lines are used to indicate the disulfide bonds. The figure is adapted from reference [12].

The main reason accounts for the antimicrobial activity of knottin-type peptides were thought to be associated with their amphipathic nature, allowing the membrane-binding interactions [12]. The characteristic that hydrophobic patches surrounded by several hydrophilic residues are observed in the surface plots of cyclotides such as PAFP-s and kalata B1. Unlike plant thionins and plant defensins, which are highly cationic-charged CRPs, cyclotides are generally neutral or weakly positive at physiological pH. In the previous study, detergent dodecylphosphocholine was used to study the interaction between cyclotide kalata B1 and membranes *in vitro* [102]. Upon binding the detergent, the structure of kalata B1 did not

significantly change. Importantly, the binding was mediated by the interaction between its loops and lipid tails of the detergent as well as the interaction between its weak positive charge and polar head of the detergent. However, studies also show that the membrane-binding mechanism varies in different cyclotides due to the different location of their hydrophobic patches [103].

### 1.4.5. α-Hairpinin

α-Hairpinins are a group of CRPs that consist of lysine- or arginine-rich plant defense peptides. The presence of $CX_3CXnCX_3C$ motif and a helix-loop-helix secondary structure is a common characteristic of these peptides (Figure 1.10). In their tertiary structures, the α-helices are oriented antiparallel and stabilized by two disulfide bonds. This unique structure distinguishes α-hairpinins from CRPs with β-strand decoration such as thionins, defensins, and knottins [12].

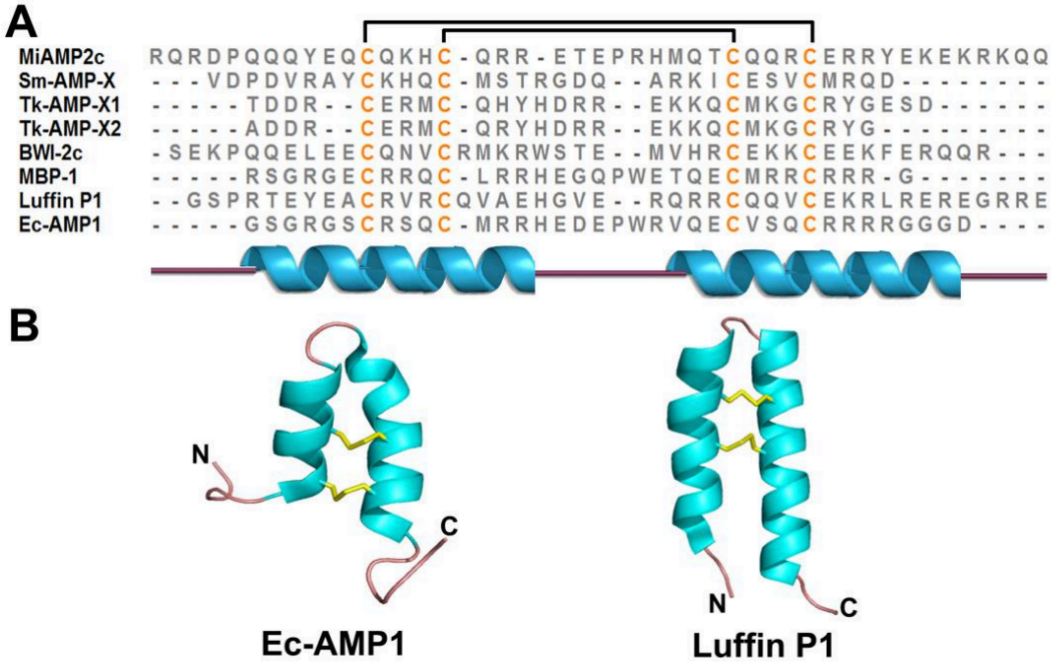



**Figure 1.10. (A) Sequences and (B) structures of representative α-Hairpinins.** The α-helix is shown in cyan, β-strand is represented in magenta, random coil is highlighted in pink and disulfide bonds are displayed in yellow. The figure is adapted from reference [12].

Until now, only a few members of the α-hairpinin family have been identified from crops. Some of them were shown to be processed from multinodular precursor proteins. For example, MBP-1 isolated from maize kernel, MiAMP2 isolated from the nut kernel and Ec-AMP1 from the barnyard grass seeds possess antifungal activities against a few plant pathogenic fungi [104]. Confocal microscopic analysis showed that Ec-AMP1 binds to the fungal conidia surface,

internalizing, and accumulating in the cytoplasm without destroying the membrane integrity [105]. Tk-AMPs isolated from wheat and Sm-AMP-X from chickweed seeds are another two members from the α-hairpinins family that show antifungal activity. Other than antifungal activity, some α-hairpinins display other biological functions. For example, VhT1 isolated from the *Veronica hederifolia* is a trypsin inhibitor while Luffin P1 shows anti-HIV-1 activity *in vitro* [12].

### 1.5. *In silico* sequence data mining of CRPs

In protein and peptide world, sequence homology can always provide clues about the biological function and evolutionary relationship of a newly sequenced gene [106]. The Basic Local Alignment Search Tool (BLAST) is a tool for performing similarity searches of DNA and protein sequence by an algorithm with high sensitivity [107].

It is crucial to translate protein-encoding DNA sequences into protein sequences before performing sequence comparison, which can help to identify related genes and correct different codon usage more efficiently. In addition, the database needs to filter out low-complexity regions to provide a significant match, which means the output of sequences should have a small expected value and reasonable alignments with the query sequences.

Different methods for building additional relationships among the matched sequences should also be considered. For example, a phylogenetic analysis may help to reveal which sequences found in an organism is most closely related to a query sequence. Therefore, this may be a group to have the same function as the query sequence and described as an ortholog of it [108].

### 1.6. Biosynthesis of CRPs

CRPs are mini-proteins and gene-encoded products that are synthesized ribosomally. Generally, the precursor sequences contain a mature CRP domain along with an N-terminal signal peptide. The presence of the signal peptide suggests their roles as secretory proteins which are destined to the endoplasmic reticulum (ER). During the translocation, the signal peptide will be removed by the membrane-bound type I signal peptidases (SPase I), and subsequently, the signal peptide will direct the synthesized preprotein to the desired destination [109]. Signal peptides are usually 16-30 amino acid in length, containing a hydrophilic N-terminal region followed by a hydrophobic central region and a C-terminal region where the SPase I cleavage site locates [110].

The procedure for releasing mature CRPs involves multiple steps. Firstly, the signal peptide is removed by SPase I after targeting to the ER lumen. Subsequently, in the highly oxidative environment, disulfide bonds are formed. The correct folding prepropeptides are assisted with

protein disulfide isomerase and chaperone, while the misfolded peptide will remain inside the ER lumen and be degraded. [111]. After the correct folding, the propeptides will leave the ER and be transported to the pre-vacuolar compartment and the vacuole, where the propeptides will be cleaved by proteases to release the mature CRPs.

There are various precursor gene arrangements observed in plant CRP families with a two to five domain architecture (Figure 1.11). The three-domain organization is the most common precursor arrangement, which can be found in thionins, defensins, HLPs, and knottin-type peptides. Thionins contain a prothionin domain that is flanked by the N-terminal signal peptide and a C-terminal acidic tail [112]. Their mature domain is more conserved than the terminal domain in the preproprotein because of evolutionary pressure. Similar to thionins, HLPs contain a typical three-domain precursor, containing a signal peptide domain, a mature peptide domain followed by a C-terminal domain. However, no C-terminal tail is observed in 6C-HLPs. The lengths of the C-terminal domain in 8C and 10C-HLPs are highly variable, ranging from 13 to 254 amino acids in length [57]. The long C-terminal domains consist of a hinge region followed by a protein cargo such as class I chitinase-like domain while this domain is absent in the short C-terminal tail. However, in our recent study, the discovery of ginsentides belonging to the non-chitin-binding 8C-HLPs revealed a new precursor arrangement, which comprises a signal peptide, a pro-domain, and a mature peptide domain [59]. In plant defensins, there are two types of precursor arrangement have been identified, wherein the majority groups adopt a two-domain architecture containing an N-terminal domain and a mature peptide domain when the minor group contains an additional C-terminal tail. The additional tail was thought to be associated with the vacuolar sorting mechanism [113]. For knottin-type peptides, generally, they contain a typical three-domain architecture as other plant CRPs whereas the precursor arrangement of the cyclic knottins, cyclotides, are different. They contain a signal domain, pro-domain, one (or more) mature cyclotide domain(s) and a C-terminal tail [76]. However, variations of precursor arrangements are observed in cyclotides. A recent study on cliotides showed that they originate from chimeric precursors, which contain albumin -1 chain A and mature cyclotide domains [77]. The highly diverse precursor arrangement present in plant CRPs suggests that plants may utilize different defense mechanisms to adapt and survive in nature.

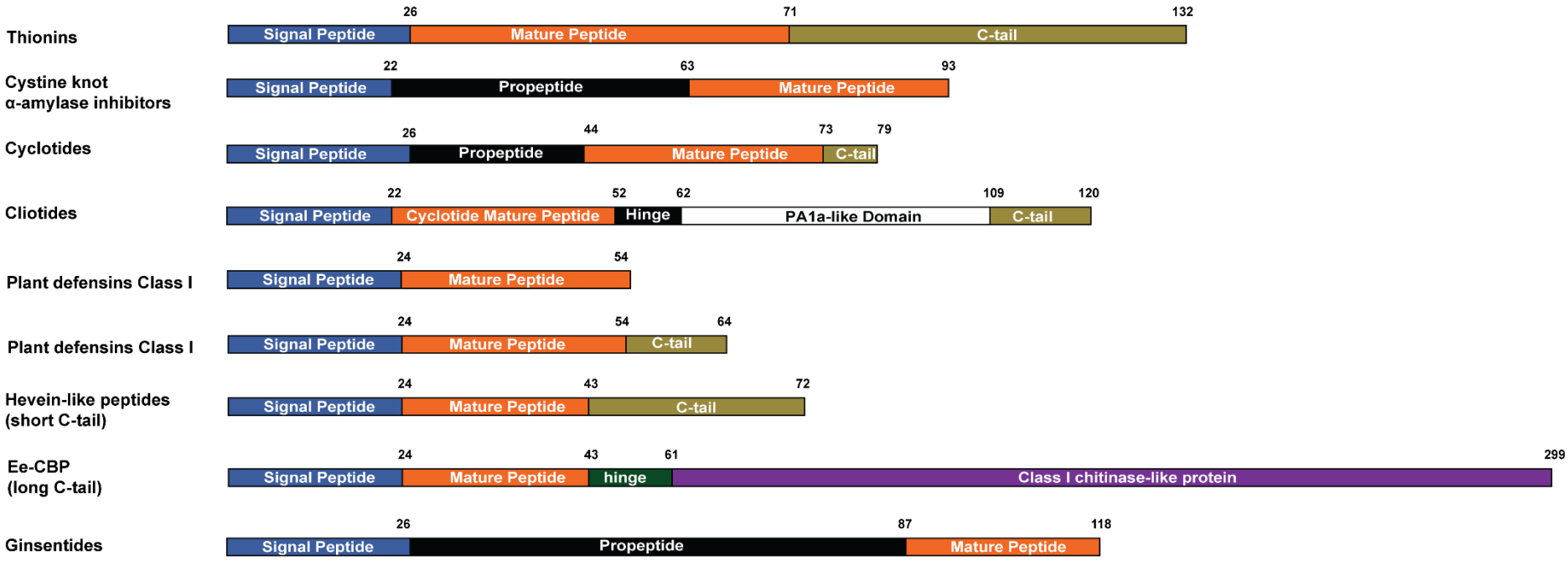



**Figure 1.11. Precursor arrangement of major plant CRP families.** Thionins and HLPs produce a similar three-domain precursor organization, where the mature peptide is located in between the signal peptide and the C-terminal domain. In HLPs, based on the length of the C-tail, they can be classified as HLPs with protein cargo and HLPs without protein cargo. In addition, the non-chitin-binding HLPs like ginsentides exhibit a different three-domain architecture where a pro-domain is present. Plant defensins adopt two classes of precursor arrangement that defensin class I contains two domains, signal peptide, and mature peptide, while defensin class II contains an additional C-terminal tail. Linear knottin-type peptides usually contain a typical three-domain architecture while cyclotides contain a four to five domain gene precursors, due to the possible presence of a chimeric arrangement.

### 1.7. Solid-phase peptide synthesis and oxidative folding of CRPs

In 1963, Bruce Merrifield pioneered solid-phase peptide synthesis (SPPS) [114], which gradually become a routine procedure for synthesizing peptides with >30 amino acids, producing good yields and high purity [115]. The method uses resin as solid support to immobilize a functioned amino acid to perform the chemical synthesis of peptides reiteratively for many cycles. In SPPS, the first amino acid of a peptide is usually attached to the resin by a stable covalent bond. Then it is deprotected to facilitate the coupling of the next amino acid in the sequence. The process is repeated until the desired sequence has been assembled after which it is cleaved. Subsequently, the free peptide in the solution can be isolated and purified by suitable procedures [114]. The general schema is shown in Figure 1.12. There are three major advantages of SPPS: It can simplify and accelerate the synthesis by carrying out all the reactions in one single vessel; It avoids the large losses involved in the isolation and purification of intermediates; It can reach a high yield of the synthesized products by using excess reactants to force the individual reaction to complete [114].

A disulfide bond is a covalent linkage between two cysteines [116]. In the 1960s, Anfinsen's team revealed that in the presence of oxygen, disulfide bonds could form spontaneously *in vitro,* which indicated that this process spontaneously occurs *in vivo* [117]. Oxidative folding globally is the method of choice of folding of CRPs. This reaction involves concurrent reduction (disulfide bond breakage) and oxidation (the formation of disulfide bonds) in the presence of a pair of redox reagents that include reducing as well as oxidizing reagents, which can reshuffle and rearrange the disulfide pairs into disulfide bonds and presumably one of the stable conformational states [118]. Upon the folding process, misfolding and aggregation procedures will compete with the correct folding pathway [119]. For disulfide bond formation, with the increasing number of cysteine residues, the possibility of combination increases [120], but the correct disulfide bond formation is biased by free energy gained upon the formation of the correct native conformation (Figure 1.13). Reduced and oxidized glutathione (GSH, GSSG) are utilized as universal redox reagent due to the low efficacy of disulfide bond formation by oxidation with molecular oxygen thiol-disulfide [121]. With low molecular weight, other thiols such as cysteamine/cystamine or di-β-hydroxyethyl disulfide/2-mercaptoethanol, are superior for disulfide bond formation depending on the respective inclusion body protein [122].

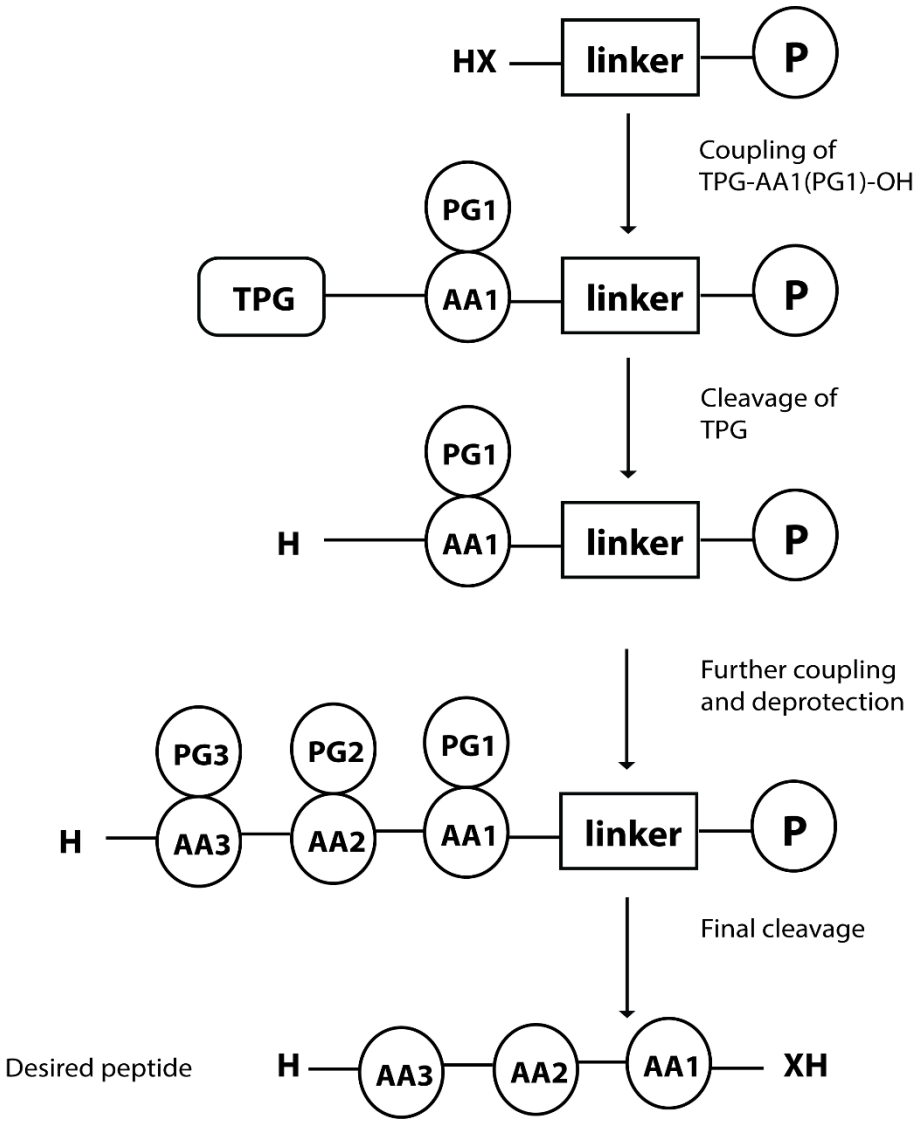



**Figure 1.12. Scheme of solid-phase peptide synthesis.** X = O, NH; AA = Amino Acid; PG = Protecting Group; P = Polymer Support; TPG = Temporary Protecting Group. The figure was adapted from BACHEM (https://www.bachem.com/fileadmin/user_upload/pdf/Catalogs_Brochures/Solid_Phase_Peptide_Synthesis.pdf)

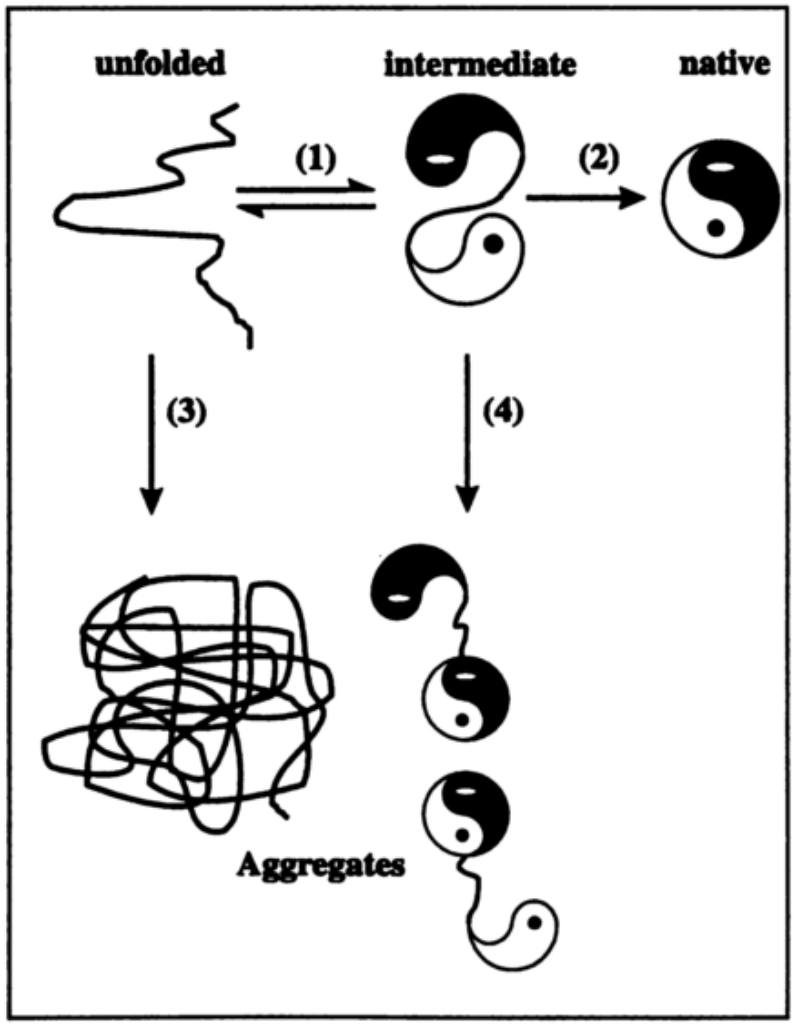



**Figure 1.13. Folding and aggregation during protein denaturation.** The correct folding reaction leads to (1) and (2). Irreversible aggregation reaction leading to renaturation process (3) and (4). The figure was adapted from reference [122].

There are several key parameters to determine the correct *in vitro* oxidative folding. For example, buffer conditions including pH, ionic strength, and salt content, have a significant influence in the folding process. Previous works have reported that pH ~8.5 is the most suitable pH condition for the efficient recovery of the native protein. Also, the selection of different redox agents can influence the kinetics of oxidative folding [123]. When oxidative agents such as GSSG or cystamine accelerate the disulfide-bond formation and lead to an accumulation of fully oxidized isomers, reducing agents such as GSH or cysteamine will promote disulfide reduction and further reshuffling of fully oxidized isomers to native structure [124]. Hence, it is important to figure out the optimal concentrations of redox agents for achieving efficient oxidative folding of CRPs *in vitro*.

### 1.8. CRPs as drug lead and scaffold

Plants have been used as medicine for a variety of diseases since prehistoric times, and these medications are still the primary sources for healthcare for 80% of the world population [125]. Importantly, approximately 80% of the herbal drugs showed a usage identical to their ethnopharmacological use. However, the principal components that are responsible for the mechanism still remain unclear, and it is a challenge to purify the complex mixture to a single compound for assays [125]. The previous studies showed that the bioactive compounds in plants or herbal medicines are mainly small-molecule metabolites such as flavonoids, alkaloids, and quinine, with a molecular weight (M.W.) < 1 kDa, such as flavonoids, alkaloids, tannins and polyphenols [126]. Examples include anti-malarial drug quinine, which is isolated from *Cichona officinalis* and morphine, an analgesic isolated from *Papaver somniferum*. These small-molecule drugs are preferred by pharmaceutical companies due to their high bioavailability and low production cost. However, their small footprints with off-target side effects have shifted the interest to biologics in recent years [127], although there is a bias that protein and peptide biologics are not suitable potential therapeutics due to their easy degradation upon heating, being digested by enzymes and low oral bioavailability. In the gastrointestinal tract, the linear peptides are easily degraded into inactive fragments by enzymes and then filtered out the body via the kidney. Therefore, injection remains the most suitable method for peptidyl drug administration due to the limitations. The approach to develop peptidyl drugs with the stability of small molecules and specificity of protein drugs could fill in the gap and serve as a promising drug candidate.

To develop such therapeutics, several studies have utilized strategies such as cyclization [128], side-chain modification [129], inserting unusual amino acids [130] or grafting bioactive peptides into natural hyperstable scaffold [131]. CRPs are a promising peptide grafting target due to the presence of multiple disulfide bonds which confers them strong stability against thermal, acidic and enzymatic degradation during the herbal decoction procedure or the digestion in the gastrointestinal tract. Moreover, CRPs display a variety of biological functions. The most well-known function is their defense-related roles in plants, especially in peripheral cells of vulnerable, nutrient-rich tissues such as flower and seed. The antimicrobial activity such as anti-fungal, anti-bacteria activities are commonly found features and thus those CRPs with antimicrobial activity are often regarded as antimicrobial peptides. Other reported functions of CRPs include toxicity against mammalian cells [132] and insect larvae [133], inhibition of α-amylase activity [49], or non-defensive roles such as reproductive regulation to growth and development [134]. The diverse medicinal value and the hyperstablity displayed

by CRPs suggest that CRPs may act as the bioactive compounds responsible for the medicinal values of herbal medicine and thus could be considered as a potential candidate for novel peptide therapeutics discovery.

The first study on developing the peptide grafting approach was conducted on scorpion toxins, named charybdotoxin in 1995. A metal-binding site was engineered on this cysteine-rich toxin [135]. Another example showing the capability of a cystine-knot scaffold is illustrated by a scorpion toxin called BmBKT×1, which as modified into an anti-tumor cell-penetrating mini-protein by grafting a p53 inhibitor and substituting multiple cationic residues [136]. This study showed that the introduction of additional peptide did not cause changes in the peptide structure and the formation of disulfide bonds.

Our laboratory has developed two orally-active bradykinin B1 receptor antagonists using the peptide grafting strategy. Clarence T. T. Wong has grafted a linear bradykinin B1 receptor antagonist DALK and DAK into the stable cyclotide kalata B1 (Figure 1.14A). Bradykinin B1 receptor is shown to be involved in the procedure of chronic inflammation stimulation, and the inhibition of this receptor may assist in releasing inflammatory pain. However, the application of linear antagonist is limited in the clinical application due to their susceptibility against peptidases. This study provided an approach to solve the problem, which is by grafting the linear bradykinin antagonists into the stable scaffold to confer the hyperstablity as well as increased oral bioavailability to the engineered bradykinin antagonists. In addition, they also display potent pain inhibition in writhing assay, which was done on mice with oral administration compared to its linear analogs [131]. Another example can be illustrated by the development of a bradykinin receptor antagonist TIBA, which is orally active and metabolically hyperstable [137]. The bradykinin B1 receptor antagonist was fused into the Sunflower Trypsin Inhibitor-1 (SFT1), which generated a bifunctional cyclic peptide with both bradykinin B1 receptor inhibition and trypsin inhibition effects (Figure 1.14B).

Taken together, these studies suggest that CRPs are highly amicable for modifications due to their tolerance for amino acids within their intercysteinyl loops and high stability. They were also shown to maintain the biological activity of the linear peptide epitopes during the modifications. With the increasing demand for peptidyl drugs, the discovery and applications of the CRP scaffolds are gaining increasing importance.

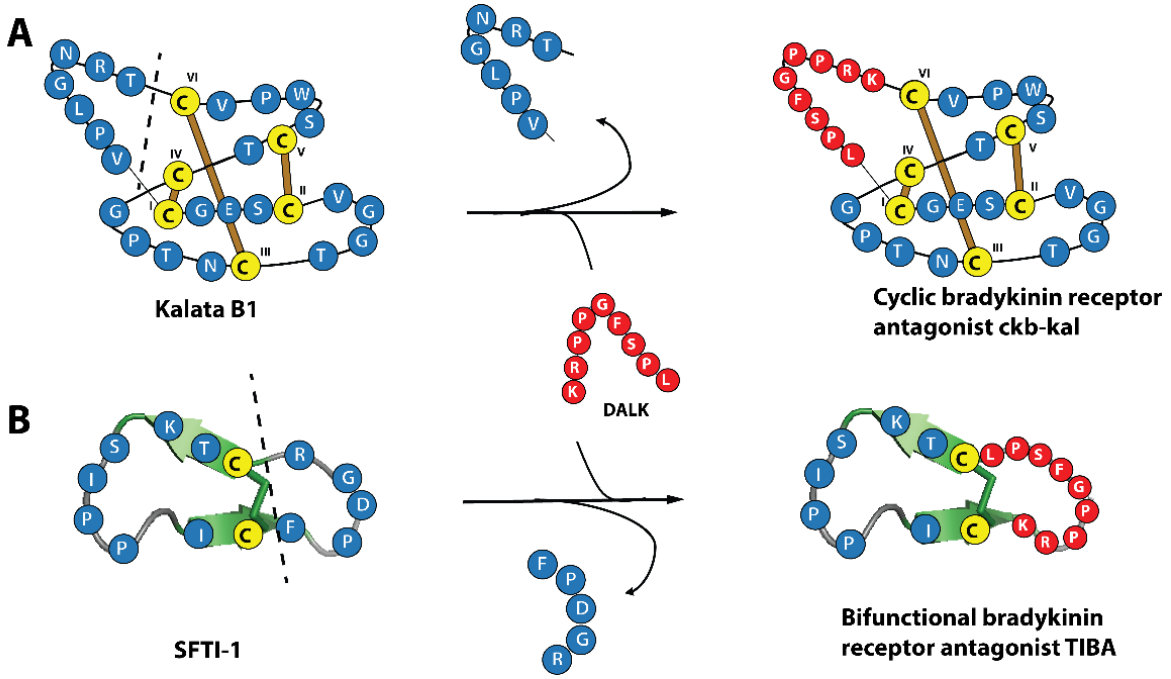



**Figure 1.14. Scheme design of two examples on engineering orally active and chimeric bradykinin receptor antagonists.** (A) The scheme of designing the cyclic, orally active peptidic bradykinin receptor antagonist ckb-kal, which is achieved by the replacement of the loop 6 of kalata B1 by the peptidic B1 receptor antagonist DALK. (B) The scheme of designing of a bifunctional peptidic bradykinin receptor antagonist TIBA, which is achieved by the replacement of the loop of Sunflower trypsin inhibitor-1 (SFTI-1) by DALK. The figure was modified based on the combination of reference [137] and [131].

### 1.9. Quality control of herbal medicine

The use of herbal medicines has been widely embraced in the world due to the reasons that herbal remedies could promote healthier living and are viewed as a balanced approach to cure individuals who use them as home remedies [138]. Herbals are generally considered harmless and can be consumed without prescription. However, some of the herbal medicines can cause serious side effects, some are toxic, and some may interact with other drugs [139]. In addition, misidentification and adulteration of herbal medicines are also important problems which sometimes are accidental and may not be noticed by those who use the plant materials. Misidentification and mislabelling of the medicinal plants sometimes may be responsible for the adverse events. For example, both *Aristolochia fangchi* and *Stephania tetrandra* S. Moore are traditional Chinese medicine known as “Fang Ji”. Although geographical identification classifies *S. tetrandra* as “Guang Fang Ji” and *A. fangchi* as “Fen Fang Ji”, their similar name and morphology still cause confusion [140]. The roots of *S. tetrandra* was used as a safe

diuretic and anti-rheumatic agent [141] while *A. fangchi* roots are not clinically favored due to the presence of aristolochic acid, which may cause renal failure. Thus the misidentification of *S. tetrandra* with *A. fangchi* will lead to aristolochic severe acid nephropathy [142]. In addition, these incorrect identifications sometimes may cause serious toxicity-related issues. For example, the traditional Chinese medicine *Acanthopanacis cortex* known as "wujiapi", is often misidentified with *Periplocae cortex,* which is known as "xiangjiapi". Although share similar morphology and name, the efficacy of the two herbs are different. Importantly, clinical studies showed that *Periplocae cortex* contains toxic compounds and may cause poisoning issues [143]. Thus, there is an urgent need to develop tools for the authentication of herbal medicines.

### 1.9.1. Qualitative and quantitative methods

Traditionally, quality control methods of herbal medicines include botanical identification based on the macroscopic and microscopic evaluation, chemical composition examination, and biological activity evaluation of the whole plant [139]. Macroscopic identification is based on sensory evaluation parameters such as shape, size, color, texture, taste, and surface characteristics. However, due to the subjective judgment and the presence of adulterants that may resemble the genuine materials, microscopic analysis, or physicochemical analysis is need. Microscopic evaluation is indispensable for the herbal medicines that are in broken or powdered form, and usually requires to be coupled with other analytical methods [144]. Chemical evaluation, on the other hand, covers screening, identification, and purification of the chemical compounds. The biological evaluation was also applied for the standardization of herbal medicines, which was indicated by their effect on the living animals or their isolated organs [145]. A systematic scheme of herbal medicine quality control was summarized in Figure 1.15.

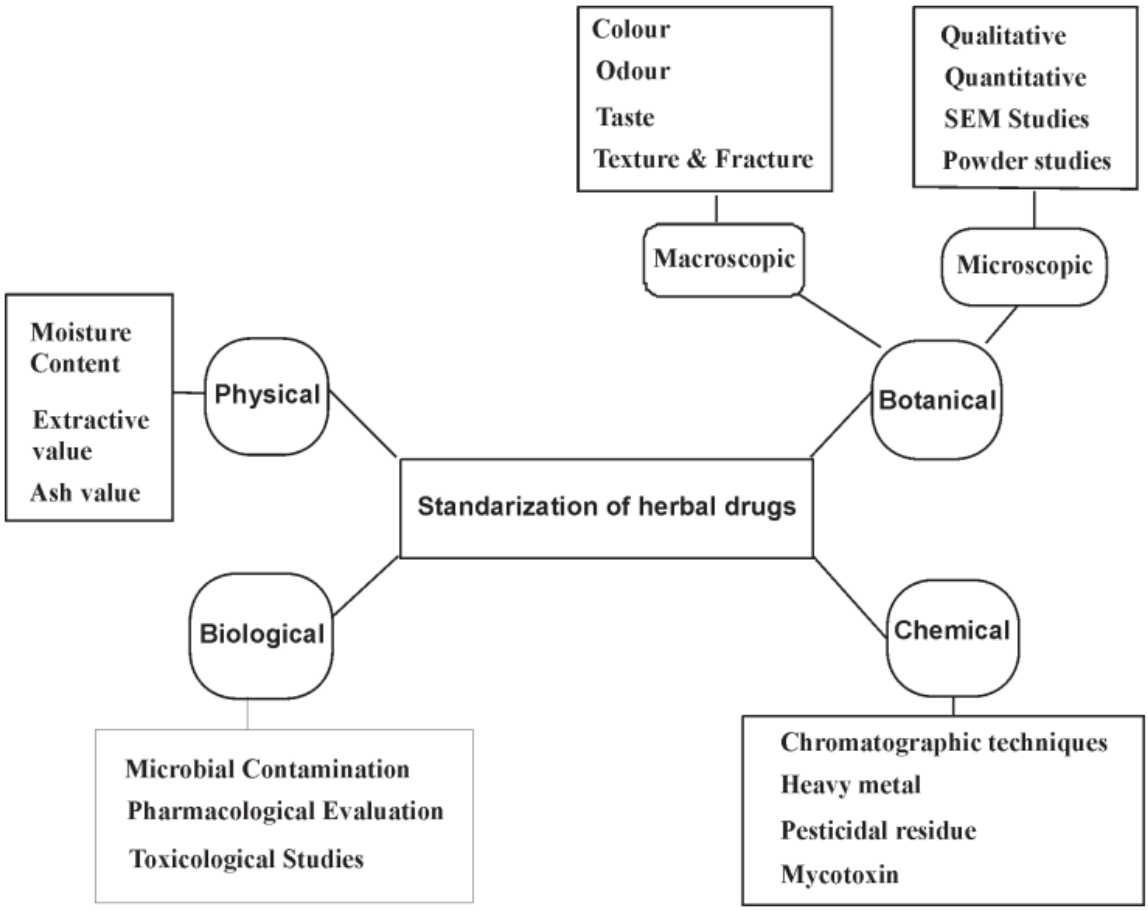



**Figure 1.15. A schematic summary of herbal medicine quality control methods.** The figure was adapted from reference [139].

The most traditional and practical approach for herbal medicine quality control is based on published monographs in a pharmacopeia [144, 146]. When the monographs are not available, other analytical methods based on chromatographic, electrophoretic, and spectroscopic techniques are developed for quality control and authentication of herbal medicine [147].

Thin-layer chromatography (TLC) is s chromatographic analytical approach used to authenticate herbal medicine, which is easy to handle with a low-cost. It can provide a unique picture-like image for visualizing the profiles of a herb and especially, with the advances in digital scanning as well as documentation software, TLC can give a comprehensive authentication of herbal medicines [148]. Previous studies have applied the method into the differentiation of species, such as the authentication of Ginseng and *Radix Puerariae* species [149]. For herbal samples containing volatile components, GC and GC-MS are the most well-known quality control method due to the high sensitivity and stability [147]. For example, the

GC-MS method was used to identify and quantify the chemical components present in polyherbal oil [150]. However, GC is not suitable for application when the samples contain polar and non-volatile components. In this case, lipid chromatography becomes another alternative analysis method. High-performance liquid chromatography (HPLC) is a 'golden standard' for authentication due to its easy handling, high accuracy properties. Since it has no limitation for sample composition, it can be used to analyze all components in herbal medicine. The most extensive use of this method is to determine and quantify molecules with similar or different structures [151]. In addition, HPLC is capable of coupling with multiple detectors such as UV, DAD, ELSD, FLD, RID, and MS, to provide higher sensitivity and better reproducibility. Furthermore, the occurrence of capillary HPLC and ultra-performance (UPLC) have increased the efficiency of analysis [152]. Another advance in the HPLC system is the development of comprehensive two-dimensional LC, leading to higher capacities, resolution, and separation. For example, by applying this technique, more than 54 compounds in a complex prescription, *Qingkaiqing*, were separated [153].

The most widely used electrophoretic method is capillary electrophoresis (CE), which can be used for nearly every kind of charged compounds, showing the complexity of a sample. The advantages of this technique are high resolution, minimal sample, and solvents consumption with short analytical time [154]. Based on different solvent and operation systems, different types of CE have been employed for the authentication of herbal medicine, including capillary zone electrophoresis (CZE), micellar electrokinetic chromatography (MEKC) and capillary electrochromatography (CEC) [155]. However, the main disadvantage of CE lies in their short optical path length and low sensitivity caused by small injection volume. When applying CE in herbal medicine authentication, the chromatographic conditions have to be optimized to obtain good separation results.

Spectroscopic techniques such as Fourier transform infrared spectroscopy (FT-IR), near-infrared spectroscopy (NIR) and nuclear magnetic resonance (NMR) are also commonly employed in herbal medicine authentication. These spectroscopic techniques usually focus more on the integrative and holistic characteristics of the herbs [147]. It is fast and straightforward to utilize these techniques since they require no pre-preparation of samples. Traditionally, FT-IR was employed to characterize the functional groups of chemical components. However, with the development of this technique, it has been broadly employed to authenticate herbal medicine. The use of this technique usually contains three steps: conventional FT-IR, second derivative spectroscopy, and 2D-IR [156]. Its ability to simplify the complex spectra containing overlapped peaks, enhance the spectral resolution and provide

dynamic information of functional groups in the system, enables the discrimination of different plant species and the detection of adulteration in herbal medicine [157]. Compared to FT-TR, NIR is more accurate and contains a more straightforward sample preparation procedure. Due to the limitations that the classification and quantitation models have to be constructed before sample analysis, a few multivariate calibration regression methods were employed to construct a consistent NIR method [158]. Like other spectroscopic techniques, the development of 2D-NIR has increased the spectral resolution, simplified the spectrum and provided more information about spectral intensity. For example, such 2D-NIR method has been employed to successfully authenticate the fruits of *Lycium* spp. from other different cultivation regions [159]. Another traditional spectroscopic identification technique NMR is based on structural analysis, which is especially used for metabolomics studies. Compared with other analytical methods, NMR contains a simpler sample handling procedure, higher stability, and reproducibility. Based on the non-selective characteristic, NMR shows outstanding discrimination ability [160]. The application of NMR technique in herbal medicine quality control mainly focuses on $^{1}H$ NMR when computer-based techniques are often coupled to reduce the complexity of NMR data. In addition, NMR spectroscopy is employed for quantitative identification (qNMR) of chemical compound composition in herbal medicine with a higher precision and lower requirement for expensive reference chemical compounds [161].

A reliable method for herbal authentication should not be affected by various conditions such as age, storage methods, physiological and environmental factors. The DNA-based analysis for herbal medicine meets the criteria and is one of the most reliable methods [162]. There are two major approaches used for the DNA-based authentication: the first is to determine the nucleotide sequence of one or more genetic loci in the plants while the second one characterizes the "fingerprints" of the whole genomic DNA [147]. Nowadays genomic fingerprint has been commonly used in authentication of plant species, homogeneity analysis, and adulterants identification. However, degradation and contamination of DNA templates, PCR inhibitors existence, poor reproducibility, and time-consuming are the limitations of this technique.

### 1.9.2. Comprehensive methods

Fingerprinting analysis usually highlights the comprehensive chromatographic and spectroscopic features of samples. Nowadays, US Food and Drug Administration (FDA), State Food and Drug Administration of China (SFDA) and the European Medicines Agency (EMEA) has accepted fingerprint analysis as one of the efficient quality control methods for herbal medicine [163-165]. Techniques such as TLC, HPLC, GC, CE, IR, NMR and DNA fingerprinting are widely accepted for the analysis, while hyphenated technologies such as GC-

MS, HPLC-MS, DAD-MS, and LC-NMR are usually used to obtain much more information [166].

However, it is not incomprehensive and not reliable to employ a single fingerprint to analyze complex herbal medicine. Therefore a few 2D techniques were developed to analyze their chemical characteristics. For example, Chen employed a 2D fingerprint to analyze multiple *Qingkailing* injections by using HPLC/DAD data at various wavelengths [167]. In addition to the binary fingerprinting, the combination of HPLC/DAD and GC/MS was used for fingerprinting of total alkaloids from *Caulophyllum robistum* [168].

In addition to chemical profiling, biological fingerprinting offers distinct advantages and can be used not only for a screening of the bioactive compounds but also can provide direct quality control for the herbal medicine [169]. These biological fingerprints focus on the correlation studies between chemical compounds and pharmacological data. For example, Li et al. used a combination of HPLC fingerprint with a bacteriostatic activity test for studying *L. japonica.* The regression analysis revealed that the two data sets could show the correlation between chemical components and the bacteriostatic activity [170].

### 1.10. Chemometrics

Chemometrics or multivariate analyses is a method combining arithmetical and statistical techniques to increase the interpretation and the correlation of chemical data and the analytical instrument data [171]. Traditionally, after the fingerprinting analysis, the chemical composition is analyzed directly or compared to the similarity based on the correlation coefficient. However, with the development of chromatographic techniques and the hyphenated instruments, much more complex chromatograms are obtained. The quality control of herbal medicine should be achieved based on the combination of analytical instruments application and chemometrics tools [171].

The application of chemometrics generally contains three parts: exploratory data analysis, regression, and classification. Usually, the first step of chemometric processing is exploratory data analysis, which summarizes the main features of data and makes a simplification [172]. It provides an overall view of the data and often uses a projection technique to obtain possible similarities/dissimilarities among samples. The most commonly used method is principal component analysis, which is a projection model providing the best fit of the data distribution. PCA has been widely used for drug quality control in the market, and the most popular application is fraud detection. For example, PCA has been used in bulk NIR spectroscopy to detect counterfeit drugs, in quality check of formulations and in obtaining composition information of different pharmaceuticals [171].

Due to the unsupervised nature of exploratory data analysis, it only provides a few unbiased pictures of the data distribution but lacks formulating prediction on new observations. Usually, a quantification or qualitative prediction of a specific compound contained in pharmaceutical products is often needed, which can be achieved by combining instrumental data with chemometric regression models [173]. Among the regression models, partial least squares (PLS) regression is the widely used model used to find the fundamental relations between two matrices. In general, these regression methods are often combined with analytical instruments to develop rapid approaches for the quantification of active components in formulations. For example, a PLS model was coupled with UV-Vis spectroscopy to quantify three analytes in a synthetic ternary mixture and different formulations [174]. Also, PLS regression was coupled with FT-Raman spectroscopy to estimate the amount of captopril and prednisolone in tablets [175].

Classification approaches aim at identifying the region in the multivariate spaces and predicting the category of the samples, which are classified into two different subfamilies: discriminant and class-modeling methods [171]. Briefly, discriminant classification models usually focus on identifying the boundaries in the multivariate space. In contrast, class-modeling approaches aim at looking for the similarities among individuals belonging to the same category and defining a subspace where with certain possibility, samples from the group under investigation can be found (Figure 1.16). Applications of classification models have been widely applied in quality control of pharmaceuticals. For example, NIR and fluorescence spectroscopy were coupled with multiple classification models to differentiate pure tablets from the adulterants [176]. In 2008, de Peinder et al. proposed a method to spot counterfeits of a specific cholesterol-lowering medicine by using NIR spectra coupled with PLS-DA [177].

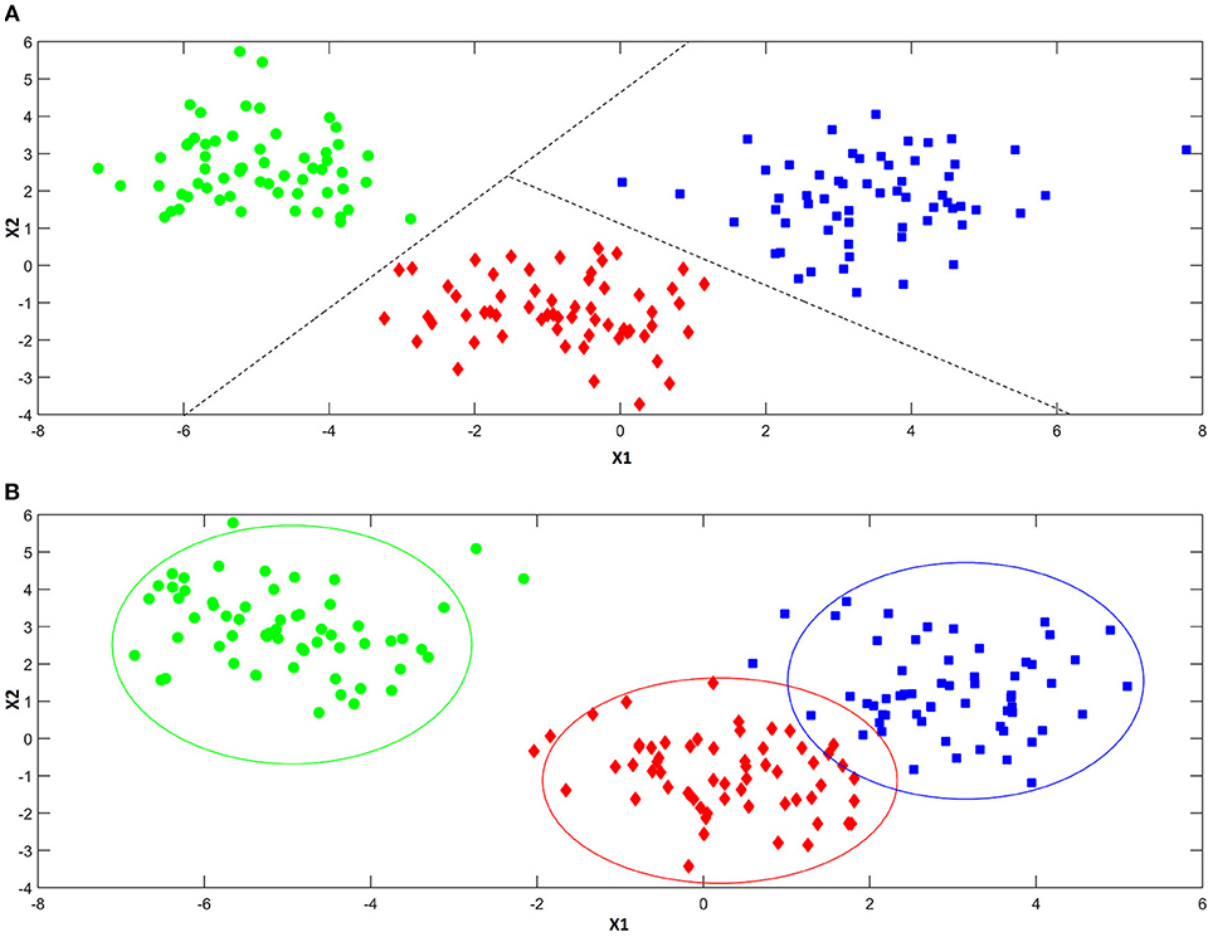



**Figure 1.16. Illustration of (A) discriminant and (B) class-modeling methods.** Discriminant analyses focus on dividing available hyperspace into many regions while modeling techniques build a separated model for each category. The figure is adapted from reference [171].

## 1.11. *Coffea*

*Coffea* belongs to the Rubiaceae family with a commonly known name as coffee. It is a genus of flowing plants whose seeds (coffee beans) are used to make a variety of coffee beverages and products. Coffee plants are shrubs, or small trees grow in tropical regions [178] (Figure 1.17). There are more than 100 species of *Coffea*, including *arabica*, *canehora*, *liberica* and *racemose*. Among these species, *C. arabica* is the most commonly used species for coffee production while *C. canephora* ranks the second. *C. arabica*, which is believed to be the first coffee to be cultivated, mainly grows in Africa, Latin America and Brazil between 1300 m and 1500 m altitude. Different from *C. arabica*, *C. canephora* (synonyms: *C. robusta*) is more economical for it is a high-yield crop and it is mainly cultivated in Vietnam in recent years. Apart from these two most popular species, *C. liberica* is a less popular species but deserves to mention. It can grow to 9 m tall and have bigger cherries. Although first being discovered

in West Africa, it was introduced to Indonesia to replace Arabica trees by the end of the 19th century and can be widely grown in Southeast Asia, including Singapore. With similar tasting characteristics to *C. canephora*, it is often used to make instant coffee.

**Figure 1.17. Coffea plant.** The picture was adapted from https://m.gafei.com/views-129262.

### 1.11.1. Commercial Importance

Coffee is the world's second-most traded commodity, following the crude oil [179]. Since 80 countries, it has become a source of income for millions of people and was considered as an important source of foreign currency for developing countries [180]. According to the International Coffee Organization (ICO), approximately 130 million bags of coffee brew were produced in 2011-12 seasons [178]. There are over 2.25 billion cups of coffee sold in the world every day, and more than 25 million small producers rely on coffee for a living worldwide. Furthermore, 52% of the population in the US consumed coffee [181].

### 1.11.2. Coffee Processing

There are two major ways for coffee processing in the industry. One is wet processing, which dries the seeds after the removal of the covering fruit pulps. At this stage, coffee pulps are generated. In contrast, the dry method requires the spreading of the freshly harvested fruits, which were stirred and ridged to get thoroughly dried. At this moment, coffee cherry husks will be generated. After processing, the fruit is subjected to the roasting stage, which will change the physical and chemical properties of coffee beans and lead to the coffee silverskin

production. The last step is brewing, in which heterophase is ranging from a smooth pure solution to emulsion. At this stage, lots of spent waste will be produced [178] (Figure 1.18).

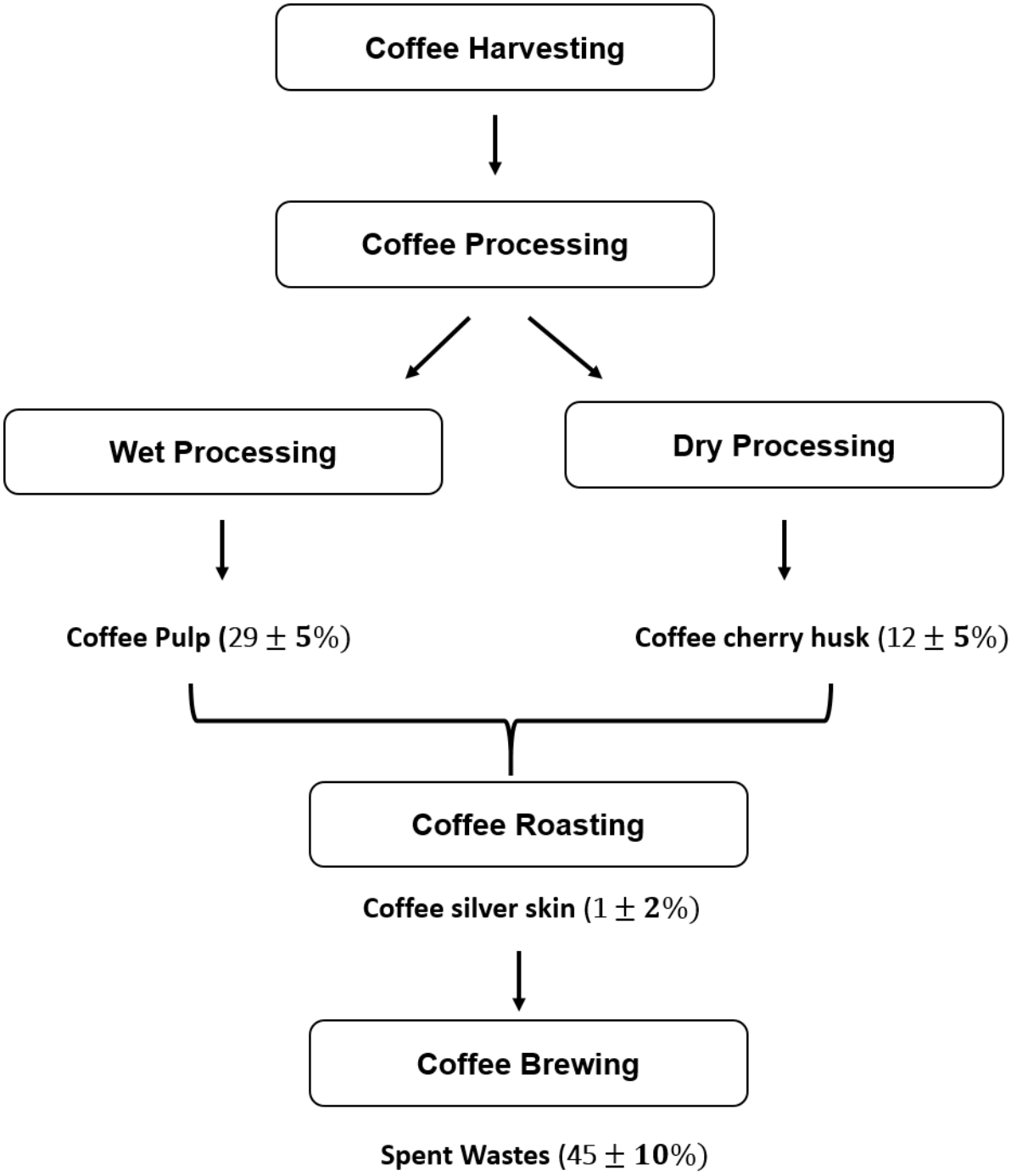



**Figure 1.18. Coffee processing procedure in the industry.** The figure was adapted from reference [178].

### 1.11.3. Traditional medical uses and secondary metabolites

Different parts of the coffee plant have been used as medical treatments around the world for a long time. The coffee seeds decoction has been used to treat influenza in Brazil and orally administrated by nursing mothers to increase milk production in Mexico and used as a cardio-tonic as well as a neuro-tonic in Thailand [179]. In Haiti, leaf decoction was used to treat

anemia, edema, asthenia, and rage, while the leaf poultice can be used for fever in Mexico [179]. For the coffee fruit, its decoction can be used for hepatitis in Haiti and used as a stimulant for sleepiness and drunkenness and as an antitussive in flu and lung ailment in Peru [179].

The major components in coffee are antioxidants such as flavonoids, polyphenols, and compounds belonging to the hydroxycinnamic acid family (caffeic, chlorogenic, and coumaric acids) [179]. Namba and Matasuse reported that coffee could eliminate the physiologic damage, which may arise during viral infections [182]. In addition, their secondary metabolites such as caffeic acid, chlorogenic acid, and protocatechuic acid are reported to possess antibacterial activities [183]. A study showed that green coffee bean extract could help to decreases in total plasma homocysteine levels [184]. Also, other studies showed that regular coffee intake could potentially decrease the susceptibility of low-density lipoprotein to oxidation and the malondialdehyde levels [185]. Dates back to 1970s, research has revealed that coffee consumption can lead to the reduction of plasma glucose levels [186]. A recent study has shown that coffee drinkers have a lower risk of having type 2 diabetes during the next few years than those who never drink coffee [187], suggesting that coffee might be regarded as the "functional food" for the prevention of metabolic diseases [188].

There are over a thousand compounds in coffee, producing unique taste and smell of it [189]. Thanks to the extensive efforts of previous research, the essential ingredients are some small-molecule metabolites, like caffeine, diterpene, chlorogenic acid, and other polyphenols. Caffeine is a natural purine alkaloid present in coffee beans [190]. Generally, it exerts its biological effects through antagonism of the adenosine receptor, which is an endogenous neuromodulator with inhibitory effects [191]. Caffeine administration is associated with some physiological effects, such as central nervous system stimulation, acute elevation of blood pressure, and diuresis [192]. Cellular studies have confirmed that caffeine has potential *in vivo* anti-inflammatory activities that are beneficial to the heart. Additionally, in coffee products, two diterpenes cafestol and kahweol can increase the level of serum cholesterol [193]. An *in vitro* cellular study revealed that diterpenes reduced the activity of LDL receptors and thus caused the extracellular accumulation of LDL, which will reduce the amount of cholesterol [194]. Moreover, chlorogenic acid is a secondary metabolite found in coffee species, which is a family of esters formed between quinic and *trans*-cinnamic acids [195]. It has antioxidant properties, which is due to the ability to scavenge various free radicals when tested *in vitro* [196].

### 1.12. *Astragalus membranaceus* (Fisch.) Bje

*Astragalus membranaceus* (Fisch.) Bje is a perennial herb belonging to plant Fabaceae family, which is approximately 50 to 150 cm height with straight, long roots and erect stems. The plant is commonly found in Korea and Northern China (Figure 1.18) [197]. The roots of *A. membranaceus,* known as Huangqi (黄芪) in Chinese and Radix Astragali in Latin, is commonly used in Traditional Chinese Medicine (TCM) to increase the overall vitality, strengthening the immune system and treating weakness, anemia, fever, chronic fatigue and fever [198]. Clinically, it is used for the treatment of renal diseases, chronic phlegmatic disorders, and gastrointestinal disturbances [197]. To date, there are more than 100 components identified from Radix Astragali. The major compounds include polysaccharides, flavonoids, saponins, alkaloid, amino acids, and other components, showing great bioactivities such as antioxidant, anti-diabetes, antitumor, and antiviral activities *in vivo* or *in vitro* [199].

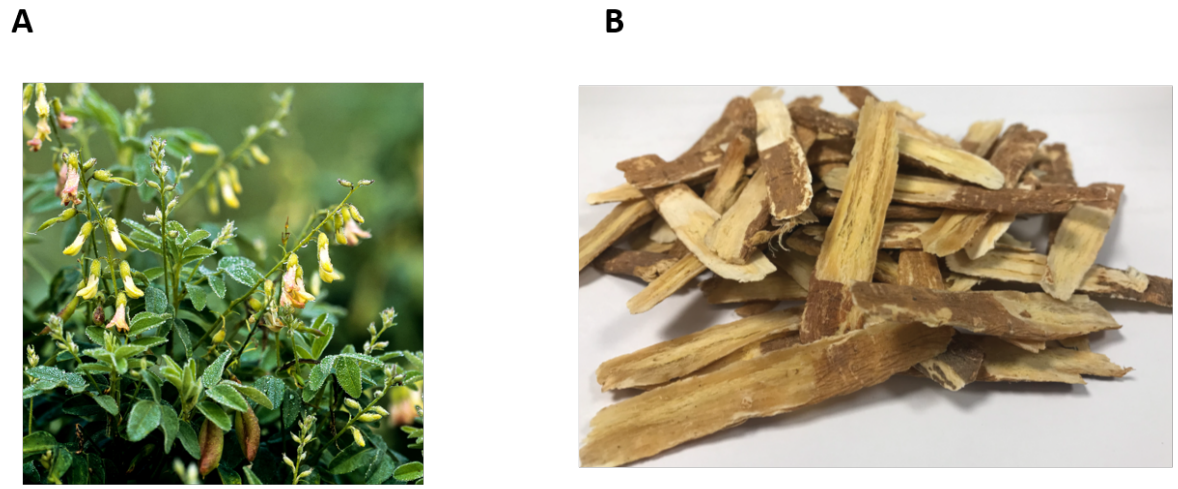



**Figure 1.19. *A. membranaceus* plant.** Currently, it is being cultivated mainly in Northern China and in Korea. (A) *A. membranaceus* plant (The picture was adapted from https://vielajoie.com/produit/astragale-huang-qi-astragalus-membranaceus/). (B) The roots of *A. membranaceus* (Radix Astragali).

More than 40 types of triterpenoid saponins were isolated from Radix Astragali and its related plants. Early years, the major saponins in Radix Astragali are divided into four groups, including astragaloside I-VIII, acetyl astragaloside, Isoastragaloside I-IV, and soysaponin [200]. In 2007, Yu et al. isolated and identified two new saponins components, which are mongholicoside A and mongholicoside B [201]. Later years, more saponins such as isoastragaloside IV, asernestiosideC, and β-daucosterol were isolated from Radix Astragali

[202]. Another dominant class of components present in Radix Astragali is flavonoids. For example, kaempferol, quercetin, isorhamnetin, calycosin-7-O-β-D-glucoside-6″-O-malonate have been isolated from Radix Astragali [197]. Polysaccharides are also important compounds of Radix Astragali, which are classified into glucans and heteropolysaccharides. Jin *et al.* have reported that there are more than 24 types of polysaccharides were identified from Radix Astragali [203].

Traditionally, Radix Astragali was used as Traditional Chinese Medicine to treat general weakness and chronic diseases. Its pharmacological effects such as improving sensitivity to insulin, modulating immune systems, displaying antiviral, antitumor activities, and enhancing cardiovascular functions have been reported in many studies [197]. It is reported that the polysaccharides in Radix Astragali can activate the β cells of mouse and interact with membranous immunoglobulin to stimulate macrophages [204]. Choi *et al.* also reported that the crude extract of polysaccharides in Radix Astragali exhibits mitogenic and co-mitogenic activities on both mouse splenocytes and human lymphocytes [205]. In addition, two saponins macrophyllosaponin b (Mac B) and astragaloside VII (Ast VII) also possess the inflammatory cytokines-inducing activity and may play essential roles for the immunostimulating and anticancer activity of Radix Astragali [206].

The crude extract of Astragali Radix is involved in clinical diabetes mellitus and early diabetic nephropathy. The polysaccharides can significantly reduce the blood sugar level, insulin resistance, and increase the serum high-density lipoprotein levels in type 2 diabetic rats [207]. This protective mechanism involves a decrease in blood glucose concentration and HbA1C levels. Moreover, the saponins astragaloside I and IV in Astragali Radix also display similar protective effects in diabetic rats, based on the inhibition of glycan end-products accumulation in both erythrocytes and nerve cells [208].

Studies have shown that Astragali Radix is a potent antioxidant for patients with severe heart and liver diseases. The flavonoids are the primary active components responsible for the antioxidant activity, which displays a significant inhibitory effect against superoxide action [209]. Jiangwei *et al.* suggested that the consumption of Astragali Radix shows effects on improving lipid profiles, inhibiting peroxidation, and increasing antioxidant enzymes activities [210].

There is an increasing amount of active compounds that have been isolated from Astragali Radix. Also, their phytochemical and pharmacological activities have been comprehensively studied. However, the quality control of Astragali Radix remains poor. Further studies need to be conducted to clarify their adulterants and different cultivated species as well as to determine

their action mechanisms, including synergistic effects and contraindication with other medications.

### 1.13. Aims and significance of the study

Medicinal plants play an important role in treating and managing human diseases. From the perspective of drug discovery, there are two major clusters of chemical spaces that are highly explored in medicinal plants and used as drugs, including small molecules with molecular weight < 1000 Da and proteins which are >10,000 Da. The chemical spaces in between are occupied by peptides. In pharmaceutical companies, peptides are usually not favored due to their easy degradation by enzyme and heat. However, the bias can be corrected by the highly underexplored CRPs, which are a group of peptides representing potential orally active therapeutics. CRPs usually contain three to five crosslinking disulfide bonds which confer them metabolic stability against harsh conditions while they also retain the specificity and potency like larger biologics to inhibit protein-protein interactions.

However, the quality control of medicinal plants and traditional herbal medicine still remains a problem due to the presence of adulterants as well as species misidentification. Nowadays, most authentication of herbal medicine is obtained by fingerprint analysis based on spectroscopic or chromatographic techniques. When compared to the conventional authentication method using HPLC technique, new analytical techniques such as MALDI-TOF MS requires less analytical time and the minimal amount of analytes. These methods mainly focus on the quantification of small molecules present in herbal medicine. However, no such study has employed CRPs, which are well-annotated compounds more easily detected by MALDI-TOF MS compared to small molecules, as standard markers for the authentication of herbal medicines. Therefore, it is of great importance to discover and characterize CRPs from medicinal plants for drug development and also to employ them as chemical markers for quality control of herbal products.

The specific aims of my thesis are:

1. To employ CRPs as unique chemical markers for the authentication of herbs and herbal products based on MALDI-TOF MS analysis coupled with chemometrics.
2. To explore the generality and the wide distribution of CRPs in medicinal plants by using *A. membranaceus* and *Coffea* species as examples.

3. To discover and characterize novel CRPs from the *A. membranaceus* roots, a commonly used herb in Traditional Chinese Medicine and explore their potential functions.
4. To discover and characterize novel CRPs form the second-most traded commodity *Coffea canephora* and *Coffea liberica* and explore their potential biological activities.

In my thesis, I focus on the highly underexplored chemical spaces, cysteine-rich peptides with M.W. ranging from 2 to 6 kDa, which are hyperstable peptides widely expressed in plants. A rapid and general method for herbal authentication based on the hyperstable CRPs was described and designated as CRP fingerprinting. Screening of 100 herbs and herbal products revealed that CRP fingerprinting produces consistent results regardless of the morphology, chemical composition, and origins of the plants. *A. membranaceus* roots and *Hedysarum polybotrys* roots, which possess similar Chinese names and morphology but different biological activities, were used as examples to validate the method. Coupling with multivariate analyses, CRP fingerprinting was able to differentiate these two species. Our study reveals that CRP fingerprinting is rapid, and its classification ability is comparable to that of the conventional method using UPLC. Thus our method could be used as a general approach for quality control and authentication of herbal and natural products.

To further the understanding of the distribution and functions of CRPs, a combination of transcriptomic, proteomic, and genomic analyses were employed to identify and characterize CRPs from several medicinal plants. *Coffea* is the species used to produce coffee drinks, which is the second-largest commodity traded in the world. Its different parts have been used to treat influenza, anemia, edema, asthenia, and fever. A family of CRPs, namely coffeetides, were isolated from *Coffea canephora* and *Coffea liberica.* Transcriptomic data mining revealed that coffeetides are widely present *in Planta* and belong to the family of non-chitin-binding 8C-HLPs. NMR analysis showed that coffeetides adopt a pseudocyclic structure with highly acidic residues on the surface. The preliminary biological assay revealed that coffeetides are non-cytotoxic, ion-binding peptides that can promote cell migration and neuron cell metabolism. Additionally, two CRPs, designated as α- and *β*-astratides were isolated and characterized from the *A. membranaceus* roots. *α*-Astratide aM1 represents the first Pea Albumin-like peptide isolated from medicinal plants, and the biological assays showed that it possess insecticidal activity as well as insulin-modulating activity. *β*-Astratide bM1, on the contrary, belongs to the plant defensins family and possesses potent antifungal activity. Its unique CXCXC motif also distinguishes it from other reported plant defensins and thus expands the existing library of

plant defensins. My findings suggest that CRPs are highly underexplored as unique fingerprints and bioactive components in natural products.

# Chapter 2 Materials and Methods

## 2.1. Materials

### 2.1.1. Chemical reagents

All the chemicals and reagents used in this study were of molecular biology or analytical grade and purchased from the following companies:

| | |
|---|---|
| Acetic acid | Merck |
| Acetonitrile (ACN) | Fisher |
| Agarose | Bio-Rad |
| Ammonium bicarbonate ($NH_4HCO_3$) | Sigma-Aldrich |
| $C_{18}$ reverse phase silica powder | Grace Davison Discovery Sciences |
| Calycosin | Chengdu Biopurify Phytochemicals, Ltd. |
| Calycosin-7-O-beta-D-glucoside (>98%) | Chengdu Biopurify Phytochemicals, Ltd. |
| Dichloromethane (DCM) | Merck |
| Dithiothreitol (DTT) | Sigma-Aldrich |
| Diisopropylcarbodiimide (DIC) | Merck |
| dNTP nucleotide mix | Fermentas |
| Ethanol (EtOH) | Merck |
| Formic acid (FA) | Sigma-Aldrich |
| Formononetin (98%) | Chengdu Biopurify Phytochemicals, Ltd. |
| Iodoacetamide (IAA) | Sigma-Aldrich |
| Isopropanol | Fisher |
| Medicarpin (>98%) | Chengdu Biopurify Phytochemicals, Ltd. |
| Methanol | Merck |

| | |
|---|---|
| N, N- dimethylformamide (DMF) | Merck |
| Ononin (>98%) | Chengdu Biopurify Phytochemicals, Ltd. |
| Sodium chloride | Sigma-Aldrich |
| Sulfuric acid ($H_2SO_4$) | Sigma-Aldrich |
| Trifluoroacetic acid (TFA) | Merck |
| Triton-X | Bio-Rad |

### 2.1.2. Enzymes

The enzymes were used for *de novo* sequencing, stability assays, and molecular cloning. Trypsin and chymotrypsin used for MS sequencing were of sequencing grade and purchased from Roche (Switzerland). Pepsin and aminopeptidase I used for proteolytic stability tests were purchased from Sigma-Aldrich (Missouri, USA). The enzymes used in gene cloning were of molecular biology grade and were purchased from Fermentas (Massachusetts, USA), Promega (Wisconsin, USA) and NEB (UK).

### 2.1.3. Plant materials

Dried *A. membranaceus* roots were purchased from a local herb distributor (Hung Soon Medical Trading Ltd., Singapore). Husks of *Coffea canephora* was collected in Feb 2016 from Vietnam. *Coffea liberica*, *Theobroma cacao*, *Triticum aestivum*, and *Eleutherococcus trifoliatus* were obtained from Singapore botanic garden and grown in NTU herb garden, Singapore. *A. membranaceus*, *C. canephora*, *C. liberica*, *T. cacao*, *T. aestivum* and *E. trifoliatus* were authenticated by Mr. Ng Kim Chuan. Voucher specimens were deposited in the Nanyang Herbarium with accession numbers of AMR-20171010, CCH-20160523, CLL-20170425, TCS20190218, TAS20190218, ETR20190218 and ESR20190218. Forty Radix Hedysarum and fifty-one Radix Astragali samples were collected from herbal pharmacies in various regions of China and Singapore (Table 2.1).

**Table 2.1. The sample code and collection location of the RH and RA samples.**

| *H. polybotrys* | | *A. membranaceus* | |
|---|---|---|---|
| Sample code | Collection location | Sample code | Collection location |
| RH1 | Gansu, China | RA1 | Anhui, China |
| RH2 | Gansu, China | RA2 | Anhui, China |
| RH3 | Gansu, China | RA3 | Anhui, China |
| RH4 | Gansu, China | RA4 | Anhui, China |
| RH5 | Guangdong, China | RA5 | Anhui, China |
| RH6 | Gansu, China | RA6 | Anhui, China |
| RH7 | Gansu, China | RA7 | Jiangsu, China |
| RH8 | Anhui, China | RA8 | Anhui, China |
| RH9 | Gansu, China | RA9 | Anhui, China |
| RH10 | Jiangsu, China | RA10 | Anhui, China |
| RH11 | Guangdong, China | RA11 | Anhui, China |
| RH12 | Gansu, China | RA12 | Anhui, China |
| RH13 | Anhui, China | RA13 | Anhui, China |
| RH14 | Gansu, China | RA14 | Jiangsu, China |
| RH15 | Guangdong, China | RA15 | Gansu, China |
| RH16 | Anhui, China | RA16 | Gansu, China |
| RH17 | Guangdong, China | RA17 | Gansu, China |
| RH18 | Gansu, China | RA18 | Gansu, China |
| RH19 | Gansu, China | RA19 | Jiangsu, China |
| RH20 | Gansu, China | RA20 | Inner Mongolia, China |
| RH21 | Inner Mongolia, China | RA21 | Guangdong, China |
| RH22 | Gansu, China | RA22 | Hong Kong, China |
| RH23 | Gansu, China | RA23 | Gansu, China |
| RH24 | Anhui, China | RA24 | Guangdong, China |
| RH25 | Jiangsu, China | RA25 | Guangdong, China |
| RH26 | Shanxi, China | RA26 | Shanxi, China |
| RH27 | Guangdong, China | RA27 | Heilongjiang. China |
| RH28 | Guangdong, China | RA28 | Inner Mongolia, China |
| RH29 | Sichuan, China | RA29 | Ningxia, China |
| RH30 | Sichuan, China | RA30 | Beijing, China |
| RH31 | Beijing, China | RA31 | Yunnan, China |
| RH32 | Jiangxi, China, | RA32 | Yunnan, China |

| | | | |
|---|---|---|---|
| RH33 | Zhejiang, China | RA33 | Shandong, China |
| RH34 | Hong Kong, China | RA34 | Qinghai, China |
| RH35 | Hebei, China | RA35 | Shanxi, China |
| RH36 | Hebei, China | RA36 | Shanxi, China |
| RH37 | Hubei, China | RA37 | Shanxi, China |
| RH38 | Hong Kong, China | RA38 | Jilin, China |
| RH39 | Shanghai, China | RA39 | Jilin, China |
| RH40 | Gansu, China | RA40 | Jilin, China |
| | | RA41 | Jilin, China |
| | | RA42 | Fujian, China |
| | | RA43 | Sichuan, China |
| | | RA44 | Sichuan, China |
| | | RA45 | Hebei, china |
| | | RA46 | Xinjiang, China |
| | | RA47 | Xinjiang, China |
| | | RA48 | Xinjiang, China |
| | | RA49 | Liaoning, China |
| | | RA50 | Liaoning, China |
| | | RA51 | Singapore |

### 2.1.4. Kits

The kits for gene cloning were purchased from Invitrogen (Life Technologies, CA, USA), Qiagen (MD, USA), Clontech (Takara Bio, Japan) and Promega (Wisconsin, USA).

### 2.1.5. Fungal strains

Four phyto-pathogenic fungal strains were acquired from China Center of Industrial Culture Collection, namely, *Fusarium oxysporum* (CICC 2532), *Alternaria alternata* (CICC 2465), *Rhizoctonia solani* (CICC 40259) and *Curvularia lunata* (CICC 40301).

### 2.1.6. Cell lines

*Spodoptera frugiperda* ovarian cells, Sf9 cells were received from Professor Julien Lescar lab, NTU, Singapore and grown at 26 °C in Sf-900 III SFM culture medium (Sigma-Aldrich, Missouri, USA) supplemented with 2% fetal bovine serum (FBS) and 1% (v/v) Penicillin /Streptomycin (PAA Laboratories). Human umbilical vein endothelial cells (HUVEC-CS cells) were received from Professor Sze Siu Kwan, NTU, Singapore. Mouse pancreatic β cell line beta-TC-6 cells (ATCC no. CRL-11506), CHO-K1 cells from Chinese hamster ovary (ATCC no. CCL-61), Human cervical cancer (HeLa) cells (ATCC no. CCL-2), neuroblastoma SH-

SY5Y cells (ATCC no. CRL-2266), H9c2 cardiomyocyte cells (ATCC no. CRL-1446) and human carcinoma A431 cell line (ATCC no. CRL-1555) were purchased from Bio-Rev (Singapore).

## 2.2. Instrumentation

### 2.2.1. MALDI-TOF MS and MS/MS

Mass spectrometry was performed on an ABI 4800 MALDI-TOF/TOF system (Applied Biosystem, MA, USA). A saturated solution of CHCA in 80% ACN with 0.1% TFA was used as a MALDI matrix. Samples were mixed thoroughly at the ratio of 1:1 (v/v) with matrix and spotted on the target plate.

Both MS and MS/MS spectra were obtained using a dual-stage reflectron mirror with laser intensity from 4000 to 5000. The average spectra for MS and MS/MS were accumulated up to 1000 and 5000 shots with an accelerating voltage of 20 kV and 8 kV, respectively.

### 2.2.2. HPLC and UPLC

Shimadzu systems (Shimadzu, CA, USA) equipped with a UV detector at 220, 254, and 280 nm were used for high-performance liquid chromatography (HPLC) and ultra-performance liquid chromatography (UPLC). Preparative, semi-preparative and analytical Reverse-Phase-HPLC (RP-HPLC) was performed on Phenomenex $C_{18}$ columns (particle size: 5 µm; pore size: 300 Å; CA, USA) with dimensions of 250 × 22 mm, 250 × 10 mm, and 250 × 4.6 mm at a flow rate of 5 mL/min, 3 mL/min and 1 mL/min, respectively. A PolyLC polysulfoethyl A column (250 × 9.4 mm and 250 × 4.6 mm) was used for semi-preparative and analytical strong cation exchange (SCX)-HPLC at flow rates of 3 mL/min and 1 mL/min, respectively.

UPLC analysis was carried out using the Nexera X2 UPLC system (Shimadzu, Kyoto, Japan) coupled with an Aeries TM PEPTIDE XB-C18 column (3.6 µm, 100 mm × 2.1 mm, Phenomenex, CA, USA) at a flow rate of 0.3 mL/min.

### 2.2.3. LC-MS/MS

A Dionex UltiMate 3000 UHPLC system (Thermo Fisher Scientific, Bremen, Germany) coupled with an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) was used to perform LC-MS/MS analysis. An Acclaim PepMap RSL column (75 µm × 15 cm; 2 µm particles, Thermo Scientific, Bremen, Germany) was used.

## 2.3. Genomics

### 2.3.1. RNA extraction

Plant materials were homogenized using liquid nitrogen. Subsequently, 1 mL TRIzol® Reagent (Life Technologies, CA, USA) was mixed with 100 mg of plant material. An incubation of chloroform (200 µL) with the sample was performed for 3 min, followed by centrifugation at 12,000 rpm for 15 min. The aqueous layer was collected, and an equal volume of isopropanol and 25% of salt solution (1.2 M sodium chloride + 0.8 M sodium acetate) were added and incubated for 10 min at room temperature to precipitate RNA. The sample was incubated at -20 °C for 30 min, followed by centrifugation at 12,000 rpm for 10 min. The RNA pellet was washed with 75% ethanol twice. The pellet was dried and resuspended in diethylpyrocarbonate (DEPC) water (20 µL). Concentration and purity of the extracted RNA were measured using NanoDrop 2000UV-Vis spectrometer (Thermo Scientific, MA, USA) and 1.0% agarose gel electrophoresis.

### 2.3.2. Rapid amplification of cDNA ends (RACE) and PCR analysis

Total RNA extract was used as a template to generate the 5'RACE cDNA library using 5' RACE (Rapid Amplification of cDNA ends) system (Invitrogen, CA, USA) according to the manufacturer's instruction. Degenerative primers were designated using GeneRunner software and were used with universal adaptor primers (UAP/AUAP) to amplify the desired sequence using a T100 thermal cycler (BioRad, CA, USA). The PCR products were run on 1.0% agarose gel electrophoresis, and the target fragments were excised and purified with Wizard® SV Gel and PCR Clean-up System (Promega, WI, USA). Then the target bands were cloned with pGEM®-T Easy Vector System (Promega, WI, USA) into JM109 high-efficiency competent cells. The insert-containing plasmids were sequenced via 1$^{st}$ Base Company's service. The specific primers to walk upstream were derived from partial coffeetide sequences targeting TCRGNC- (5'-TCAGCAGTTACCACGACAGG-3') and -3' UTR (5'-AGCTTGGAGCTTTAGCTTGAT-3'), respectively, and prepared with SMARTer$^{TM}$ RACE cDNA Amplification Kits (Clontech, Takara Biotechnology, Dalian, China) (Figure 2.1).

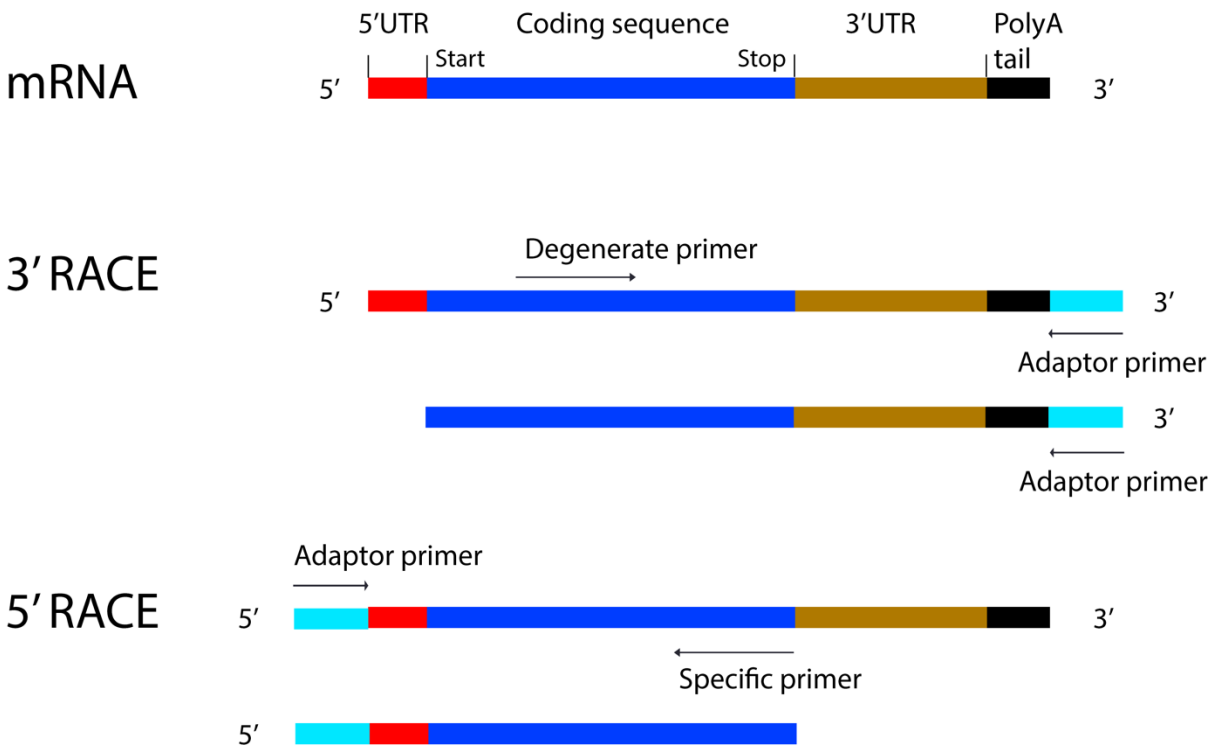



**Figure 2.1 Genetic cloning at mRNA level.** The general organization of mRNA starts with 5' untranslated region (5' UTR), followed by the coding sequence and ends with 3' UTR followed by a poly A tail at the end of the transcription process. A degenerate primer and an adaptor were employed on the 3' RACE cDNA template to first get the 3' end partial sequence. Subsequently, the segment became the base to design a specific primer to walk upstream 5'RACE cDNA in order to obtain the complete transcript.

The PCR set up was as follows: Initial denaturation at 94 °C for 5 min, main amplification for 35 cycles (denaturation at 94 °C for 30 s, annealing at Tm -3 °C for 30 s, elongation at 72 °C for 45 – 60 s), final elongation at 72 °C for 10 min, and cooling to 4 °C. PCR products were tested on 1.0% agarose gel electrophoresis, purified, cloned into pGEM-T Easy Vector, transformed into competent cells and sequenced as previously described. The whole procedure was summarized in the flowchart (Figure 2.2).

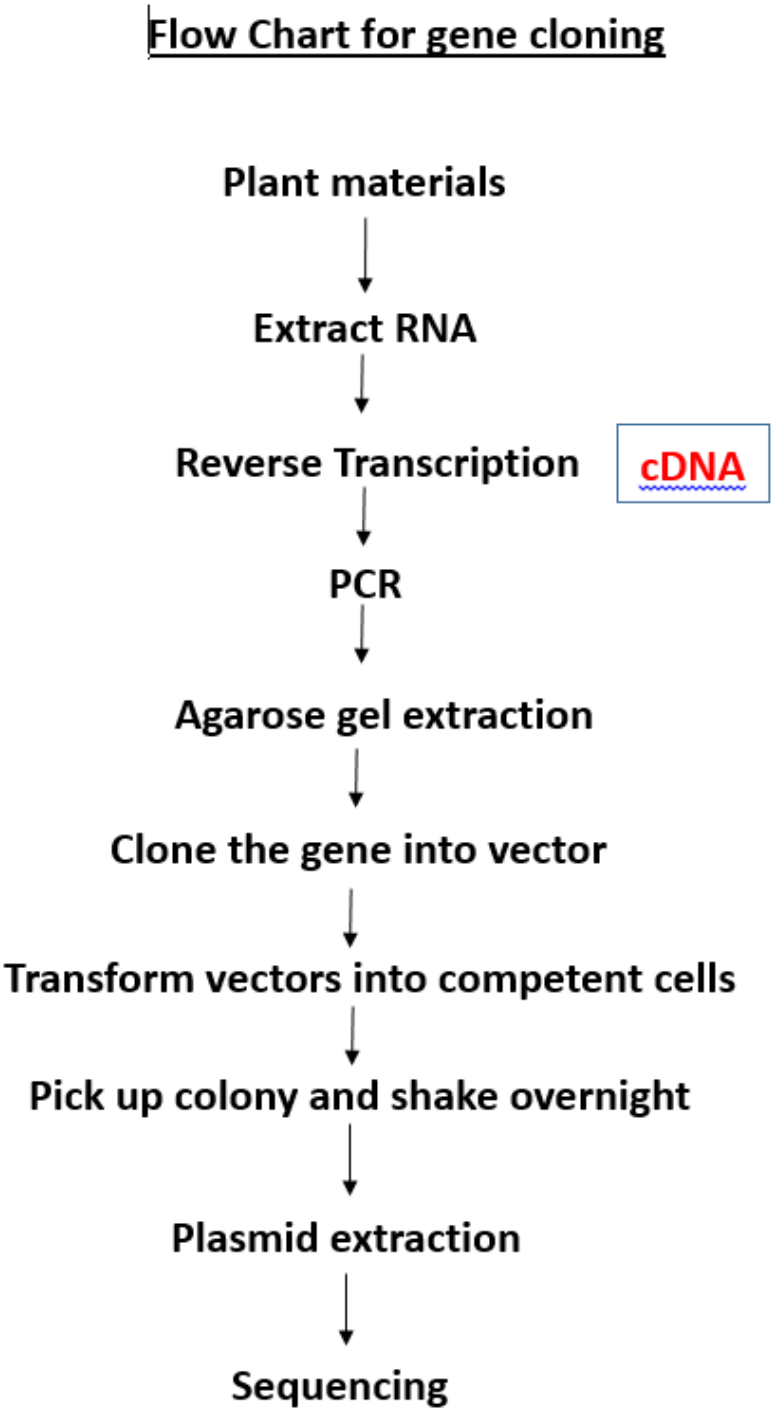



**Figure 2.2 Flowchart of gene cloning.**

### 2.3.3. Sequence analysis

The sequence analysis was performed using BioEdit software. Prediction of the amino acid sequences of the clones was done by ExPasy translate tool [211]. The open reading frame was

defined as the region between the specified start (ATG) and stop (TAA, TAG, TGA) codons. The SignalP server was used to determine the cleavage site of the signal peptide in the precursor sequence [212]. Logo sequences were built using WebLogo V3 server [213].

## 2.4. Proteomics

### 2.4.1. Screening of plant materials

100 mg of *Coffea canephora* husks, *Coffea liberica* leaves, *A. membranaceus* roots, and *Hedysarum polybotrys* roots were extracted with 1 mL water and then centrifuged at 10000 × g for 10 min to remove the plant materials. 400 mg of ammonium sulfate was added to the supernatant and shake for 1 h. After centrifugation, the pellet was dissolved in 10% acetonitrile (ACN) and subjected to Zip-tip $C_{18}$ (Millipore, MA, USA) prior to mass spectrometry analysis from 2 to 6 kDa.

### 2.4.2. CRP fingerprinting

150 mg of plant materials were extracted with 1.5 mL of Milli-Q water. The mixture was vortexed at room temperature for 1 h and centrifuged at 10,000 × g for 15 min. The supernatant was filtered through Whatman No. 1 filter paper. A Strata -X Polymeric Reversed-Phase microelution 96-well plate (Phenomenex, CA, USA) was used for sample preparation of the crude extracts for mass spectrometry analysis. Each well was percolated with water, and the filtrate of samples was loaded onto different wells under vacuum. The desired peptides were eluted with 80% (v/v) ACN and subjected to MS analysis using a detection range of 2000–5000 Da. The extraction and MS scan were performed triplicate for each sample.

### 2.4.3. Protein extraction and purification

Dried herbal samples (1-3 kg) of *C. canephora* husks, *C. liberica* leaves, *A. membranaceus* roots and *H. polybotrys* roots were homogenized and extracted with water in a ratio of 1:10 (v/v). The mixture was centrifuged at 10,000 × g for 20 min to remove the plant material. Subsequently, the supernatant was filtered through 1 µm and 0.45 µm pore-size filter papers and loaded onto a $C_{18}$ flash column (Grace Davison, Columbia, MD, USA). The sample was eluted with increasing concentrations of ethanol (20-80%). The eluted fraction that showed positive signals in the desired range from 2 kDa to 6 kDa were combined to be further purified. For *C. canephora*, the eluents were loaded onto to a flash column containing 100 mL slurry of Q-sepharose Fast Flow anion-exchange resin (GE Healthcare, California, USA) prior to RP-HPLC purification. The ion exchange flash column was percolated with 5% ACN in 20 mM

$NaH_2PO_4$ buffer (pH 7.0). The desired peptides were eluted with 5% ACN in 1 M NaCl and 20 mM $NaH_2PO_4$ buffer (pH 7.0). Eluents that contained coffeetides were pooled and purified by multiple rounds of preparative RP-HPLC. Peptides from *C. liberica*, *A. membranaceus,* and *H. polybotrys* were pooled and purified through several dimensions of RP-HPLC and SCX-HPLC. SCX-HPLC was conducted with a linear gradient from 0-100% with buffer A (5% ACN, 20 mM $KH_2PO_4$; pH 3) to buffer B (5% ACN, 0.5 M KCl, 20 mM $KH_2PO_4$; pH 3). A linear gradient from 10% to 60% with buffer A (0.1% TFA in water) and buffer B (0.1% TFA in 100% ACN) was used for RP-HPLC.

### 2.4.4. *De novo* sequencing with MALDI-TOF MS/MS

The number of cysteines present in each CRP was elucidated from the mass differences before and after *S*-reduction and *S*-alkylation. Approximately 50 µg of purified peptide was dissolved in 100 mM ammonium bicarbonate ($NH_4HCO_3$) buffer (pH 8.00). 50 mM Dithiothreitol (DTT) was then added and the reaction was incubated at 37 °C for 1 h. The reduced disulfide bonds were alkylated with 100 mM iodoacetamide (IAA) in room temperature for 1 h. Enzymatic digestion of alkylated samples was performed with trypsin or chymotrypsin at a ratio of 1:5 (peptide: enzyme) at 37 °C for 15 min and then examined by MALDI-MS followed by MALDI- MS/MS analysis. *De novo* sequencing of peptides was analyzed based on both *b*- and *y*-ion series in the MS/MS profiles. The assignment of Leu/Ile and Lys/Gln were based on the transcriptomic analysis.

### 2.4.5. LC-ESI-MS/MS analysis

*S*-alkylated sample was desalted using Millipore Zip-tip and lyophilized and re-dissolved in 0.1% formic acid (FA) before subjecting to LC-MS/MS analysis. It is performed using a Dionex UltiMate 3000 UHPLC system and coupled online to an LTQ Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). Elution was performed from eluent A (0.1% FA) to eluent B (90% ACN/ 0.1% FA), with a flow rate of 0.3 µL/min.

Data were acquired by using LTQ Tune Plus software (Thermo Fisher Scientific, Bremen, Germany) to set the mass spectrometer to a positive ion mode. The spray was generated by a Michrom's Thermo CaptiveSpray nanoelectrospray ion source (Bruker-Michrom, Auburn, USA). It is alternated between a Full FT-MS (350-3000 m/z, resolution 60,000, with 1 µscan per spectrum) and a FT-MS/MS scan applying 27%, 30% and 32% normalized collision energy in high-energy collisional dissociation (110–2000 m/z, resolution 30,000, with 2 µscan averaged per MS/MS spectrum). The three most intense ions with charge >2+ were isolated with a 2 Da mass isolation window and fragmented. A source voltage of 1.5 kV with a 250 °C

capillary temperature was set and the automatic gain control for Full MS and MS2 was set to 1× 106 and the reagent automatic gain control was 5×105. Data were processed using PEAKS studio version 7.5 (Bioinformatics Solutions, ON, Canada), applying a 10 ppm MS and 0.05 Da MS/MS tolerance. The false discovery rate is 0.1%.

### 2.4.6. Disulfide mapping

20 mM tris (2-carboxyethyl) phosphine (TCEP) was used to partially reduce the peptide samples (0.2 mg) in 500 µl of 100 mM citrate buffer (pH 3.0) at 55 °C for 50 min. Later, N-ethylmaleimide (NEM) powder was added to the mixture to make a final concentration of 50 mM and incubated at 55 °C for 30 min. The reaction was quenched by an immediate injection into HPLC with a $C_{18}$ column (250 × 4.6 mm). With a linear gradient from 45% to 60% buffer B (0.1% TFA in ACN), intermediate species were separated. The fractions were collected for MALDI-TOF MS analysis to verify the number of NEM-alkylated intermediate species. Subsequently, NEM-alkylated intermediate species were fully reduced with 50 mM DTT and incubated at 37 °C for 1 h, followed by the alkylation with 100 mM IAA at room temperature for another 1 h. The reaction was stopped by injecting into HPLC, and the samples were analyzed directly by MALDI-TOF MS/MS.

### 2.4.7. Spectrophotometric determination of peptide concentration

The Beer-Lamber law was used to calculate concentrations of purified peptides according to the equation:

$A = \varepsilon \times l \times c$

Where,

A: the absorbance at 280 nm of peptide solution in milliQ water measured on the Nanophotometer (Implemen, Germany)

ε: Molar absorption coefficient ($M^{-1}cm^{-1}$)

l: cell path length (cm)

The theoretical ε value of a peptide at 280 nm was calculated as follows:

$\varepsilon_{280} = (5500 \times n_{Trp}) + (1490 \times n_{Tyr}) + (125 \times n_{ss})$

Where $n_{Trp}$: Tryptophan numbers

$n_{Tyr}$ : Tyrosine numbers

$n_{SS}$ : Cysteine numbers

## 2.5. Chromatographic analysis

### 2.5.1. UPLC validation

The method for UPLC analysis was validated according to the guideline of International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceutical for Human Use (ICH) [24]. To evaluate the linearity and range, a serial dilution of standard compounds including medicarpin, formononetin, calycosin-7-O-beta-D-glucoside, calycosin, and ononin with 80% methanol was used to generate calibration curves. The authentic standard compounds at multiple concentrations (0.02− 3000 μg/mL) were subjected to UPLC for generating calibration curves, which are obtained by plotting the average peak area versus concentration of each standard compound. According to the ICH guideline, the limit of detection (LOD) was calculated using the formula 3.3*σ/slope, and the limit of quantification (LOQ) was calculated as 3.3*σ/slope while σ was defined as standard deviation.

The accuracy and precision were measured by analyzing triplicate of quality control (QC) samples at low QC (LOQ), medium QC (MQC) and high QC (HQC) concentrations. (calycosin-7-O-beta-D-glucoside: 250, 500, 1000 μg/mL; formononetin: 50, 100, 200 μg/mL; calycosin: 160, 320, 640 μg/mL; medicarpin: 200, 400, 800 μg/mL; ononin 375, 750, 1500 μg/mL). By injecting standards at different QC concentrations six times within one day, the intra-day accuracy and precision each standard were determined. By analyzing the QC samples on three consecutive days, in which the standards were injected six times daily, the inter-day precision and accuracy were determined. The relative standard deviation (RSD (%)) was used to show the precision. The relative error (RE (%)) was employed to show accuracy. The formulas were determined as:

Relative standard deviation (RSD) % = (SD/mean) × 100

Accuracy % = [(mean observed concentration – spiked concentration)/spiked concentration] × 100

### 2.5.2. UPLC measurement

The method used was modified from the Hong Kong Chinese Materia Medica Standards monographs (Hong Kong Chinese Materia Medica Standards Office, 2017a, 2017b). 50 mg of plant materials were mixed with 1 mL of 80% methanol for extraction and sonicated for 1 h prior to the centrifugation at 3000 × g for 5 min. Subsequently, the supernatant was filtered through a 0.45 μm PTFE filter. All the extracts were evaporated to dryness for approximately 3 h in an Eppendorf Concentrator Plus TM (Eppendorf, Hamburg, Germany). The dried

residues were re-dissolved in 100 µL of 80% methanol for a sonication prior to the centrifugation at 8000 × g for 10 min.

The analysis was carried out using the Nexera X2 UPLC system. A binary gradient elution method at a flow rate of 0.3 mL/min was employed using buffer A (Milli-Q water with 0.1% TFA) and buffer B (CAN with 0.1 % TFA), as follows: 10% B at 0.00−3.00 min, 10−30% B at 3.00−20.00 min, 30−38% B at 20.00−42.00 min, 38−80% B at 42.00−42.01 min, 80% B at 42.01−44.00 min, 80−10% B at 44.00−44.01 min, 10−10% B at 44.01−46.00 min. The detection wavelength was set to 230 nm. The chromatograms were documented and analyzed by Shimadzu LabSolutions Data software.

## 2.6. Structural analysis

### 2.6.1. Structure prediction

The I-TASSER program was used to predict the 3D structures of CRPs [214]. The solution structure of PA1b was obtained from the Protein Data Bank (PDB). A cartoon representation and electrostatic surface of the PA1b (PDB: 1P8B) and predicted CRP structures were created using PyMOL [215]. TM-align algorithm was used to calculate the structural similarities between the peptides [169].

### 2.6.2. NMR spectroscopy

The assignment and structure determination were performed on a Bruker 800 MHz NMR spectrometer (Bruker, IL, USA). The temperature for the NMR experiment was 298 k. Approximately 0.5 mM peptide was dissolved in 500µL of 30% Deuterated DMSO / 70% $D_2O$. Nuclear Overhauser effect spectroscopy (NOESY) experiments were performed with mixing times of 200 and 300ms in collecting NOE spectra [216, 217]. Total correlation spectroscopy (TCOSY) data were recorded with a mixing time of 69 or 78 ms using MLEV17 spin lock pulses [218]. 12 ppm and 4.375 ppm were set as the width of the spectrum and the center, respectively. A NMRPipe program was used to analysed the spectrums [219]. All 2D NMR data were recorded in the phase-sensitive model using the time-proportional phase increment method [220], with 2048 data points in the t2 domain and 512 points in the t1 domain. The assignment of NOE cross-peaks were determined using Sparky 3.12 software [221]. The chemical shifts of the proton were referenced to internal sodium 3-(trimethylsilyl)-1-propanesulfonate (DSS-$d_6$).

### 2.6.3. Structure calculations

The structure calculation was calculated by hybrid distance geometry using the software CNSsolve 1.3 [222]. Based on the intensities of the NOE peaks, the NOE distance restraints were loaded and were categorized into three groups: strong (1.8−3.0 Å), medium (1.8 − 3.4 Å) and weak (1.8−5.0 Å). Backbone dihedral angle restraints were determined based on the $^{3}J_{HN\text{-}H\alpha}$ coupling constraints in 1D $^{1}H$ NMR spectrum. The φ angle was considered to be between −100° to −160° when the coupling constant was more than 8 Hz. The structure was verified by PROCHECK program [223] and displayed using PyMOL version 1.849 [215].

## 2.7. Stability assays

### 2.7.1. Thermal and acidic stability

Purified 200 µM peptides were incubated at 100 °C in a water bath or with 2 M hydrochloride acid for 2 h and quenched with an ice bath or the addition of 1 M sodium hydroxide at different time points (0, 30, 60, 90, 120 min). Each experiment was done triplicate. Peaks from UPLC were analyzed by MALDI-TOF MS.

### 2.7.2. Proteolytic enzyme stability

Purified 200 µM peptides were incubated at 37 °C for 6 h with 4mg/mL pepsin in 100 mM sodium citrate buffer (pH 2.5) or 20 U/mL aminopeptidase I in 20 mM tricine and 0.05% bovine serum albumin (pH 8.0) at a final peptide to enzyme ratio (w/w) of 20:1 and 50:1, respectively. At each time point (0, 2, 4, 6 h), 20 µL of each sample was injected into UPLC to access degradation.

### 2.7.3. Serum-mediated Stability

Purified peptide (0.2 mg) were incubated with 25% human male serum AB-type in phenol red-free Dulbecco's modified Eagle medium (DMEM) at 37 °C for 48 h. At specific time points (0h, 12 h, 24 h, and 48 h), 50 µL of the samples were collected and quenched with 100 µL of 95% ethanol. Samples were incubated at 4 °C for 15 min and centrifuged at 13,000 rpm for 10 min to precipitate serum proteins. The extent of degradation was analyzed by the chromatograms obtained by UPLC.

## 2.8. Bioassays

### 2.8.1. Disc diffusion assay

The antifungal activity of *β*-astratide bM1 was evaluated by a radical disc diffusion assay, as previously described [224]. Four phytopathogenic fungal strains *Fusarium oxysporum*, *Alternaria alternata*, *Rhizoctonia solani* and *Curvularia lunata* were grown on potato dextrose agar plates at 30 °C until the mycelia reached 75% growth, a hole was punched and in the growing fungi and transferred to a new agar plate. The plate was incubated at 30 °C for 48 h-72 h until sufficient formation of a radical mycelial colony. Peptides (0.02 to 0.2 mg) were dissolved in MilliQ water and 6 mm discs were impregnated with 20 µL of each concentration were placed equidistant (1 cm) from the colony end and incubated for 24 h at 30 °C. Deionized water was used as a negative control. Formation of crescent-shaped inhibition zones indicated the susceptibility of fungi to test peptides.

### 2.8.2. Microbroth dilution assay

A Microbroth dilution assay was employed to determine the half-maximal inhibitory concentration ($IC_{50}$) of peptides against these four fungal strains [225]. Fungal spores were obtained from a confluent agar plate and seeded in half-strength potato dextrose broth at a final density of $2.5 \times 10^3$ cells/mL. 20 µL of different concentrations of peptides were mixed with 80 µL of spore solution in a 96-well plate and incubated at 30 °C for 24 h. After incubation, 100 µL of 100% methanol was added to fix the fungi for 30 min, and 1% methylene blue stain in 0.01 M borate buffer was added. The plate was incubated at room temperature for 30 min. Excess dye was washed with water, and hydrochloric acid (50 mM) in 50% ethanol was added to re-dissolve the stain after the plates were dried. Absorbance was read at 640 nm using an Infinite® 200 PRO microplate reader (Tecan, Männedorf, Switzerland). Half-strength media was used to treat control wells. Percentage inhibition was calculated as 100 times the ratio of absorbance of treated samples to control samples. A dose-response curve was generated using GraphPad Prism 6 built-in function. (CA, USA).

### 2.8.3. Insecticidal assay

Sf9 cells with a density of $1\times10^5$ cell/mL were seeded in 96-well plates. After incubation for 24 h, peptides at different concentrations (0.01, 0.1, 1, 5, 10, 25, 50, 100, 140 and 200 µM) were co-incubated with Sf9 cells and CHO-K1 cells for another 24 h. 0.1% of DMSO and 10% of Triton-X 100 was used as a negative and positive control, respectively. Cell viability was determined by a colorimetric assay based on the ability of viable cells to reduce 3-(4, 5-dimethylthiazol-2-yl)-2, 5-diphenyl tetrazoliumbromide (MTT). MTT solution (0.5 mg/mL) was

added to cells and incubated at 26 °C for 4 h before treatment with DMSO for 1 h. A microplate reader (Dynatech Laboratories Inc., Virginia, USA) was used to measure the absorbance at 550 nm.

For the microscopic analysis, Sf9 cells were incubated with 5 μM of aM1 at 26 °C for 15 h. At different time points (0, 1, 4, 8 and 15 h), cells were observed under a Zeiss Live Cell Observer (Zeiss, Oberkochen, Germany). The images were recorded and analyzed using axiovision rel 4.8 software (Zeiss, Oberkochen, Germany).

The Sf9 cells were labeled with PKH26 (Sigma-Aldrich, Missouri, USA) according to the manufacturer's recommendations with modifications. Sf9 cells (60,000 cells per coverslip per well in a 12-well plate) were washed with PBS three times and incubated with 4 μM of PKH26 in PBS for 4 min. An equal volume of FBS was added to stop the reaction. 10 μM sample of aM1 was added to the Sf9 cells immediately after cell membrane labeling. The medium was removed after incubation and washed with PBS twice. The fluorescence image of samples was subjected to a Zeiss 710 Confocal Microscope (Zeiss, Oberkochen, Germany). Images were processed and analyzed with ZEN image software (Zeiss, Oberkochen, Germany).

### 2.8.4. Cell-penetrating assay

HeLa cells were seeded in a 12 well chamber slide (ibidi, Martinsried, Germany) with a starting cell density of $5 \times 10^4$ cell/mL and incubated for 24 h at 37 °C. After 24 h, 200 nM of MitoTracker™ Red CMSRos (Invitrogen, CA, USA) was added to the cells and incubated for 20 min at 37 °C. PBS was used to wash the well twice before 30 μM of the peptide was added for an incubation. After 90 min, the medium was removed and the well washed with PBS twice. Subsequently, 4% ploy formaldehyde was added and incubated for 1 min before washing with PBS. 0.1% Triton X in PBS was added to permeate the cell membrane. A mounting medium Fluoroshield™ with DAPI (Sigma-Aldrich, Missouri, USA) was added the chamber slide and covered with coverslips for confocal microscopy.

### 2.8.5. Insulin secretion assay

*β*-TC-6 cells were seeded in a 96-well plate for 24 h with a starting density of 5,000 cell/mL. Different concentrations of aM1 (0.01, 0.1, 1, 10, and 100 μM) were added to the cells for a 24-h incubation. 10 μM of GW9508 (Sigma-Aldrich, Missouri, USA) and 0.1% DMSO were used as positive and negative control, respectively. An insulin Mouse ELISA Kit (Thermo Fisher Scientific, Massachusetts, USA) was used to measure insulin secretion levels.

### 2.8.6. Cytotoxicity assay

Cytotoxicity assays of the CRPs were performed on HeLa and HUVEC-CS cells. $1 \times 10^4$ cells per well were seeded in a 96-well plate and incubated at 37 °C. After 24 h, the cells were then treated with different concentrations of peptides (1, 10 and 100 μM) for overnight. After the incubation, an MTT assay was performed as previously described in section 2.8.3. 0.1% DMSO and 1% Triton-X were employed as the negative and positive control, respectively. The statistical significance was evaluated by a one-way ANOVA at a probability limit of $p<0.05$.

### 2.8.7. LDH assay

A CytoSelect™ LDH Cytotoxicity Assay Kit (CBA-241, Cell Biolabs, Inc., CA) was used for the LDH release assay. Cells were cultured in DMEM with or without the presence of tested peptide for 48 h at 37 °C and 5% $CO^2$. The culture medium was mixed with the LDH reagent at a ratio of 9:1 and incubated at 37 °C for 2 h. The colorimetric quantification was performed at 450 nm wavelength in experimental triplicates.

### 2.8.8. Migration assay

A431 cells ($5\times10^5$ cells/mL) were cultured in a 12-well plate for overnight incubation at 37 °C. After the incubation, the cell monolayer in each well was scraped in a straight line to create a "scratch" with a p200 pipet tip. The cell debris and medium were removed and various concentrations of peptides (1 μM, 10 μM and 100 μM) in DMEM were added into corresponding wells, following incubation at 37°C. DMEM served as a negative control. The gap was monitored under Nikon Eclipse Ti-s inverted microscope at 4X magnification. Images were taken at 4.5 h and analyzed using T-Scratch software [226].

### 2.8.9. Isothermal Titration Calorimetry (ITC) assay

ITC experiments were performed at 298 K with a MicroCal PEAQ ITC system (Malvern Instruments Ltd., Malvern, UK). $MgCl_2$, KCl, $CaCl_2$ and $FeCl_3$ were dissolved in 10 mM Tris buffer with 100mM NaCl (pH 6.3). The peptide was prepared using the same buffer. 400 μM of the ions were titrated into 40 μM of cC1 consisting of twenty 2.5 μL injections, and the heat evolved or absorbed was measured. Control experiments were performed when ions were titrated into 10mM Tris (pH 6.3) buffer with 100 mM NaCl to account for the heat released from dilution. The titration curves were analyzed using MicroCal PEAQ-ITC analytics software (version 1.0.0.1259, Malvern Instruments Ltd.).

### 2.8.10. Ion-binding activity assays

Three methods, which include thermal shift assay (TSA), circular dichroism (CD) assay, and microscale thermophoresis (MST) assay, were employed to investigate the binding ability of the peptides. Briefly, TSA is a method that determines a shift in melting temperature measured by changes in light scattering or by fluorescence techniques. The peptide unfolding procedure will be monitored by the fluorescent dye Sypro orange, and the experiment was conducted in the iCycler iQ Real-Time Detection System (Bio-Rad, CA, USA). A solution of 20 µL of 5 µM peptide, 50 µL of Sypro orange, 10 µL of ion compound, and 20 µL of potassium phosphate buffer were added to the 96-well iCycler iQ PCR plate. The final concentration of peptide is 1 µM and the final concentration of ion is 10 µM. The plate was heated from 20 to 95 °C. The fluorescence intensity was measured with Ex/Em: 490/530 nm [227].

A Chirascan™ CD spectrometer (Applied Photophysics, UK) was used for measuring the far-UV CD spectra of peptide and ions in sodium phosphate buffer (pH 7.5) at 20 °C. The CD spectra of peptide and the ions were recorded in a 1mm cuvette with 1nm spectral bandwidth and 1nm step size over the wavelength from 190 to 260 nm. All dichroic spectra were smoothed and corrected by background subtraction for the spectrum obtained with buffer alone. CD spectroscopy data were analyzed for secondary structure content using CDSSTR [228].

MST experiments were performed on a Monolith NT.115 system (NanoTemper Technologies). The peptides and ions were dissolved in 10 mM Tris buffer with 100mM NaCl (pH 6.3). Excitation of 12 pairs of the peptide with ions at different concentrations was carried out with 50% LED power, and the emission was detected in the range of 670 to 690 nm. The data analysis and fluorescence fitting curves were performed with the NanoTemper software NT Analysis 1.4.27.

## 2.9. Chemical synthesis

### 2.9.1. Solid-phase peptide synthesis (SPPS)

Initially, 1 mmol of 2-chlorotrityl chloride resin (CTC resin) was swelled in dichloromethane (DCM) for 30 min. Subsequently, amino acid residue was coupled by adding a pre-activated solution that contains Fmoc-L-Cys (Trt)-OH (4 eq., 0.4 mmol), benzotriazol-1-yl-oxytripyrrolidinophosphonium hexafluorophosphate (PyBop) (208 mg, 4 eq., 0.4 mmol) and *N, N*-diisopropylethylamine (DIEA) (174 µL, 4 eq., 0.4 mmol) in *N, N*-dimethylformamide (DMF). The reaction was performed in the shaker at room temperature for 1 h and repeated once to confirm the coupling reaction was completed. De-protection was performed in 20% piperidine with 0.1 M *N*-hydroxybenzotriazole (HOBt) in DMF for 10 min twice. A Kaiser

test consisting of a mixture of ninhydrin, potassium cyanide and phenol (45 µl, 1:1:1, v/v/v) was used to detect the presence of free amines. When the presence of a blue color on the resin occurred, the Fmoc de-protection was completed. After the deprotection step, the resin was washed with DMF three times. Peptide elongation was carried out automatically by Liberty Blue™ Automated Microwave Peptide Synthesizer (CEM Corporation, NC, USA), using standard protocols with Fmoc amino acid, PyBop and DIEA (1, 5 and 5 eq., respectively) in DMF for single coupling at 50 °C for 10 min each except for Cys (Trt) and Arg (Pbf), which were coupled under 50 °C for 10 min twice. Final cleavage from the resin and removal of all the side-chain protecting groups was achieved by adding a cocktail mixture of triisopropylsilane (TIS)/$H_2O$/1, 2-Ethanethiol (EDT)/TFA (2.5/2.5/2.5/92.5 v/v). After 1 h, the cleaved peptide was precipitated with diethyl ether (9 eq.) and centrifuged for 10 min at 6,000 rpm to get crude peptide. To confirm the presence of peptide, UPLC was run at a linear gradient 10-60 % ACN/0.1% TFA for 18 min. The synthesis scheme was summarized in Figure 2.3.

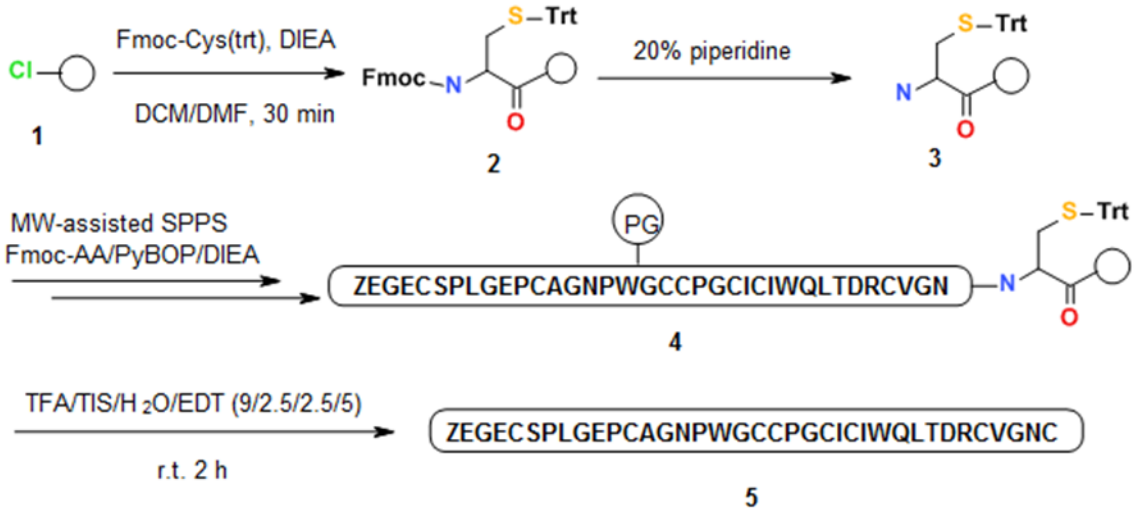



**Figure 2.3. Scheme for the synthesis for coffeetide cC1.** The peptide was synthesized through SPPS using Fmoc-chemistry. Chemical modification of resins (step 1, 2&3) was performed manually, and the subsequent elongation of the peptide (step 4) was done by an automated microwave peptide synthesizer. TFA/TIS/H2O/EDT with a ratio of 90/2.5/2.5/5 was performed to do final deprotection before adding diethyl ether to precipitate crude peptide.

### 2.9.2. Oxidative folding

18 folding conditions were investigated using different redox reagents, times, co-solvents and concentrations of redox reagent under the same environment of pH 8 in ammonium bicarbonate buffer. Two pairs of redox reagents, including reduced glutathione (GSH)/oxidized glutathione (GSSG) and cysteamine/cystamine, were used. For each condition, 0.38 mg of cC1 was

dissolved in a total volume of 100 µL of solvent to a final concentration of 1 mM. At 24 h intervals, 2 M hydrochloride acid was used to quench the reaction. The folding process was monitored by UPLC with a gradient from 20% to 80% buffer B (ACN/ 0.1% TFA) over 18 min. By comparing the peak area before and after the folding reaction, the folding yield of each condition was established. The native peptide was mixed with the folded peptide and subjected to UPLC to check the co-elution result.

## 2.10. EST-Based data mining

### 2.10.1. Translated nucleotide-based search for putative cysteine-rich peptides

Basic local alignment search tool for translated nucleotide (BLASTn) was employed to search for ESTs encoding putative CRP precursors in two databases, the National Center for Biotechnology Information (NCBI) [229] and the 1000 Plants Project (OneKP) [230]. The values of 1000 and 10 were set as the maximum target sequences and expected threshold, respectively. The criteria set to select sequences manually are as follows: (1) The open reading frame must contain a specified start (ATG) and stop (TAA, TAG, and TGA) codons; (2) No untranslated amino acid residues such as X should be contained in the translated amino acid sequence; and (3) Six, eight or ten cysteine residues should be contained in the mature peptide. Subsequently, the sequences were submitted to SignalP 4.0 [212] to identify the cleavage site of the signal peptide and aligned using Bioedit [231]. Replicate sequences from the same plants or different species of the same genus sharing an identical full-length precursor were identified using ClustalW Phylogeny and deleted from the dataset.

### 2.10.2. Data analysis

The data were analyzed by Student's t-test. The standard form of the results was described as the mean ± standard error of the mean (SEM). A p-values $< 0.05$ was considered as statically significant. The identity and similarity of the peptides were compared using EMBOSS Water Pairwise Sequence Alignment [232]. The sequence logo was generated using WebLogo 3 [213]. The aligned precursor sequences were analyzed using a neighbor-joining clustering algorithm by MEGA 6.0 [233]. The phylogenetic tree was constructed using a bootstrap test of 1,000 replications. The phylogenetic tree was displayed using iTOL v3 [234].

## 2.11. Data pre-processing for multivariate analysis

### 2.11.1. MALDI-TOF MS data matrix

The spectrums obtained from MALDI-TOF MS were converted to data points in ASCII format using Data Explorer V4.9 (Applied Biosystem, MA, USA). The mean of the three spectrums obtained from each sample was used as the final data set. Hence, a MALDI-TOF MS data matrix comprising 91 rows (sample numbers) and 30935 columns (number of data points per sample) was generated. The MALDI-TOF MS data matrix was pre-processing by several techniques. Correlation optimized warping (COW), which is a peak alignment procedure, was applied to the matrix. In addition, an algorithm for choosing the ideal reference spectrum and the most optimal segment length and the slack number was performed to optimize the procedure [235]. Other methods such as standard normal variate (SNV) and mean cantering for data pre-processing were used to remove slope variation among spectrums, and column mean from each variable of the corresponding column, respectively.

### 2.11.2. UPLC data matrix

Retention time shift is often observed in chromatographic analysis. The occurrence of this problem may due to the changes in the mobile phase, operator handling, and instrumental instability. Thus, data pre-processing is needed for the UPLC analysis. COW was applied to align the peaks in the UPLC data matrix while baseline elevation was eliminated by subtracting with a blank chromatogram. SNV and mean centering were performed before classification analysis.

## 2.12. Multivariate analyses

### 2.12.1. Unsupervised multivariate analyses

#### 2.12.1.1. Principal component analysis (PCA)

PCA is an unsupervised multivariate analysis technique to divide a great number of variables into different groups in predictive models and exploratory data. Generally, PCA reduces the large data sets by projecting them onto principal components (PCs), which are at lower dimensions. It aims to determine the best trend of the data using a limited number of PCs [236]. In this study, prior to model calibration, PCA was performed on MALDI-TOF MS and UPLC data matrix for outliner determination and sample classification. The optimal PCs for the MALDI-TOF MS and UPLC matrices were both determined as 3.

#### 2.12.1.2. Hierarchical cluster analysis (HCA)

HCA is an unsupervised multivariate method used for natural grouping among samples characterized by their features. Strategies for HCA can be classified into two main categories: agglomerative and partitional. Agglomerative methods usually start with each object being its own cluster and pairs of clusters are merged hierarchically into larger ones, while partitional method begins with a single cluster containing all objects and splits existing clusters into smaller ones [237]. Usually, agglomerative methods are more commonly used in chemometric studies. There are six agglomerative methods which include the nearest neighbor, furthest neighbor, pair-group average, centroid, median, and Ward's method, based on their inter-cluster distance and linkage rules. In this study, Ward's method and squared Euclidean distance were used, which minimized the numbers of clusters and the deviation of any two clusters for each step [238].

### 2.12.2. Supervised multivariate analyses

#### 2.12.2.1. Partial least square-discriminant analysis (PLS-DA)

The MALDI-TOF and UPLC data matrix were divided into a calibration set and validation set based on the Kennard-Stone (K-S) algorithm [239], which was applied to the RH and RA samples separately. Generally, with the greatest deviation, 60% of the samples consist of the calibration set, whereas the validation dataset is comprised of the remaining 40% of the samples. The performance of the PLS-DA model was assessed and compared after various pre-processing steps.

PLS-DA is a supervised analysis method that maximizes the separation between predefined classes rather than explaining the variation with each class. In PLS-DA, the data matrices were projected onto latent variables (LVs) to maximize the covariance between the original matrix and the predefined response classes [240]. The predictions from a PLS-DA model are qualitative and usually coded in vectors [241]. The UPLC and MALDI-TOF MS data matrices were divided into two subgroups, RH and RA, based on their botanical characteristics. Consequently, RH and RA were represented by vector numbers 0 and 1. A $y$ predicted value for every unknown sample was calculated, and the value 1 indicates that the corresponding sample belongs to the pre-defined class, whereas a 0 value means that the sample was rejected as a member of the class. The optimal LVs were determined as 2 and 1 in the MALDI-TOF MS and UPLC data matrix, respectively.

#### 2.12.2.2. K-nearest neighbors (KNN)

KNN is an instance-based algorithm that utilizes the distance between samples in the p-space as its primary criterion. The classification was performed based on the Euclidean distance between samples. Unknown samples were classified based on their distance from other data points nearest to them and the majority vote of the neighbors. K-value, the optimal number of the nearest neighbor, was established by the leave-one-out cross-validation (LOO-CV) [242]. In both UPLC and MALDI-TOF MS matrices, the optimal K values were determined as 3.

#### 2.12.2.3. Classification and regression tree (CART)

Decision trees are generally used to create a model which predicts the value of a target based on the values of independent variables. CART is a nonparametric decision tree that produces either classification of regression trees depending on whether the variables are categorical or continuous, respectively [243]. Since there were only two classes involved, there was no optimal tree size.

#### 2.12.2.4. Soft independent modeling of class analogy (SIMCA)

SIMCA is a supervised model that minimizes the hypotheses about the linearity of relationships between samples and predefined classes [149]. To build the model, each class (RH and RA) needs to be analyzed using PCA separately. Hence, a principal component model was used to account for most of the variation within each class. Because the number of PCs retained for each class is usually different, the cross-validation set was used to select the optimal numbers of PC. To classify an unknown sample, its matrix was projected to each established PCA model and the residual distance was calculated. By comparing the residual variance of the unknown sample to the average residual variance of the PCA model from each class, the unknown sample was able to be categorized [244]. The optimal PCs for the RA and RH PCA model were determined to be 5 and 1 in the MALDI-TOF MS and UPLC data matrix, respectively.

#### 2.12.2.5. Support vector machine-discriminant analysis (SVM-DA)

SVM-DA is a supervised classification method that is commonly used for binary classification. In SVM-DA, samples are represented by points in two classes. Based on this, a hyperplane boundary that separates all points to place the majority in the same class was calculated [245]. This technique aims to determine the optimal hyperplane that can maximize the distance between the two separated classes [246]. In this study, the X-block compression was set at “none” and the probability estimation was set at “on”.

### 2.12.3. Classification model performance evaluation

All models were cross-validated with Venetian blind and split into ten blocks [247]. Confusion matrices were used to evaluate and compare the performances of the classification models. In a confusion matrix, the error rate (ER), non-error rate (NER), sensitivity and specificity were calculated using the following equations:

$$Sensitivity = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Negative} ;$$

$$Specificity = \frac{Number\ of\ True\ Negative}{Number\ of\ True\ Negative + Number\ of\ False\ Positive} ;$$

$$NER = \frac{Sensitivity + Specificity}{Number\ of\ Class} ;$$

$$ER = (1 - NER);$$

True Positive: RA samples correctly classified as RA; True Negative: RH samples correctly classified as RH; False Negative: RA samples wrongly classified as RH; False Negative: RH samples wrongly classified as RA.

### 2.12.4. Software

Data processing was performed on MATLAB R2017b (The MathWorks, MA, USA). Classification modeling was analyzed on PLS toolbox version 8.5 (Eigenvector Research, WA, USA) and classification toolbox version 5.1. The K-S and COW algorithms were developed by Daszykowski *et al.* [239] and Skov *et al.* [235], respectively.

# Chapter 3 Cysteine-rich peptide fingerprinting for herbal analysis: A rapid method to differentiate Radix Astragali from Radix Hedysarum

## 3.1. Introduction

Misidentification of plant species and the presence of adulterants are major concerns in the quality control of herbal products [138]. The confusion in the identity of the herbs may be caused by several reasons: similar morphology, similar name, multiple sources, the presence of counterfeit and adulterants. A traditional way to authenticate herbal products is to quantify the major or most abundant compounds using chromatographic methods. Often, a single chemical marker is used as an indicator for quality assessment [248]. However, this approach is not able to show the complexity of the herb. Also, the chemical marker might not be unique to one single herb. Another method employed for authentication is DNA barcoding, which is based on variations in the sequence of short standard DNA region(s). However, the application of DNA barcoding has limitations because the DNA region in one plant is identical across many species, making it not a unique pattern [249]. To overcome these limitations, fingerprint analysis, which reflects the unique pattern of chemical compositions in an herb, was adopted by many global regulatory authorities [163-165]. These chemical fingerprints can be obtained by spectroscopic or chromatographic techniques, such as high-performance liquid chromatography (HPLC), gas chromatography, capillary electrophoresis, thin-layer chromatography (TLC) and Raman spectroscopy [250, 251]. Among these techniques, chromatographic fingerprints obtained from HPLC are widely used because of its high accuracy, sensitivity, and reproducibility. However, laborious sample preparation, relatively long analytical run-times, and the large volume of organic solvents consumption in HPLC hinder its application as a high-throughput screening technique [149].

Mass spectrometry is an analytical technique used to detect the mass-to-charge ratio (*m/z*) of ions derived from analytes molecules, which can provide both qualitative and quantitative information about samples [252]. The ability of mass spectrometry for analyzing non-volatile, thermally labile, intact and large biomolecules is due to the development of soft ionization techniques such as Matrix-Assisted Laser Desorption/Ionization (MALDI) and Electrospray Ionization (ESI) techniques [253]. Both ionization techniques provide a simple and efficient way for the routine mass spectroscopic analysis of peptides and proteins. However, MALDI is more robust in terms of their higher tolerance for salts than ESI. In addition, MALDI usually produces singly charges ions and thus shows lower spectral complexity than ESI [254].

Coupled with time-of-flight mass spectrometry (TOF MS), MALDI-TOF MS has been widely applied in various fields especially for the quantification of large molecules without prior chromatographic separation [255]. Fingerprint analysis has been employed to identify fungi species such as *Neoscytalidium.* and *Penicillium* based on MALDI-TOF MS technique [256]. Furthermore, it has been applied to the quality control of food products such as Brazil grape species [257] and *Campania* white wines. Peptidomic profiles derived from wine protein tryptic digests showed the unique fingerprinting of the samples [258]. Compared to HPLC, being faster and simpler, MALDI-TOF MS also has a broader detection range, has a higher tolerance to salts and buffers, and requires minimal amounts of analytes [255].

The roots of *A. membranaceus* (Radix Astragali, RA), known as Huang Qi in Chinese, are natural food supplements and popular herbal medicines used in traditional Chinese medicine (TCM) to increase overall vitality, treat diabetes and metabolic diseases [259]. However, these roots are often misidentified or substituted by the roots of *Hedysarum polybotrys* Hand.-Mazz. (Radix Hedysarum, RH), which has a similar morphology to RA. In addition, RH possesses a similar Chinese name (Hong Qi) to RA (Huang Qi). Although they share high similarity, the chemical constituents present in both species are different [260]. It is reported that RH has been shown to possess a weaker antidiabetic activity *in vivo* compared to RA [261]. In clinical practice, RH is employed to disperse swelling by external use, and incorrect use of RH in patients with diabetes may lead to fatal outcomes [262].

Irrespective of the authentication methods, small-molecule secondary metabolites with molecular weight <1 kDa are employed as major chemical markers. According to Pharmacopoeia of the People's Republic of China (PPRC) [164] and Hong Kong Chinese Materia Medica Standards (Phase III) (HKCMMS Volume I and VIII, Hong Kong), the standard chemical markers of RH are ononin and formononetin while calycosin-7-O-β-d-glucoside is the standard compound to authenticate RA. Previous studies have investigated different second metabolites such as saponins, flavonoids, or polysaccharides in RA and RH by HPLC [263] and high-speed countercurrent chromatography [264]. Additionally, DNA barcoding based on internal transcribed spacers [263] and 5S-rRNA spacer domains [265], have been used for identifying RA and RH.

Cysteine-rich peptides (CRPs) are generally hyperstable. They have well-defined structures stabilized by three or more cross-linking disulfide bridges that render them resistant to thermal, chemical, and enzymatic degradation [266]. However, the chemical spaces based on the molecular mass of CRPs of the plant-derived natural products have not been seriously used as authentication standards [27]. The hyper-stability of CRPs is essential as putative compounds

in herbal medicine because they generally require decoction or other processing steps., Our laboratory is particularly interested in CRPs with molecular weights ranging from 2 to 6 kDa, and which are readily detected by MALDI-TOF MS, a chemical space which is uncluttered by small-molecule metabolites. Another advantage of CRPs is that they are well-annotated because they are grouped into families, such as thionins, defensins, hevein-like, and knottin-type peptides based on their cysteine motifs and disulfide connectivity [23]. In addition, our studies on CRPs showed that they are widely distributed *in Planta* and could be used for authentication [20, 21, 23-27, 57, 59, 266]. We hypothesized that the unique chemical space of CRPs is suitable for discriminating different plant species, and thus can authenticate RA and differentiate it from its substitute species RH.

In addition to the instrumental analyses, multivariate data analysis techniques were introduced for quality control because of the complexity of herbal medicines to detect minor differences between closely related species. Instead of relying on the comparison to a reference compound or on quantifying a particular chemical marker, multivariate analyses usually combine mathematical and statistical techniques to increase the understanding of chemical data and also to correlate the quality parameters of physical properties of the analytical instrument data [171]. The pattern recognition models in multivariate analyses can improve the overall classification efficiency based on the chromatographic or spectroscopic fingerprint obtained.

In this chapter, I described a rapid and general method to authenticate herbs and herbal products, which is designated as CRP fingerprinting. In the case of RA and RH, we used MATLAB and classification built-in tools to extract and analyze the spectra from MALDI-TOF MS and chromatograms from UPLC. Our results suggest that this combination can provide a powerful tool for differentiating closely related plant species and herbal products.

## 3.2. Results and Discussion

### 3.2.1. CRP fingerprinting of herbs and herbal products

A mass spectrometry-driven method was applied for screening putative CRPs in herbs and herbal products in our laboratory. The small-scale screening revealed clusters of peptides within a mass range from 2 to 6 kDa are presented in 100 herbs and herbal products. To show that they are CRPs, samples were treated with a disulfide-reducing agent followed by an *S*-alkylating agent, a procedure commonly used in our laboratory [12, 26, 267]. A mass shift before and after *S*-reduction and *S*-alkylation, which results in a mass increment of 58 Da for each cysteine confirmed the presence of CRPs [25].

Usually, most authentication of plants is based on visual analysis of their morphological features, which are subjective and not accurate when many plants share similar morphological characteristics. Our results showed that by employing CRP fingerprints as chemical markers, it is able to distinguish between plants with similar morphology. For example, *Portulaca oleracea* and *Portulaca grandiflora* are two herbs with similar morphology. Using our screening procedure, we obtained the unique CRP fingerprints of these two plants, and the presence of these 'marker' peaks allows us to establish the identity of each plant (Figure 3.1). In addition, by employing CRP fingerprinting, it can distinguish between plants with similar chemical composition. For example, oleanolic acid is expressed in two herbs, Achyranthes *bidentate* and *Clematis chinensis.* By employing oleanolic acid as a chemical marker, according to the Chinese Pharmacopeia, it is difficult to differentiate these two species. In contrast, our result revealed that the CRP fingerprints present in *Achyranthes bidentate* are distinguishable from *Clematis chinensis* (Figure 3.2). Similar results can be observed in plants from same plant families (Figure 3.3), which means CRP fingerprinting can differentiate species regardless of their origins.

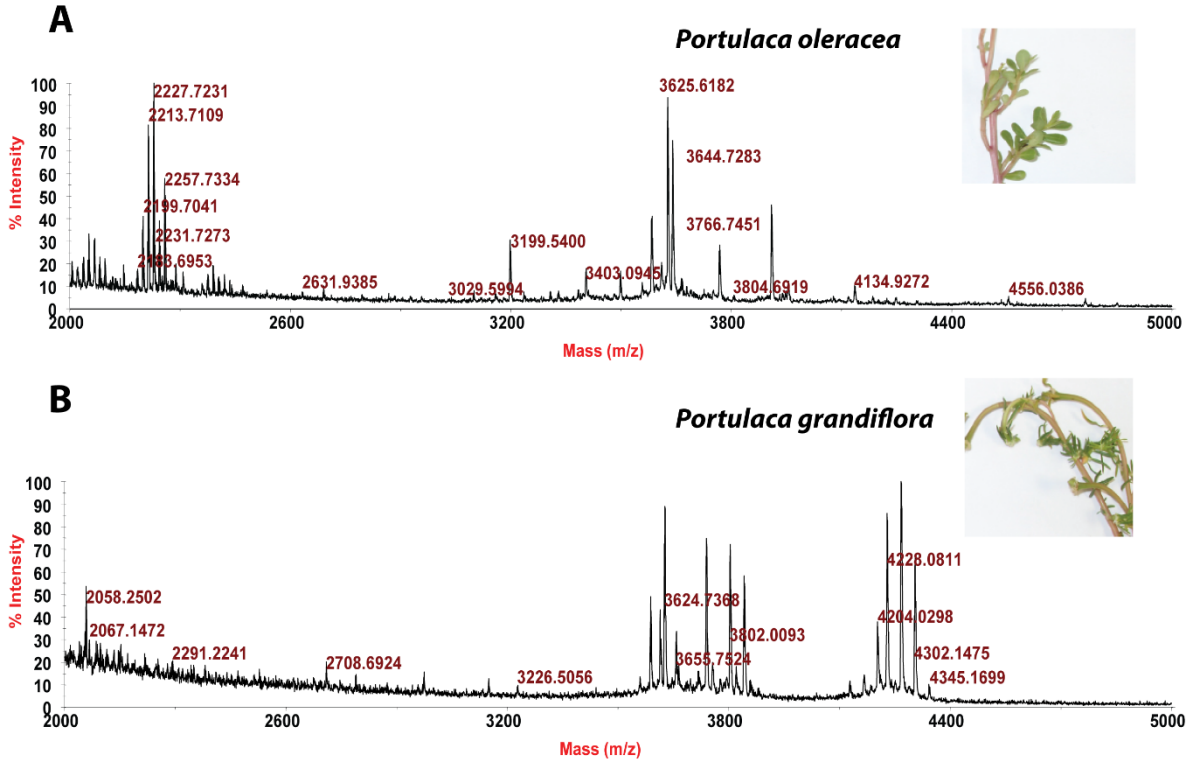



**Figure 3.1. MALDI-TOF MS profiles.** Mass spectra of two plants with similar morphology: (A) *Portulaca oleracea* and (B) *Portulaca grandiflora.*

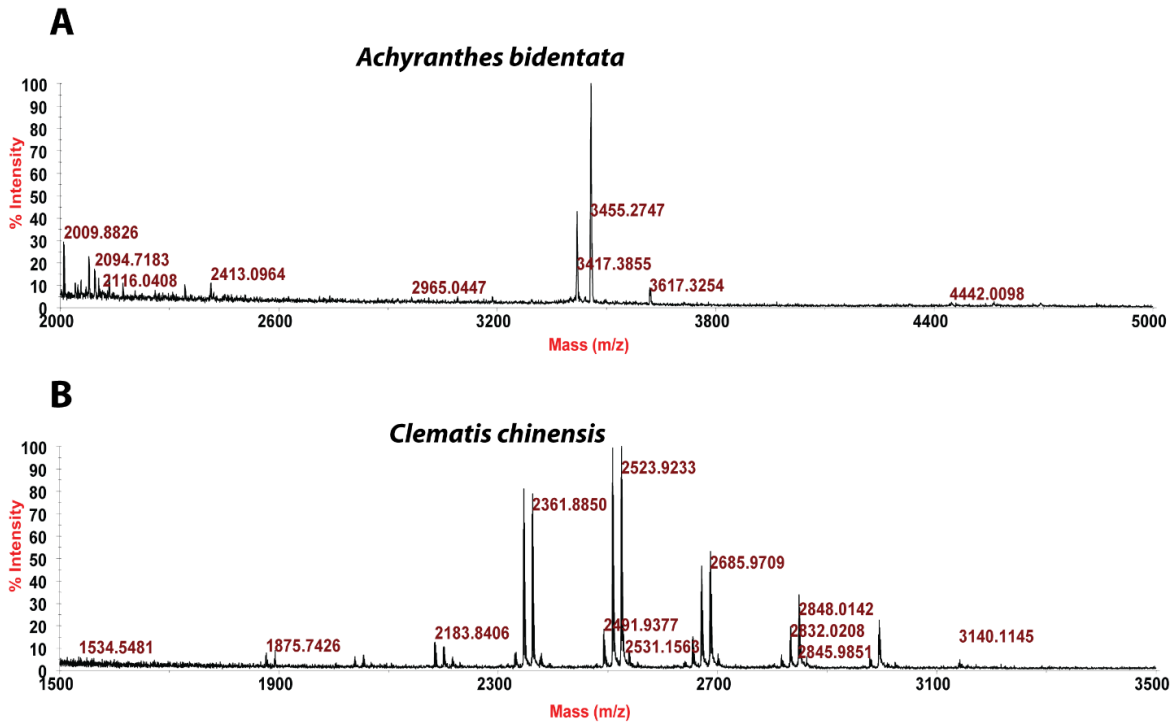



**Figure 3.2. MALDI-TOF MS profiles.** Mass spectra of two plants with similar chemical composition: (A) *Achyanthes bidentata* and (B) *Clematis chinensis*.

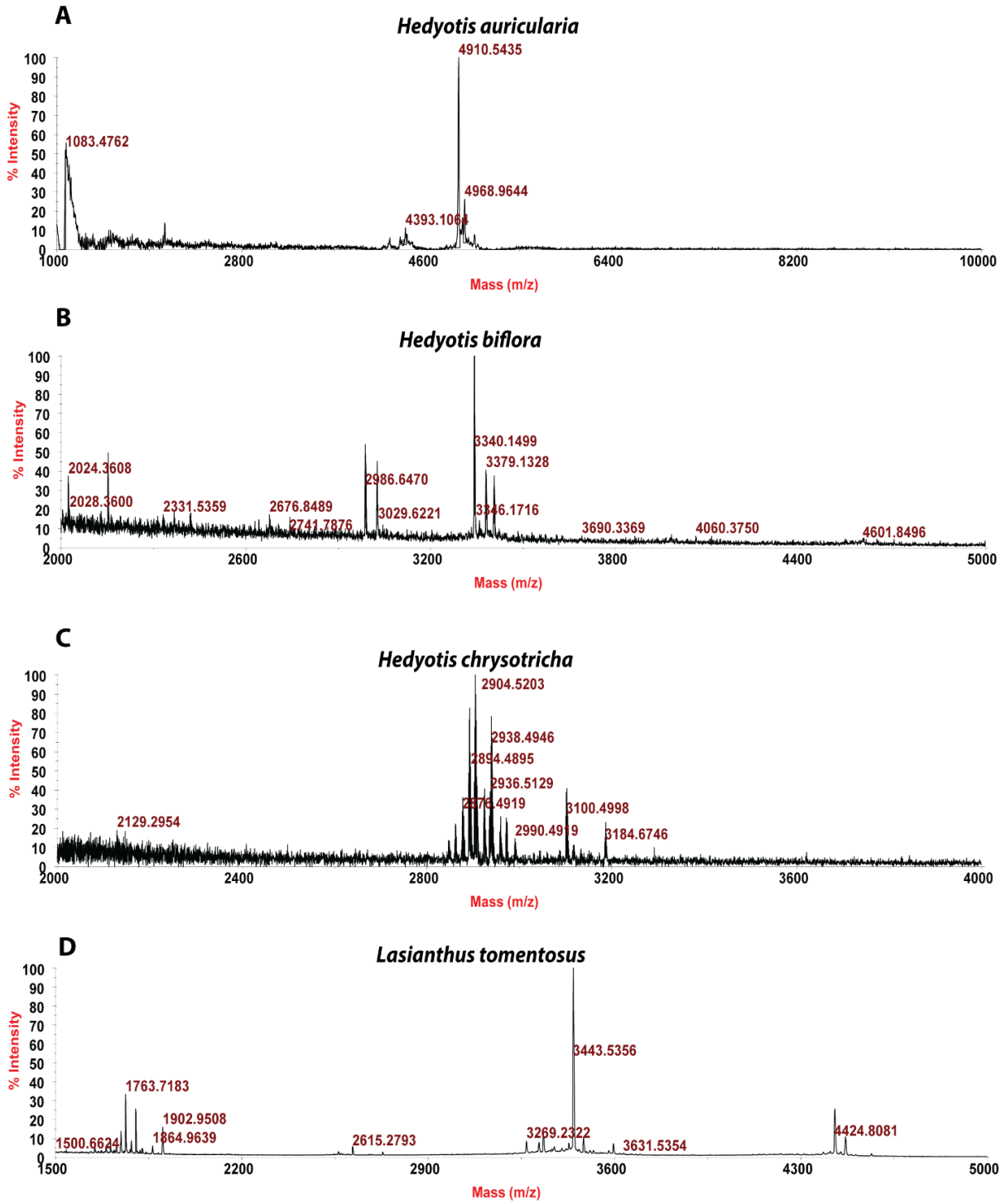



**Figure 3.3. MALDI-TOF MS profiles.** Mass spectra of four plants from Rubiaceae family: (A) *Hedyotis auricularia,* (B) *Hedyotis biflora, (C) Hedyotis chrysotricha* and (D) *Lasianthus tomentosus.*

### 3.2.2. CRP fingerprinting of RH and RA

To further validate the method, 50 RH and 41 RA samples were used as examples in this study. The small-scale screening revealed clusters of peptides with a mass range from 3 to 5 kDa in both samples (Figure 3.4). In RA samples, the identification and characterization of two major peptides aM1 and bM1 have been reported in the previous study and will be described in the next chapter, with sequences of VDCSGACSPFEVPPCGSRDCRCIPIGLVVGFCIYPTG and CEKPSKFFSGPCIGSSGKTQCAYLCRRGEGLQDGNCKGLKCVCAC, respectively [1]. Similarly, two major peptide peaks with m/z of 3944.3 Da and 4780.1 Da were observed in RH and were designated as hedytide hP1 and hP2. Their sequences were determined by the MALDI-TOF MS/MS. After trypsin digestion, hP1 yields two fragments with m/z of 3416 and 550 Da. The amino acid assignments of the two digested fragments were obtained based on the *b*- and *y*- ions detected during the MS/MS fragmentation (Figure 3.5), which gave the full sequence of hP1 (QGCNGPCTPFEQPPCGIQSCRCFPEVLFFGRCSTPSG). *De novo* sequencing of hedytide hP2 was performed in the same manner, which gave the full sequence of CEKGSEFFVGACRYSEGTQQCATLCSRGEGLQGGKCKGVRCYCSC (Figure 3.6).

Plant CRPs are divided into different families such as defensins, knottins, hevein-like peptides, and thionins, based on their different sequences, cysteine spacing, and disulfide connectivity. Astratide aM1 was shown to be a pea albumin 1 b (PA1b)-like peptide, whereas bM1 is a plant defensin [1]. It can be observed that both aM1 and hP1 are 37 amino acids in length and contain six cysteines. Sequence comparison revealed that they share a 65.7% sequence similarity and comprise the same cysteine motif of C-$X_3$-C-$X_7$-C-$X_4$-C-X-C-$X_9$-C. Similarly, both bM1 and hP2 are 45 amino acid in length and contain eight cysteines. They contain a similar cysteine motif of C-$X_{10}$-C-$X_8$-C-$X_3$-C-$X_{10}$-C-$X_4$-C-X-C-X-C and have more than 50% sequence similarity. Based on the cysteine motif and the sequence identity, we concluded that hP1 is a PA1b-like peptide similar to aM1, and hP2 a plant defensin similar to bM1.

These intra-family sequence similarity and variability were frequently observed in the legume family. It was reported that PA1b-like peptides present in more than 18 spices from the Fabaceae family, with percentages of sequence identity from 61% to 86.1% [268]. In addition, plant defensins have been identified in more than 10 species from the Fabaceae family with 46.8 − 86.7% sequence identity [1]. Although these CRPs belong to the same CRP subfamily, their sequence variability enables them unique for distinguishing one species from another. Furthermore, the ability to withstand harsh conditions during the processing stage of crude herbal medicine makes CRPs as suitable chemical markers for the authentication of RA and RH.

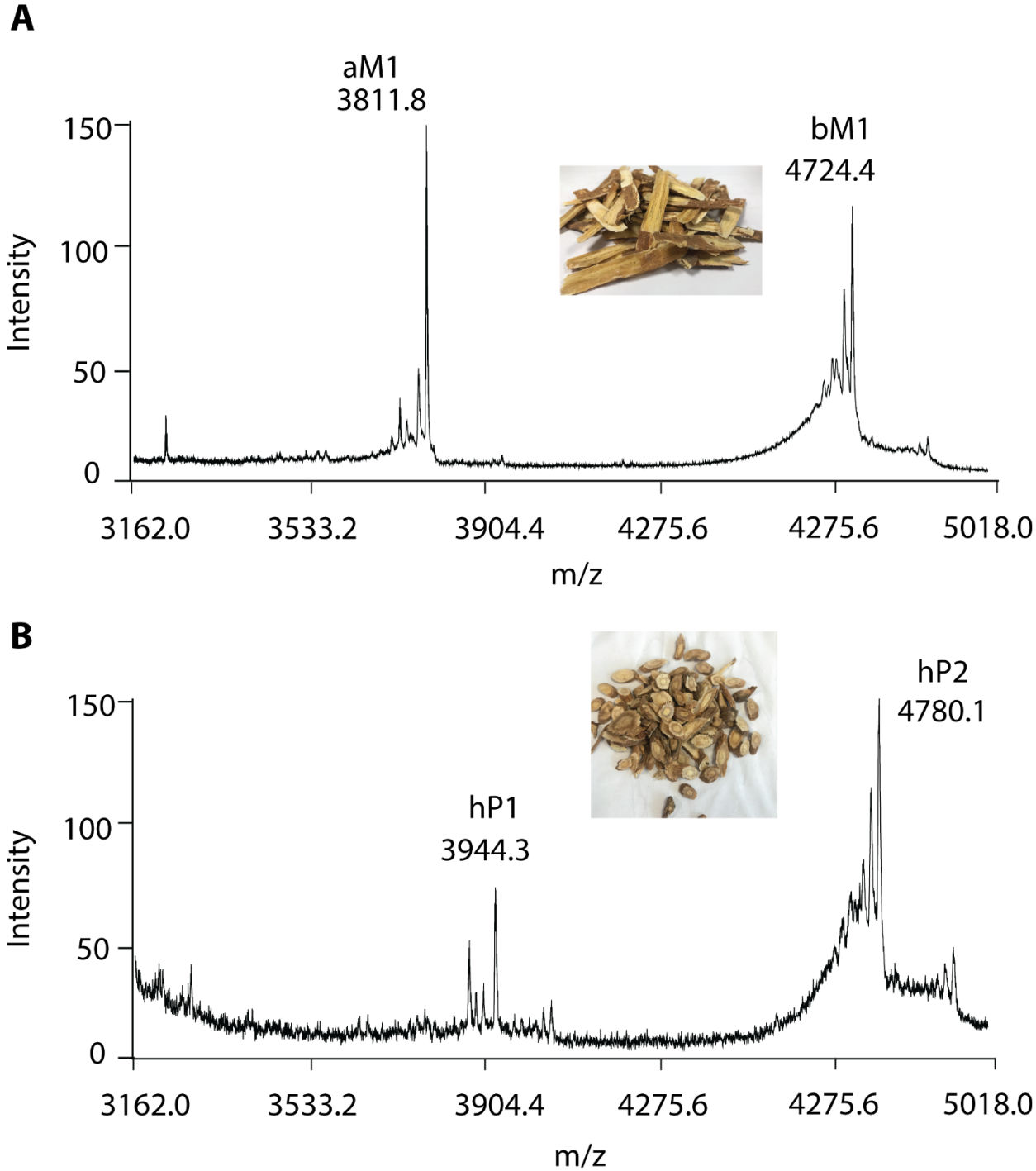



**Figure 3.4. MALDI-TOF MS profile of (A) RA and (B) RH samples.** Two major CRPs are designated as aM1 and bM1, with a molecular weight of 3811.8 and 4724.4 Da in RA samples, respectively. Similarly, with molecular weights of 3944.3 and 4780.1 Da, two major CRPs in RH samples were designated as hP1 and hP2, respectively.

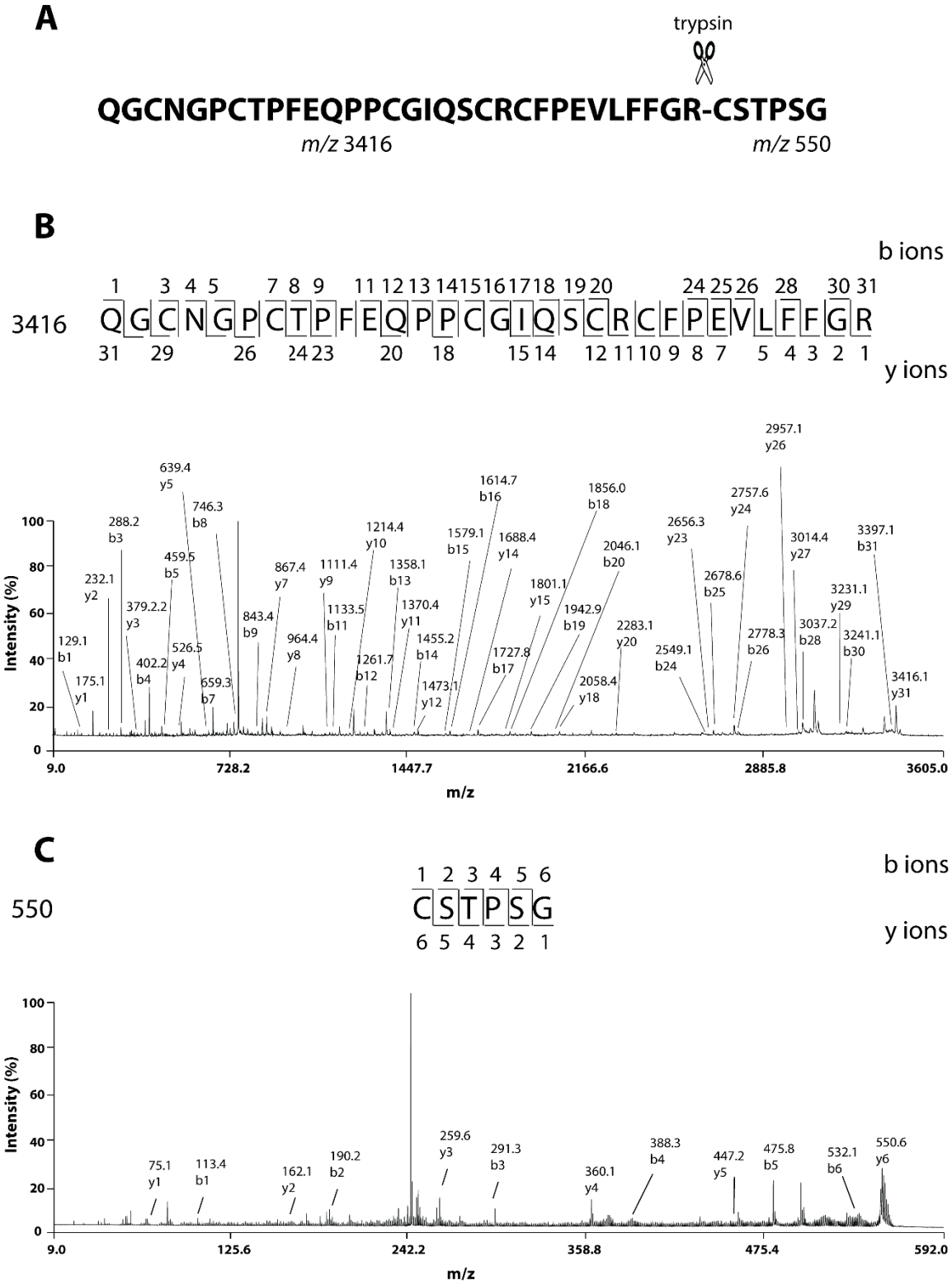



**Figure 3.5. MS/MS sequencing of hedytide hP1.** (A) Trypsin digestion of hedytide hP1 which yielded three fragments with m/z values 3416 and 550 Da; (B−C) MS/ MS spectra of digested fragments.

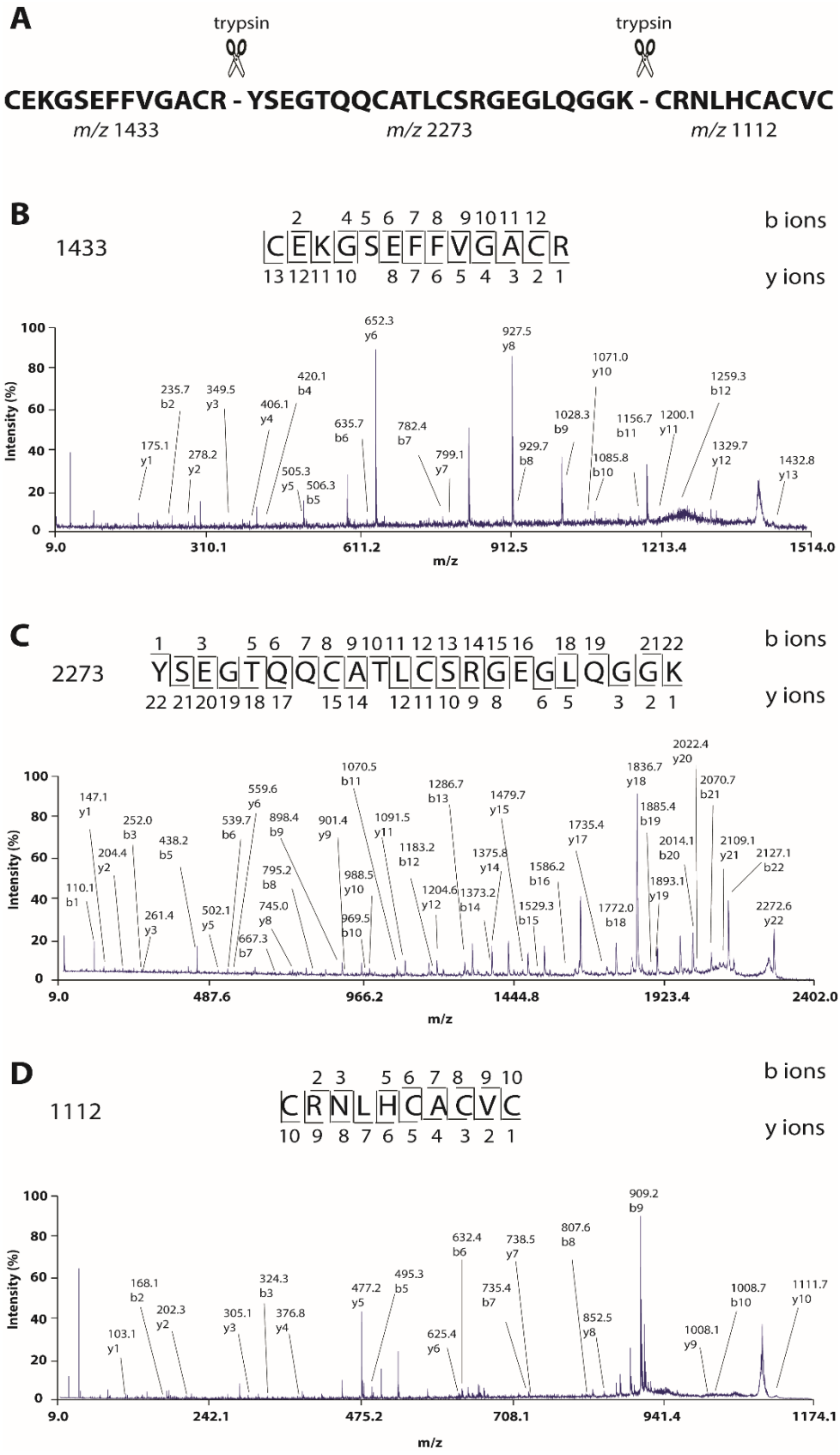



**Figure 3.6. MS/MS sequencing of hedytide hP2.** (A) Trypsin digestion of hedytide hP2 which yielded three fragments with m/z values 1433, 2272, and 1121 Da; (B−D) MS/ MS spectra of digested fragments.

### 3.2.3. UPLC method validation

To evaluate the linearity and range of the chromatographic method, LOD, LOQ, and calibration curve parameters of each standard compound were summarized in Table 3.1. It is observed that the five standard compounds obtained low LOD and LOQ values of ≤ 0.15 and 0.43 μg/mL, respectively. In addition, the high correlation coefficients ($r^2 \geq 0.9990$) and wide linear range (0.02 – 3000 μg/mL) indicated the highly correlated relationship between reference compounds and peak area.

**Table 3.1. The calibration curve parameters, LOD, LOQ of five standard compounds.**

| | Linear range (μg/mL) | Correlation Coefficient ($r^2$) | Slope | y-intercept | LOD (μg/mL) | LOQ (μg/mL) |
|---|---|---|---|---|---|---|
| calycosin | 0.02-800 | 0.9990 | 51896 | 165802 | 0.01 | 0.04 |
| calycosin-7-O-beta-D-glucoside | 0.24-1200 | 0.9990 | 39371 | -18677 | 0.11 | 0.32 |
| medicarpin | 0.49-1000 | 0.9990 | 28886 | 8705 | 0.14 | 0.43 |
| ononin | 0.24-3000 | 0.9993 | 23079 | 389359 | 0.10 | 0.31 |
| formononetin | 0.06-250 | 0.9991 | 45118 | 10206 | 0.02 | 0.07 |

In terms of recovery and precision, the intra-day and inter-day analyses for standard compounds at low, medium, and high concentrations were shown in Table 3.2. The average RSDs of intra-day LQC, MQC and HQC were 0.66%, 0.48% and 0.30%, whereas 1.42%, 1.09% and 0.93% is the average for inter-day, respectively. Moreover, intra-day recoveries at LQC, MQC, and HQC were calculated as 2.11%, 2.90% and 2.09%, whereas 1.73%, 2.73% and 1.50%, were obtained as average inter-day recoveries, respectively. The results showed that the UPLC method we developed possesses good precision and reproducibility.

**Table 3.2. Validation of the intra- and inter-day recoveries of five standard compounds at low, medium and high concentrations.**

| Compounds | Spiked concentration (µg/mL) | Intra-day (n = 6) | | | Inter-day (n = 18) | | |
|---|---|---|---|---|---|---|---|
| | | Observed concentration (µ/mL)[a] | Precision RSD (%)[b] | Recovery (%)[c] | Observed concentration (µ/mL)[a] | Precision RSD (%)[b] | Recovery (%)[c] |
| calycosin-7-O-beta-D-glucoside | 250 | 252.995 ± 0.914 | 0.361 | 1.197 | 257.999 ± 3.793 | 1.470 | 3.199 |
| | 500 | 519.093 ± 2.716 | 0.523 | 3.818 | 520.266 ± 2.424 | 0.466 | 4.053 |
| | 1000 | 967.190 ± 2.325 | 0.240 | -3.281 | 975.803 ± 7.011 | 0.718 | -2.451 |
| formononetin | 50 | 50.963 ± 0.876 | 1.719 | 1.927 | 50.081 ± 1.116 | 2.227 | 0.162 |
| | 100 | 98.997 ± 1.349 | 1.362 | -1.002 | 101.801 ± 2.857 | 2.806 | 1.801 |
| | 200 | 189.891 ± 1.312 | 0.691 | -5.084 | 193.855 ± 4.552 | 2.348 | -3.072 |
| calycosin | 160 | 161.065 ± 0.898 | 0.558 | 0.666 | 160.424 ± 0.999 | 0.623 | 0.265 |
| | 320 | 316.580 ± 0.800 | 0.253 | -1.069 | 317.512 ± 1.646 | 0.518 | -0.777 |
| | 640 | 637.987 ± 2.264 | 0.355 | -0.315 | 636.471 ± 3.514 | 0.552 | -0.551 |
| medicarpin | 200 | 209.377 ± 0.941 | 0.449 | 4.688 | 204.941 ± 4.969 | 2.425 | 2.470 |
| | 400 | 372.762 ± 0.470 | 0.126 | -6.810 | 376.282 ± 3.831 | 1.018 | -5.929 |
| | 800 | 789.892 ± 1.051 | 0.133 | -1.271 | 790.642 ± 3.315 | 0.419 | -1.169 |
| ononin | 375 | 382.684 ± 0.762 | 0.199 | 2.049 | 384.5343 ± 2.083 | 0.542 | 2.542 |
| | 750 | 763.594 ± 0.921 | 0.121 | 1.812 | 758.200 ± 4.755 | 0.627 | 1.093 |
| | 1500 | 1507.782 ± 1.069 | 0.071 | 0.518 | 1496.51 ± 9.334 | 0.623 | -0.233 |

[a] Mean ± standard deviation (SD).
[b] Relative standard deviation (RSD) % = (SD/mean) × 100.
[c] Recovery % = [(mean observed concentration−spiked concentration)/spiked concentration] × 100

### 3.2.4. UPLC fingerprinting

According to the method and monographs recorded in PPRC [164] and HKCMMS (Volume I and VIII, Hong Kong), the quality control of RH and RA samples were accessed based on the UPLC analysis of five standard compounds: medicarpin, formononetin, calycosin-7-O-beta-D-glucoside, calycosin, and ononin [269]. The chromatographic fingerprints of RA and RH samples were obtained by injecting their methanolic extraction to UPLC, whereas the same methods were applied for the five standard compounds (Figure 3.7). The result showed that the retention time of calycosin-7-O-beta-D-glucoside, ononin, calycosin, formononetin, and medicarpin were 10.0, 15.0, 17.5, 23.5, and 27.5 min, respectively. The peaks eluted from both species were labeled with numbers from 1 to 11, and peak 1, 3, 5, 8 and 11 coelute at the same time as calycosin-7-O-beta-D-glucoside, ononin, calycosin, formononetin, and medicarpin, respectively. Thus, it revealed that ononin, calycosin, and formononetin could be found in both species. The major difference between the two species is that calycosin-7-O-beta-D-glucoside is unique in RA samples, whereas medicarpin was only observed in RH. The results are in agreement with the previous study that formononetin, ononin, calycosin, formononetin-7-O-Dglucoside-6"-O-malonate and soyasaponin present in both RH and RA, while medicarpin was unique in RH species [270].

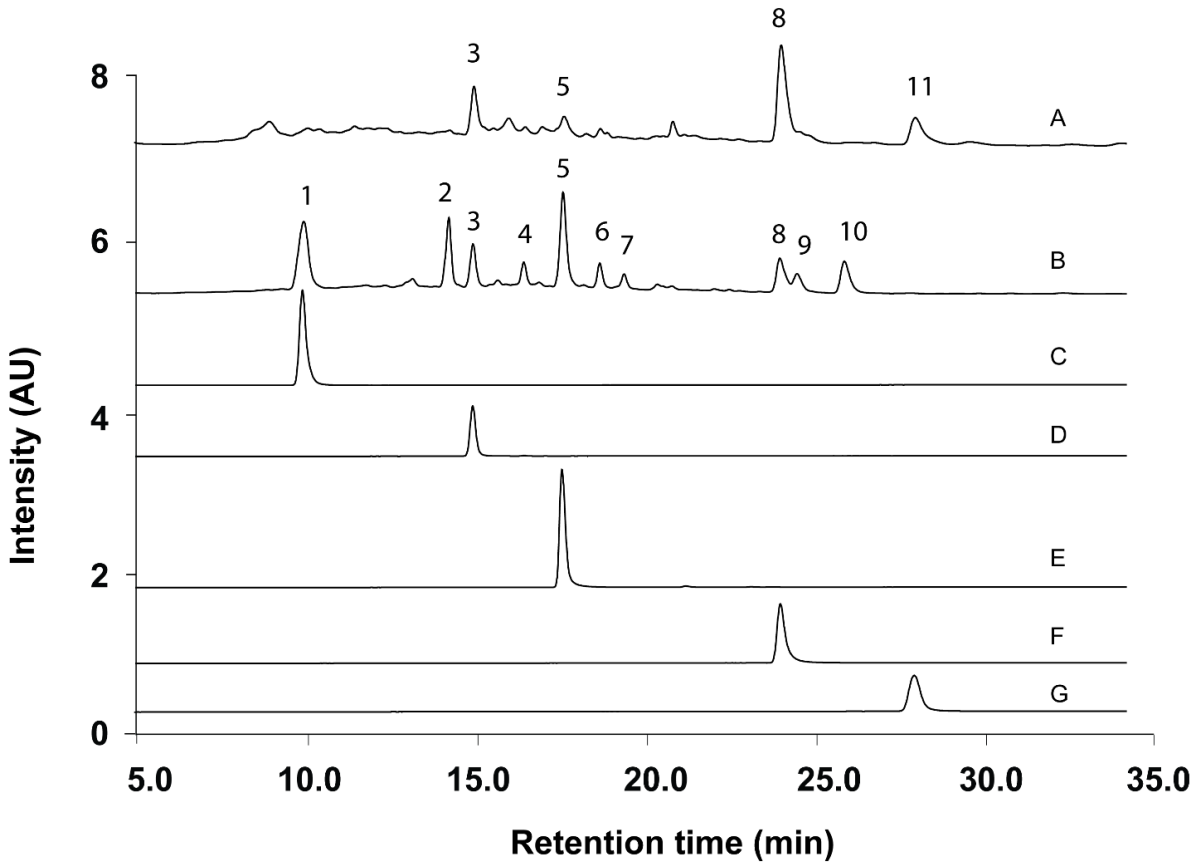


**Figure 3.7. Representative UPLC Chromatograms of samples and standards.** (A) Methanolic extract of RH. (B) Methanolic extract of RA. (C) Calycosin-7-O-beta-D-glucoside. (D) Ononin. (E) Calycosin. (F) Formononetin. (G) Medicarpin.

### 3.2.5. Data pre-processing

#### 3.2.5.1. Peak alignment

Due to the inconsistency and misalignment of the analytical instruments, COW was employed to preprocess both MALDI-TOF MS and UPLC data matrices. To perform the pre-processing, reference chromatogram, segment length, and slack numbers are needed. The parameters were optimized by the method proposed by Skov *et al*.[235] and summarized in Table 3.3. The chromatograms of forty RH samples and fifty-one RA samples before and after peak alignment were compared in Figure 3.8, whereas Figure 3.9 showed the raw MALDI-TOF MS spectra and the spectra after peak alignment.

**Table 3.3. COW pre-processing parameters and reference chromatograms or mass spectrum for UPLC and MALDI-TOF MS data matrices.**

| Data matrices | Sample | Segment length | Slack number | Reference chromatograms or mass spectrums |
|---|---|---|---|---|
| UPLC | *H. polybotrys* | 148 | 1 | RH15 |
| UPLC | *A. membranaceus* | 149 | 22 | RA3 |
| MALDI-TOF MS | *H. polybotrys* | 150 | 5 | RH28 |
| MALDI-TOF MS | *A. membranaceus* | 151 | 4 | RA40 |

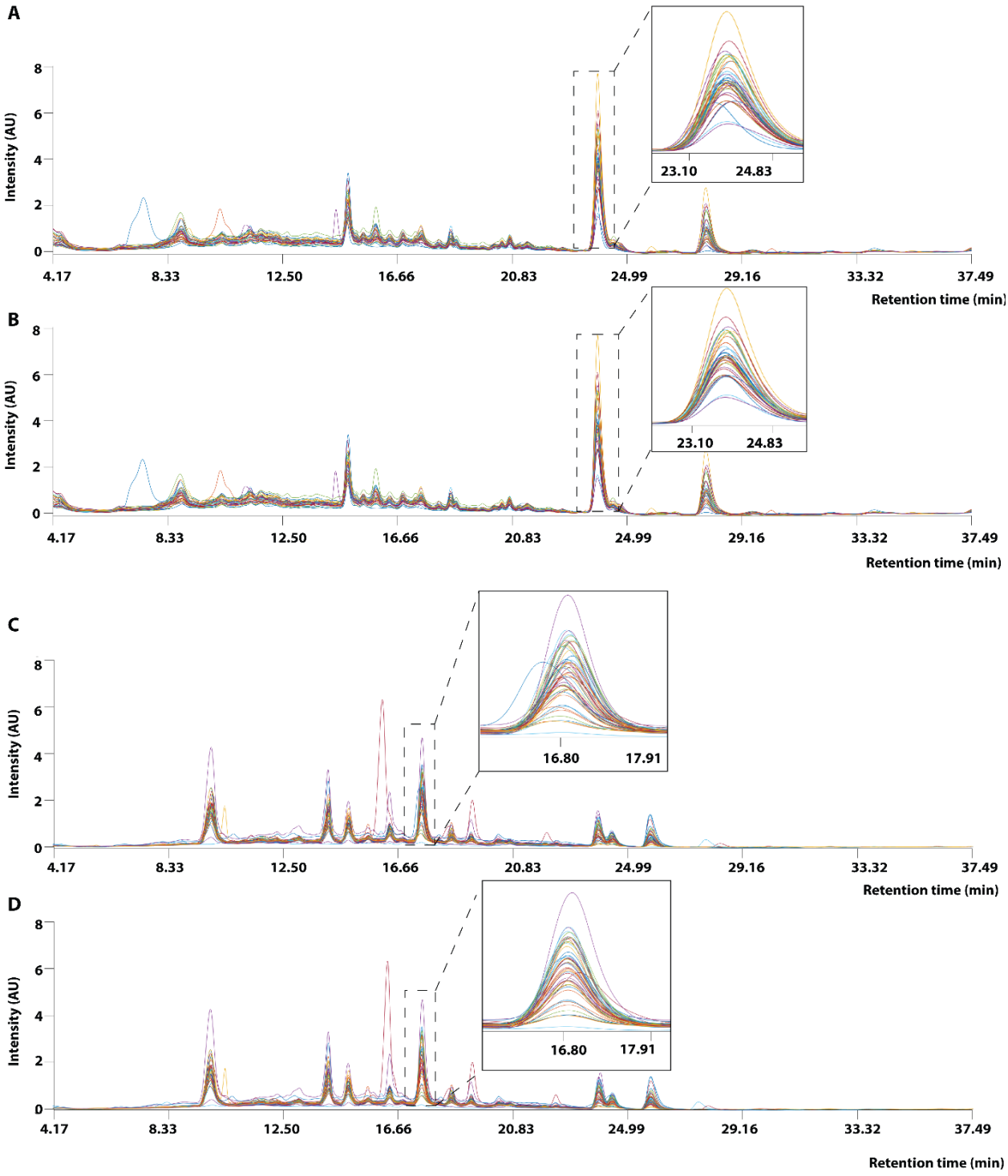



**Figure 3.8. UPLC chromatogram of 40 RH and 51 RA methanolic extracts between retention time between 4.17 and 37.49 min.** (A) raw data and (B) after Correlation Optimized Warping (COW)-corrected UPLC chromatogram (segment=148; slack=1) of RH samples. (C) Raw data and (D) after Correlation Optimized Warping (COW)-corrected UPLC chromatogram (segment=149; slack=22) of RA samples. The chromatograms were zoomed in to show peak alignment after warping.

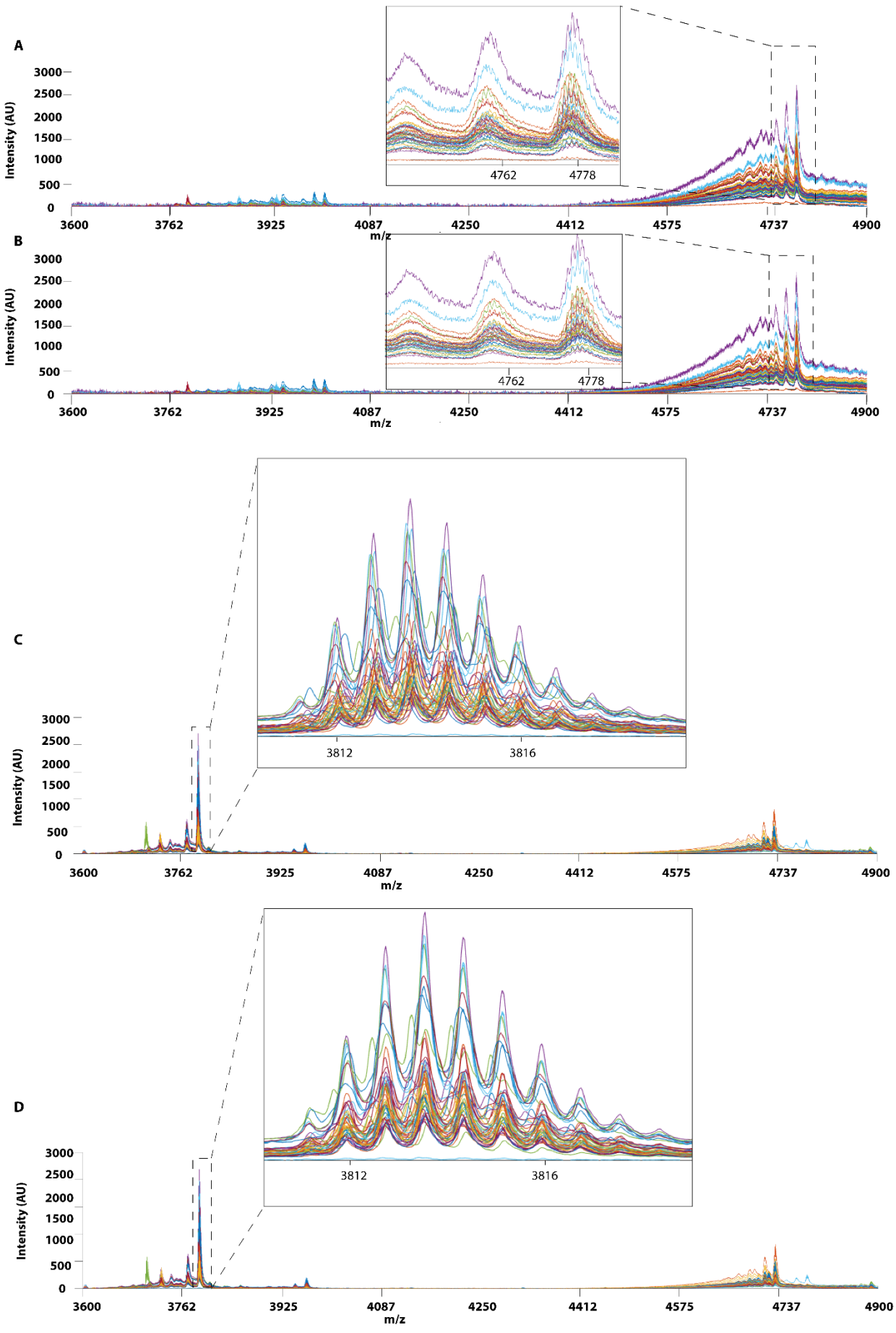

A
B
C
D
Intensity (AU)
m/z
3000
2500
2000
1500
1000
500
0
3600
3762
3925
4087
4250
4412
4575
4737
4900
4762
4778
3812
3816

**Figure 3.9. MALDI-TOF MS profiles of 40 RH and 51 RA between 3600 and 4900 Da.** (A) raw data and (B) after Correlation Optimized Warping (COW)-corrected MALDI-TOF MS spectrum (segment=150; slack=5) of RH samples. (C) Raw data and (D) after Correlation Optimized Warping (COW)-corrected MALDI-TOF MS spectrum (segment=151; slack=4) of RA samples. The chromatograms were zoomed in to show peak alignment after warping.

#### 3.2.5.2. Detection of outliners and unsupervised multivariate analyses

Outlier detection is an essential evaluation before constructing a classification model, due to the reason that possible anomalous samples in the data matrices could affect the quality of the model and therefore, should be removed beforehand. In this study, PCA is employed to identify the presence of outliers and provide an overall idea about the sample distribution. Prior to PCA analysis, the samples were preprocessed by COW baseline removal, SNV, and mean centering. The determination of outliners was accessed by Hotelling's T square verse Q residuals plot, where a sample with high Hotelling's T square and Q residual value are considered as an outlier. The outlier could donate a higher influence on the model and has a larger variation compared to the projected data and thus should be eliminated before further multivariate analysis. In this study, the MALDI-TOF MS data matrix was used as the primary data source for constructing the classification model, and thus, the outlier detection was applied mainly on the MALDI-TOF MS data. Four outliers (RH22, RH30, RA24, and RA46) were detected from the data matrix and removed for subsequent analysis. Figure 3.10 shows the Hotelling's T square versus the Q residuals plot before and after removing the outliners. It can be observed that four outliners are far from the sample major cluster (Figure 3.10A). After the elimination of the four outliners, no samples were detected to have both high Hotelling's T square and Q residual values (Figure 3.10B). Hence, the dataset was reduced to 87 samples, including 38 samples of RH and 49 samples of RA. The reason for the presence of such outliers may be due to the different culture conditions of the plants, including soil conditions, climate, and harvest season.

PCA is the oldest and most widely used method for reducing the dimensions of multivariate problems. It is useful for finding combinations of variables or factors that describe the major trends in a data set. Figure 3.10C shows the PC1-PC2 score plots of the preprocessed data after removing the four outliners. From the scatter points, the samples could be classified into two groups: RH and RA. The results suggest that these two species have distinct spectrometric characteristics.

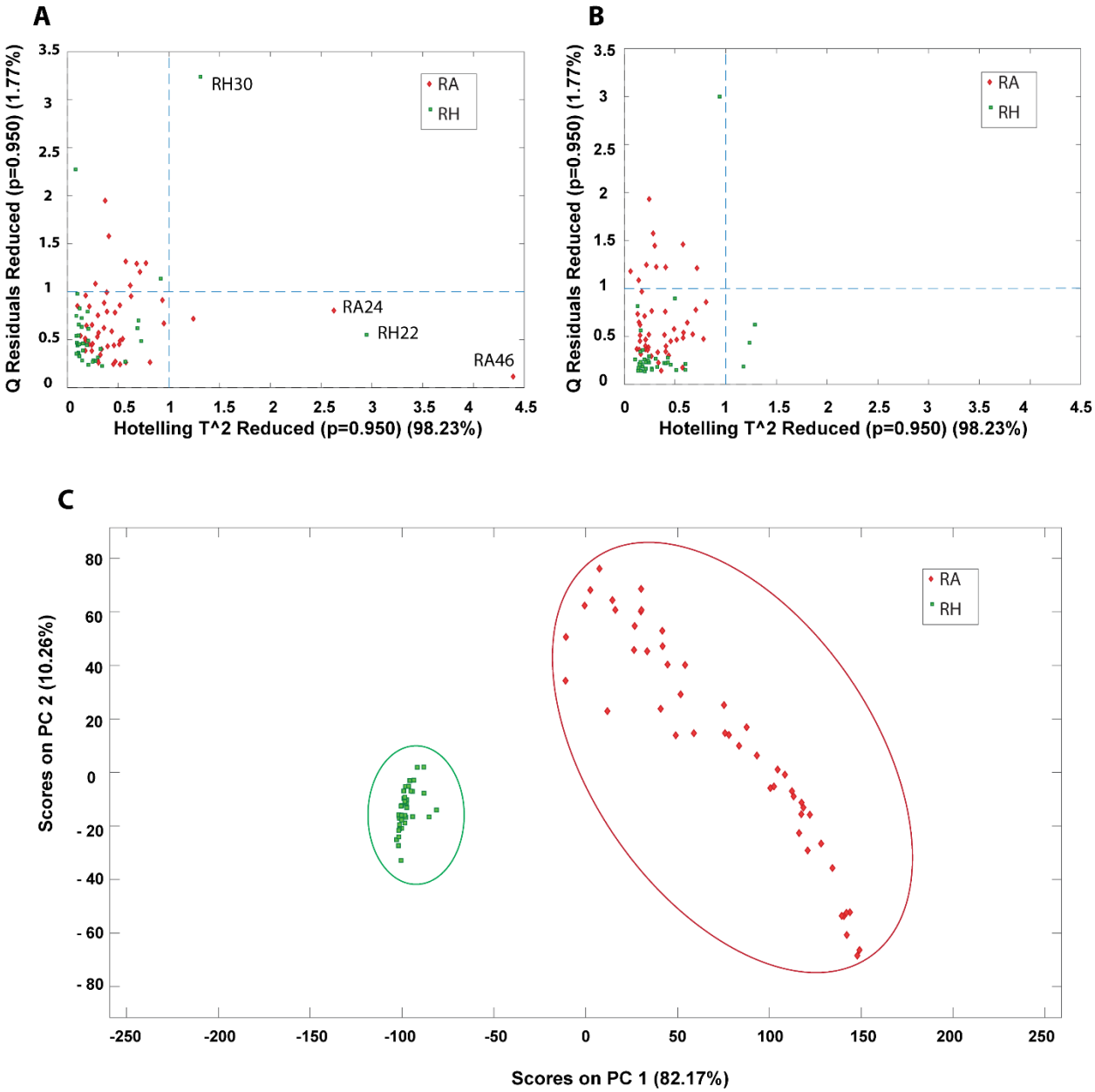



**Figure 3.10. PCA plots obtained from pre-processed MALDI-TOF MS data matrix.** Hotelling's T^2 versus Q residuals plot of (A) raw data and (B) after removal of outliners. (----) line represents 95% confidence interval; (C) PC1-PC2 scores plot after removing outliners.

HCA was performed as a continuation of PCA. With different classification algorithms, it is more promising to obtain sensitive sample classification [271]. In this study, Euclidean distance was used to measure the distance similarity, and Ward's method was applied. HCA draws a connection between RA and RH, producing a dendrogram (Figure 3.11) in which similar samples are grouped, and this similarity is calculated based on the distance between the samples. The dendrogram showed that all RA samples and RH samples are well separated into two major clusters, highlighted in green and red, respectively. Taken together, our results

showed that the clustering pattern obtained using HCA agreed with the classification results acquired from PCA, indicating that the MALDI-TOF MS-based CRP fingerprinting method can deliver a consistent classification result.

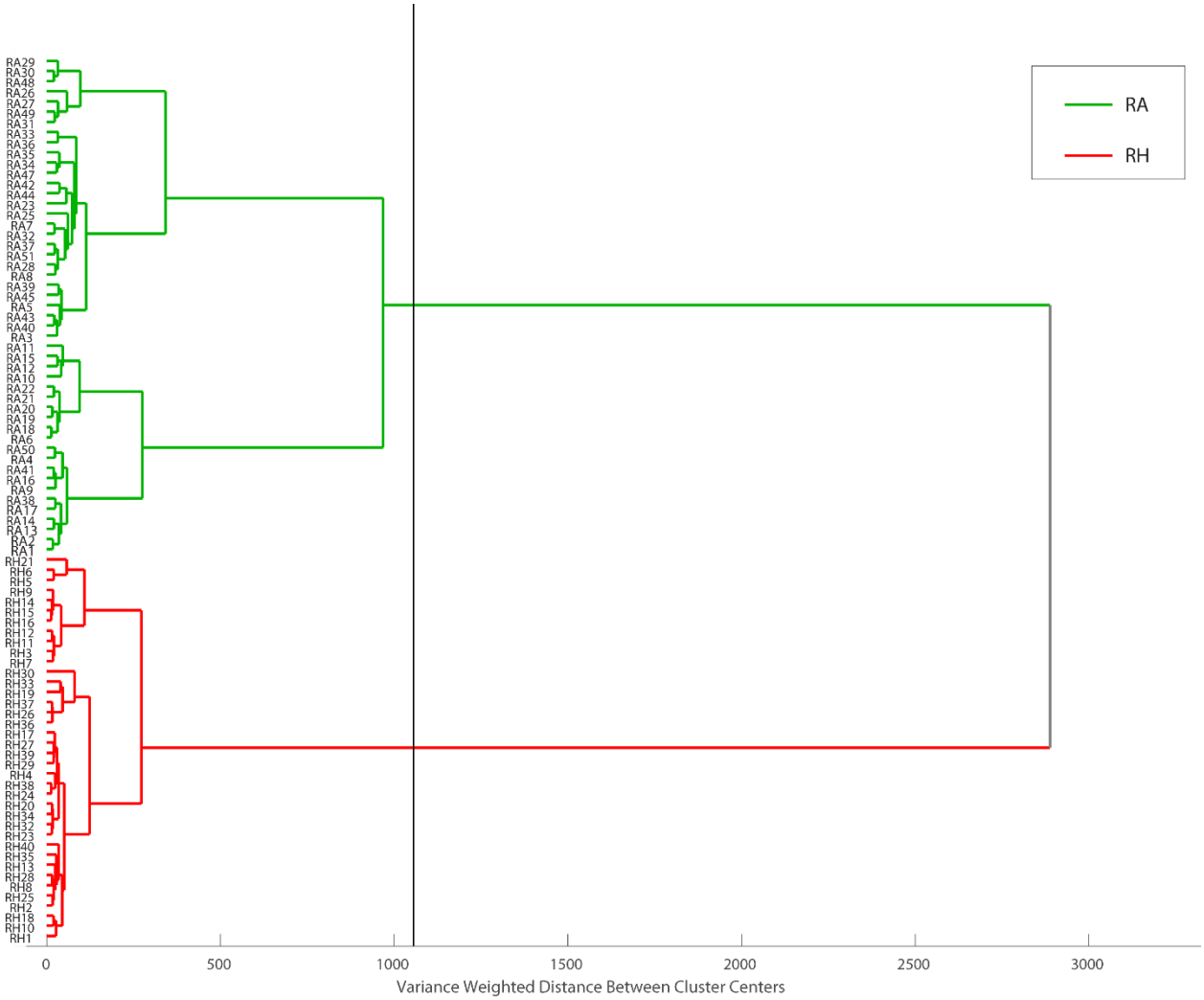



**Figure 3.11. Dendrogram representation of HCA performed on MALDI-TOF data matrix using Euclidean distance and Ward's method.** Obtained data can be divided into two main groups, RH and RA samples at the distance of about 1200 (the black line represents cut-off level).

### 3.2.5.3. Optimization of pre-processing methods

Multiple combinations of pre-processing techniques were employed in the study to improve the data quality, reduce the noise of the raw data matrices to improve the interpretability of the constructed multivariate analysis models. However, there is no well-established procedure of

applying preprocessing algorithms, and therefore, a procedure of optimization of the preprocessing techniques is needed.

In order to evaluate and compare the performance of the combination of pre-processing methods, a PLS-DA model was used. Firstly, the data matrices were separated into a calibration set and a validation set based on the Kennard-Stone (K-S) algorithm. The model was established from 23 RH and 30 RA samples consisting of the calibration set, whereas the remaining samples (15 RH and 19 RA) as a validation set after eliminating the four outliners. Preliminary pre-processing such as peak alignment and baseline correction has been applied prior to the application of mean-centering, SNV, and normalization on the data matrices. Briefly, mean-centering was used to subtract the column mean from each variable in the respective column, while normalization aims at dividing each variable by the sum of the absolute value of all variables [272]. In contrast, by removing the slope variation, SNV can normalize the variables from each chromatogram or spectrum and thus correct the non-constant variances in signal recording [273]. The statistical performances of different combinations of pre-processing techniques were evaluated by the root mean square error of calibration (RMSEC), root mean square error of cross-validation (RMSECV), root mean square error of prediction (RMSEP) and correlation coefficient from LOO-CV as shown in Table 3.4.

To be considered as a good combination of preprocessing method, it should give a low complexity (the number of LVs), a low root mean error and a high correlation coefficient. In this study, all pre-processing methods can improve the performance of the model. However, pre-processing with SNV followed by mean centering showed the highest correlation coefficient, and the smallest root mean error between the calibration and validation data set. Therefore, this combination of pre-processing methods was chosen as the optimal method and applied to the MALDI-TOF MS and UPLC data matrices in the subsequent analysis.

**Table 3.4. Comparison of the statistical performance of PLS-DA model after applying various preprocessing methods on the MAIDI calibration and validation data set.**

| Pre-processing method(s) | LV(s)[a] | RMSEC[b] | RMSECV[c] | RMSEP[d] | Deviation between RMSEP and RMSEC (%)[e] | $r^2$ cal[f] | $r^2$ CV[g] | $r^2$ val[h] |
|---|---|---|---|---|---|---|---|---|
| None | 3 | 0.2260 | 0.2481 | 0.1944 | -16.2551 | 0.7966 | 0.7572 | 0.7474 |
| SNV[i] + mean centering | 2 | 0.0687 | 0.0749 | 0.0671 | -2.3845 | 0.9808 | 0.9772 | 0.9837 |
| Normalization + mean centering | 2 | 0.1034 | 0.1114 | 0.0938 | -10.2345 | 0.9565 | 0.9495 | 0.9654 |
| Mean centering + normalization | 2 | 0.1538 | 0.1738 | 0.1469 | -4.6971 | 0.9046 | 0.8778 | 0.9206 |
| Mean centering + SNV | 2 | 0.1324 | 0.1507 | 0.1518 | 12.7799 | 0.9288 | 0.9076 | 0.9139 |

[a] Optimal number of latent variables; [b] Root mean square error of calibration; [c] Root mean square error of cross-validation; [d] Root mean square error of prediction; [e] Deviation between RMSEC and RMSEP; [f] Correlation coefficient of calibration set; [g] Correlation coefficient of cross-validation set; [h] Correlation coefficient of validation set; [i] Standard normal variate.

### 3.2.6. Comparison of various classification models

In this study, we constructed the model using the MALDI-TOF MS matrix as the reference and compared it with the chromatographic matrix. The chromatographic matrix provides high stability and superior model performance, as chromatographic fingerprint was used as a gold standard for quality control of herbal products according to the World Health Organization (WHO) [146] and the United States Food and Drug Administration [165]. However, this method is time-consuming and challenging to handle. Hence, our study focused on exploring the possibility of an alternative and efficient approach for quality control is needed to be analyzed. Models based on UPLC and MALDI-TOF MS data matrix were established using different classification models, including PLS-DA, KNN, CART, SIMCA, and SVM-DA. Both data matrices were preprocessed and separated into calibration (53 samples with 23 RH and 30 RA) and validation (15 RH and 19 RA) sets, which were selected by the K-S algorithm. The calibration data set was used to establish and train the classification model while the validation dataset used to compare and evaluate the predictive ability of the calibrated model in the final step.

Table 3.5a and 3.5b compare the confusion matrices of the preprocessed MALDI-TOF MS and UPLC fingerprints, based on different classification techniques. The results are demonstrated by the cross-validation set (CV) and the perdition set (Pred), which indicate the model interpretability and predictability, respectively. For the MALDI-TOF MS data, the classification model constructed by KNN, PLS-DA, and SVM-DA provided the greatest interpretability (100.00%) and predictability (100.00%) for both the cross-validation and prediction data sets, whereas some misidentification was found in the SIMCA and CART models and thus led to worse performance. In terms of interpretability, SIMCA has higher accuracy (96.20%) compared to CART (87.00%), in which a total of 7 samples (4 RH and 3 RA) were misidentified. However, CART has shown greater predictability (97.00%), compared to SIMCA (88.90%), in which four samples were not assigned to the prediction set.

Generally, compared with other analytical techniques, UPLC or HPLC data will provide a relatively high prediction ability of classification performance [274]. The classification results based on the UPLC data matrix were summarized in Table 3.5b. In terms of interpretability, KNN showed the highest accuracy (100.00%) whereas PLS-DA and SVM-DA generated slightly lower interpretability (96.20%), both with one RH sample misidentified as the RA sample. SIMCA (89.20%) and CART (96.00%) showed relatively low interpretability on the cross-validation set, which is similar to the results obtained from the MALDI-TOF MS data. Additionally, with 5 (2 RH and 3 RA) samples not assigned, SIMCA provided the worse

interpretability. However, all five models generated a 100.00% correct rate in prediction set, which indicated the high predictive ability of the UPLC data matrix using different classification models.

To obtain a better understanding of the classification ability of different classification models, the error rate (ER), the non-error rate (NER) sensitivity and specificity of each model were calculated and compared (Table 3.5c). The four parameters demonstrated the ability of the models to correctly classify samples belonging to a specific class, and the ability to reject the samples from all other classes. Since the sensitivity was symmetrical to specificity while RH and RA were the only two classes in this study, hence only the sensitivity and specificity of RA were described. Based on the MALDI-TOF MS data, KNN, SVM-DA, and PLS-DA generated a perfect score in both specificity and sensitivity. The minimal deviation between sensitivity and specificity indicated that no particular trend in the models in recognizing either RA or RH. In contrast, SIMCA and CART showed relatively low sensitivity and specificity values, respectively. The low sensitivity value in SIMCA (0.89) suggested that the model is preferable in discriminating RH than RA. On the contrary, the low specificity value in CART (0.93) indicated the greater ability of this model to differentiate RA than RH. Similar to the results in the confusion matrix, all models based on the UPLC data showed a perfect score with zero error rate, suggesting that all classification models were able to classify the samples based on the UPLC data correctly.

The classification performance based on the UPLC data matrix generated a perfect result on all five classification models, in which all RH and RA samples were classified in the prediction set correctly. Importantly, the various classification models established from the MALDI-TOF MS data matrix demonstrated comparable classification ability to the UPLC data matrix, with prediction accuracies more than 89.00% in all models. These results suggested that MALDI-TOF MS can also be applied as a reliable alternative analytical technique for sample authentication.

Based on the characteristics of classification models, KNN and CART were the preferable algorithms for differentiating RH from RA since it required minimal data handling procedures, less parameter optimization, and short running time. However, PLS-DA and SVM-DA are preferable if the study focused on the distribution of the classes and the relationships between the variables, for the reason that these models provided detailed information about the data matrices as shown in the score and loading plot. Furthermore, among all the classification models, SIMCA is the least suitable technique for distinguishing RH and RA because of its

lowest accuracy and high error rate in prediction and requirements in multiple steps for data optimization.

**Table 3.5a. The confusion matrices obtained from the prediction set of the pre-processed MALDI-TOF MS data.**

| | CV[a] | | | | | | Pred[b] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted class | | | | | Predicted class | | | |
| | True class | N[c] | RH | RA | NA[e] | Accuracy (%) | N | RH | RA | NA | Accuracy (%) |
| KNN | RH | 23 | 23 | - | - | 100.00 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 30 | - | | 19 | - | 19 | - | |
| PLS-DA | RH | 23 | 23 | - | - | 100.00 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 30 | - | | 19 | - | 19 | - | |
| SIMCA | RH | 23 | 23 | - | - | 96.20 | 15 | 13 | - | 2 | 88.90 |
| | RA | 30 | - | 29 | 1 | | 19 | - | 17 | 2 | |
| SVM-DA | RH | 23 | 23 | - | - | 100.00 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 30 | - | | 19 | - | 19 | - | |
| CART | RH | 23 | 19 | 4 | - | 87.00 | 15 | 14 | 1 | - | 97.00 |
| | RA | 30 | 3 | 27 | - | | 19 | - | 19 | - | |

[a] Venetian blind cross-validation set; [b] Prediction set; [c] Number of samples; [d] Not assigned

**Table 3.5b. The confusion matrices obtained from the prediction set of the pre-processed UPLC data.**

| | CV[a] | | | | | | Pred[b] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Predicted class | | | | | Predicted class | | | |
| | True class | N[c] | RH | RA | NA[d] | Accuracy (%) | N | RH | RA | NA | Accuracy (%) |
| KNN | RH | 23 | 23 | - | - | 100.00 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 30 | - | | 19 | - | 19 | - | |
| PLS-DA | RH | 23 | 22 | 1 | - | 96.20 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 30 | - | | 19 | - | 19 | - | |
| SIMCA | RH | 23 | 21 | - | 2 | 89.20 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 27 | 3 | | 19 | - | 19 | - | |
| SVM-DA | RH | 23 | 22 | 1 | - | 96.20 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | - | 30 | - | | 19 | - | 19 | - | |
| CART | RH | 23 | 22 | 1 | - | 96.00 | 15 | 15 | - | - | 100.00 |
| | RA | 30 | 1 | 29 | - | | 19 | - | 19 | - | |

[a] Venetian blind cross-validation set; [b] Prediction set; [c] Number of samples; [d] Not assigned.

**Table 3.5c. The classification parameters of the pre-processed UPLC and MALDI-TOF MS data obtained from the prediction set.**

| | MALDI-TOF MS | | | | UPLC | | | |
|---|---|---|---|---|---|---|---|---|
| | | | RA | | | | RA | |
| | ER[a] | NER[b] | Specificity | Sensitivity | ER[a] | NER[b] | Specificity | Sensitivity |
| KNN | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| PLS-DA | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| SIMCA | 0.06 | 0.94 | 1 | 0.89 | 0 | 1 | 1 | 1 |
| SVM-DA | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| CART | 0.03 | 0.97 | 0.93 | 1 | 0 | 1 | 1 | 1 |

[a] Error rate; [b] Non-error rate.

Similar results can be observed in previous studies comparing the efficacy of different classification models. For example, Wong et al. have conducted a study on the differentiation of Puerariae Lobatae Radix and Puerariae Thomsonii Radix using HTPLC coupled with a seven classification model. The results showed that the classification error rate of the model established from SIMCA delivered the worst performance, with a 0.5 error rate and 60.00% accuracy in predicting the class of sample, compared to KNN, PLS-DA and SVM-DA [149]. In addition, CART also showed a worse performance with low sensitivity value (0.38) and prediction rate (64.29%) than the other classification models. Another study by Aderval S. Luna et al. on the authentication of transgenic and nontransgenic soybean oil using NIR spectroscopy illustrated similar results on the classification models [245]. SVM-DA (CV: 100.00%, Pred: 95.00) and PLS-DA (CV: 97.50%, Pred: 90.00%) showed a higher correct classification rate, whereas SIMCA provided lower results in class modeling. However, not all studies showed the same results. For example, Martins et al. revealed that the SIMCA model established from HPLC data showed only a 12.00% correct rate when identifying six different *Phyllanthus* species. On the contrary, 100.00% of the rice seed samples were correctly classified while a SIMCA model based on NIR data was constructed, whereas only 80.00% accuracy was demonstrated by both PLS-DA and KNN models [275]. Taken together, it revealed that the ability of a classification model might not be the same when constructing models on different data sets or applying different analytical methods.

## 3.3. Conclusion

Traditionally, small-molecule metabolites quantification using HPLC analysis is employed to authenticate herbs and herbal products. In this chapter, a rapid and general method for herbal authentication based on the CRPs, which are hyperstable chemical spaces between 2 and 6 kDa, was described and designated as CRP fingerprinting. Our screening results on 100 herbs and herbal products revealed that CRP fingerprinting produces consistent results regardless of the morphology, chemical composition, and origins of the herbs and herbal products. In addition, CRP fingerprinting was further validated by the differentiation of RA from its substitute species RH. Using the MALDI-TOF MS technique, we showed that RA contains unique CRPs such as astratides aM1 and aM2 while hedytides hP1 and hP2 are novel CRPs that only found in RH species. *De novo* sequencing revealed that astratides and hedytides contain different amino acid composition. Compared to the conventional quality control method using chromatographic fingerprinting, CRP fingerprinting based on MALDI-TOF MS analysis is 500-fold faster. Unsupervised multivariate analyses such as PCA and HCA showed that RA and RH could be separated into two clusters based on their CRP fingerprints. In addition, the

classification ability of CRP fingerprinting coupled with five supervised multivariate analyses had comparable classification accuracy to that of UPLC. In terms of the performance of classification models, KNN, PLS-DA, and SVM-DA from CRP fingerprinting showed a perfect correct classification rate (100.00%) while minor classification errors (3.00%) were found in the CART model. With 88.90% sensitivity and 94.00% correct rate of classification, SIMCA performed worse and thus became the least preferable classification model. Overall, with simple handling procedure and accurate classification results, CRP fingerprinting can be used as a novel and general approach for quality control and authentication of herbal and natural products.

# Chapter 4 Discovery and characterization of insulin-modulating, insecticidal and antifungal cysteine-rich peptides from *Astragalus membranaceus*

## 4.1. Introduction

Medicinal plant-derived medicine is important for treating human diseases and maintaining health. Until now, it is reported that plant-derived small-molecule metabolites with M.W. < 1000 Da account for approximately 45% of all clinically approved drugs and inspire the development of > 37% of the synthetic drugs [276]. On the contrary, peptides derived from natural products are usually not clinically approved as drugs. Thus, medicinal plant-derived peptides are highly under-explored.

Our laboratory is particularly interested in the plant-derived cysteine-rich peptides (CRPs) with 3 – 5 disulfide bonds, which occupying the chemical spaces from 2 – 6 kDa. The highly cross-linking disulfide bridges of CRPs confer them high metabolic stability against temperature, pH, and environmental changes and thus reveal their potential to be orally active therapeutics [266]. Generally, compared to small-molecule drugs, CRPs contain larger footprints which confer a higher on-target pharmacological specificity and selectivity [59, 277].

Pea Albumin 1 subunit b (PA1b), a hormone-like CRP with 37 amino acids, is discovered from *Pisum sativum* by Higgins *et al.* in 1986 [278]. Typically, its structure is characterized by a triple-stranded anti-antiparallel β-sheet and a cystine-knot core [279]. This structure provides its high metabolic stability against proteolytic, temperature, and chemical changes [280]. PA1b was found to exhibit a variety of bioactivities in insect, plant and mammalian world. Previous studies revealed that PA1b could bind to the plasma membrane in the insect midgut [281]. Thus, it is cytotoxic to insect Sf9 cells and lethal to cereal weevils (*Sitophilus granaries*) [282]. In plants, leginsulin, which is the homolog of PA1b, is capable of enhancing cell proliferation, suggesting that the PA1b homologs may act as plant hormones to regulate plant signal transduction [283]. In the mammalian world, PA1b was reported to be able to interfere with glucose homeostasis [281, 284-286].

Defensins are a superfamily of CRPs with host defense functions that are broadly distributed in plants, animals, and microbiomes [287]. In 1995, peptides previously known as γ-thionins were corrected as “plant defensins” because studies revealed that antifungal peptides Rs-AFP1 and Rs-AFP2 identified from radish seeds, share higher similarities with insect and mammalian defensins than with plant thionins [288]. In general, plant defensins containing 45-54 amino

acids in length are cationic peptides with a broad range of bioactivities such as antifungal, antibacterial, and enzyme inhibitory activities [12].

*Astragalus membranaceus* belonging to the plant Fabaceae family is a perennial plant usually found in China and northern Asia. The roots of *A. membranaceus* (Chinese: Huang Qi), is a famous health-promoting Traditional Chinese Medicine in China. As documented in the Shennong's Classic of Materia Medica, the roots of *A. membranaceus* are commonly employed to boost overall vitality, eliminate weakness and diabetes symptoms [197]. Previous studies showed that polysaccharides, saponins, flavonoids, and sterols are the major chemical compounds present in the *A. membranaceus* roots [289]. A study on *A. membranaceus* decoction revealed its ability to reduce blood glucose level in both rats [290] and patients with diabetes mellitus [291]. However, the principal ingredients responsible for such anti-diabetic activity remain unclear.

In the previous chapter, I have reported that CRPs are widely distributed in Plant kingdom and could be used as unique chemical markers to differentiate species. Therefore, the following two chapters will be focusing on the detailed discovery and characterization of CRPs from specific medicinal plants. In this chapter, two CRPs, $\alpha$-astratide aM1 and $\beta$-astratide bM1, were discovered and identified from the roots of *A. membranaceus*. The combination of proteomic and transcriptomic analyses revealed that aM1 contains 37 amino acids with six cysteines, which belongs to the family of PA1b-like peptides. In contrast, $\beta$-astratide bM1 containing 47 amino acids with eight cysteines is a plant defensin. In addition, data mining of the biosynthetic precursors of aM1 and bM1 revealed that they share the same precursor arrangement with PA1b-like peptides and plant defensins, respectively. Until now, PA1b-like peptides are only found in the peas' family, and aM1 is the first to be discovered in medicinal plants. Similarly, phylogenetic tree revealed that bM1 belongs to a new subfamily of plant defensins. In addition, biological assays showed that astratides are bioactive peptides with functions including insecticidal, anti-fungal, and insulin-modulating activities, suggesting that they not only play a role in plant host-defense system but also serve as a putative drug lead.

## 4.2. Results and Discussion

### 4.2.1. Screening of CRPs in *Astragalus membranaceus* roots

*A. membranaceus* roots were extracted with water and semi-purified by ZipTip before MALDI-TOF MS analysis. Peptides with M.W. of 2 - 6 kDa were revealed in the mass spectrum (Figure 4.1). Two major clusters of peptides were detected and designated as $\alpha$-astratides and $\beta$-

astratides based on their m/z intensity from 3.5 - 4.1 kDa and 4.6 - 4.9 kDa, respectively. The most abundant CRPs in these two clusters were designated *α*-astratide aM1, with relative monoisotopic molecular masses $[M+H]^+$ of 3811.2 Da and *β*-astratide bM1 of 4724.1 Da.

To show that these peptides are CRPs, samples were treated with a disulfide-reducing agent followed by an *S*-alkylating agent, a procedure commonly used in our laboratory [1, 23, 26, 27, 59, 77, 80, 267]. After the *S*-reduction of the disulfides with DTT and *S*-alkylation of the free thiols with IAA, a mass shift of 348 Da was observed in this cluster of peptides. Since the procedure will cause a 58 Da mass increment for each cysteine residue, our result suggests that there are six cysteine residues and three disulfide bonds present in aM1 (Figure 4.2) [75]. The same method was used to determine the number of cysteines present in *β*-astratide bM1 and the mass shift of 464 Da indicates that bM1 is an 8C-CRP (Figure 4.3).

In the clusters of *α*-astratides, the $[M+H]^+$-57 peak was later shown to be a truncated version of aM1 with a missing glycine at the C-terminus. In contrast, the $[M+H]^+$-17 peaks were observed in both *α*-astratides and *β*-astratides. The peaks could have several origins, for example, a dehydration reaction or a proline to asparagine substitution. The cysteine numbers within the same CRP family are absolutely conserved. Thus, *α*-astratides and *β*-astratides belong to different CRP families since they possess different numbers of cysteine residues.

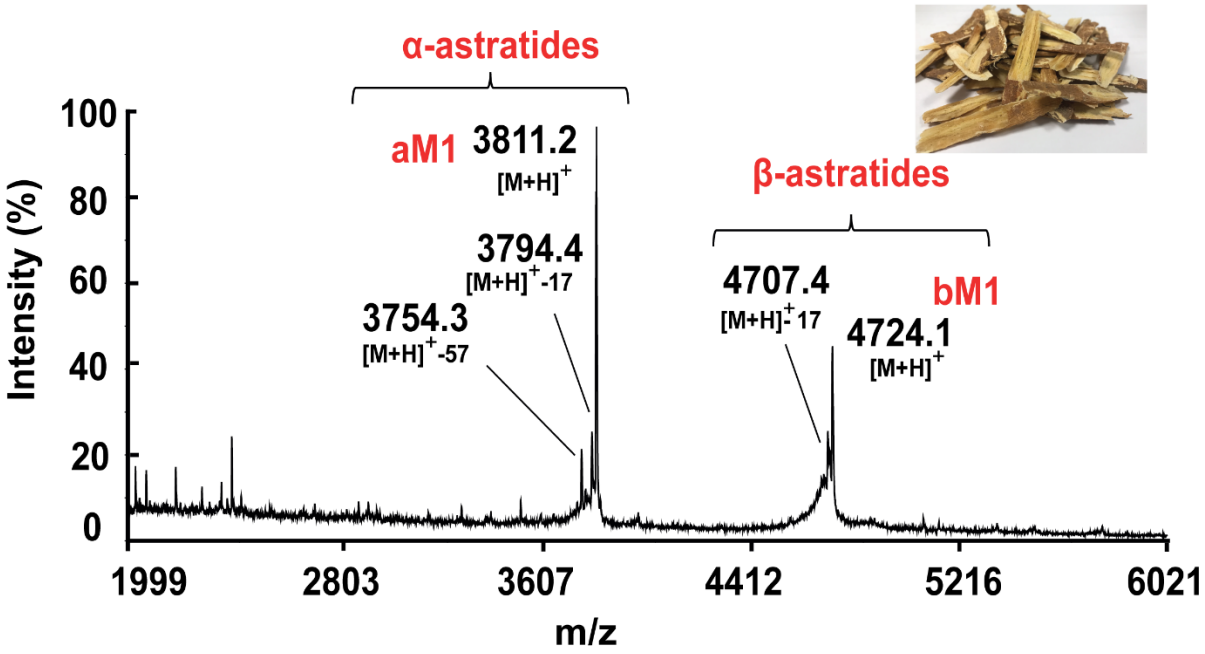



**Figure 4.1. MS spectrum of the aqueous crude extract of *A. membranaceus* roots.** The two major clusters of peptides were designated as *α*-astratides and *β*-astratides. (Photograph on top right corner).

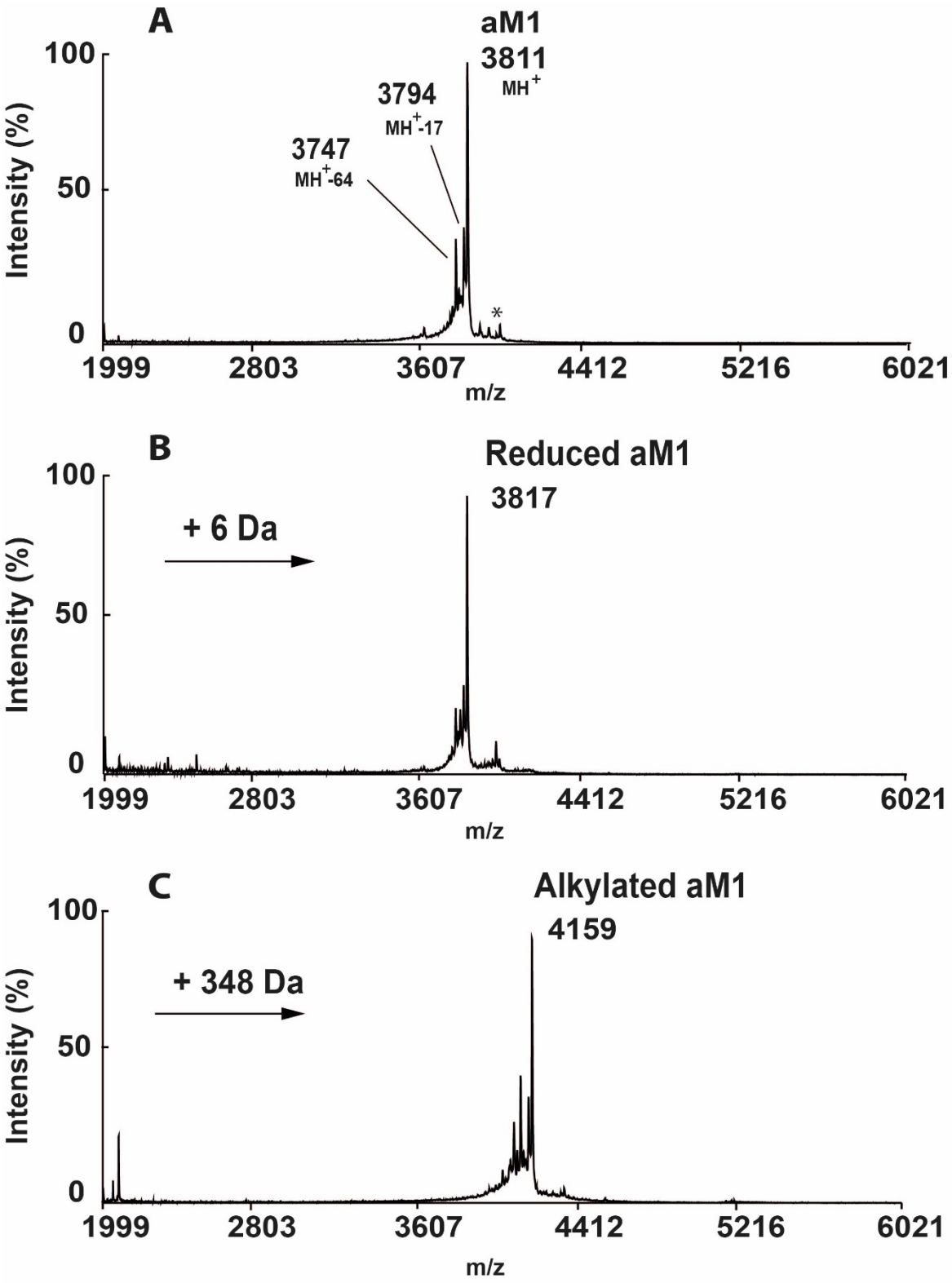



**Figure 4.2. MS profiles of *α*-astratides.** (A) MS spectrum of native *α*-astratides. (B) MS spectrum of α-astratides after *S*-reduction. (C) MS spectrum of *α*-astratides after *S*-alkylation.

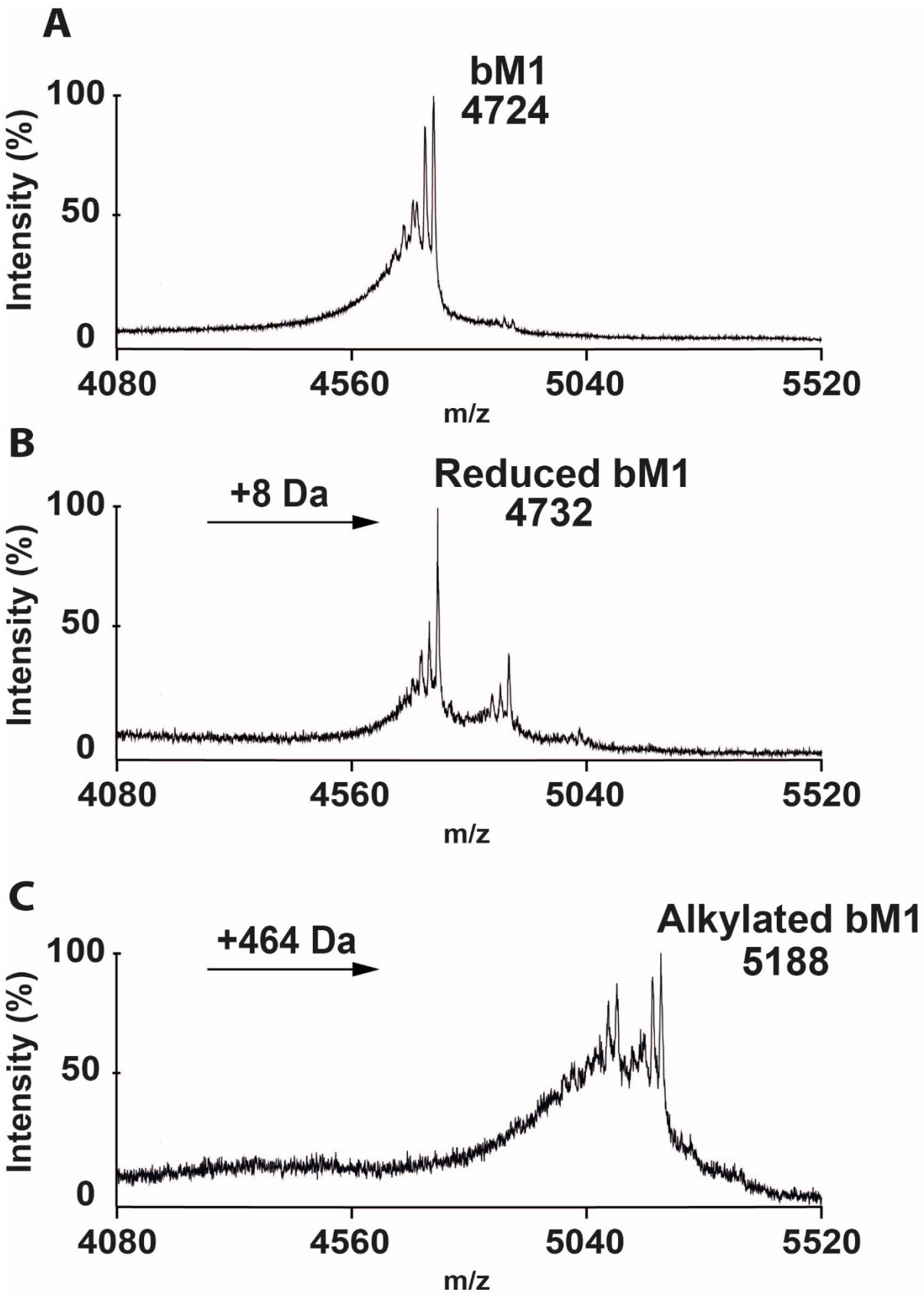



**Figure 4.3. MS profiles of *β*-astratides before and after *S*-reduction and *S*-alkylation.**

### 4.2.2. Isolation and sequence identification of astratides

To further characterize astratides aM1 and bM1, 2 kg of *A. membranaceus* roots was used for a scale-up isolation. The two CRPs were extracted and purified by $C_{18}$-reversed-phase HPLC to >97% purity (Figure 4.4). The extraction yield of aM1 and bM1 were approximately 0.5 mg and 2 mg per kg of dried plant material, respectively.

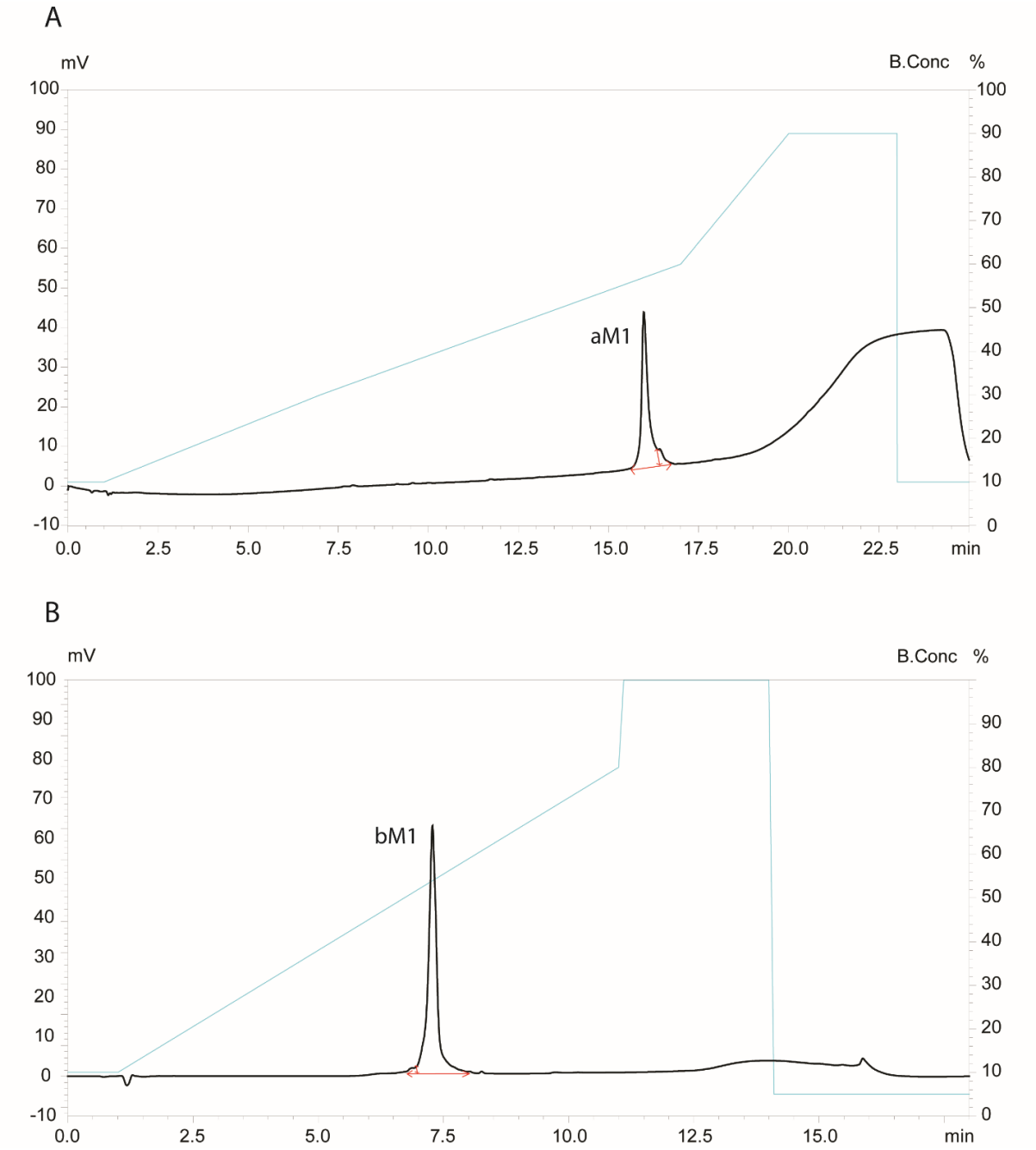



**Figure 4.4. UPLC profiles of isolated (A) *α*-astratide aM1 and (B) *β*-astratide bM1.** The integrated peak areas were highlighted with red lines. The purity of the sample was calculated based on the percentage of the pure peptide of the whole integrated peak area.

After the fully *S*-alkylation, a nanospray MS/MS was performed for *de novo* sequencing of *α*-astratide aM1. The mass differences between *b*- and *y*- ions were employed to elucidate the peptide sequences. Tandem MS analysis revealed that aM1 contains 37 amino acids, with a primary sequence of VDCSGACSPFEVPPCGSRDCRCIPIGLVVGFCIYPTG (Figure 4.5), where the isobaric residues Ile/Leu assignments were confirmed by transcriptome. The same method was used to determine the sequence of *β*-astratide bM1. Tandem MS revealed that bM1 contains 45 amino acids, with a primary sequence of CEKPSKFFSGPCIGSSGKTQCAYLCRRGEGLQDGNCKGLKCVCAC (Figure 4.6).

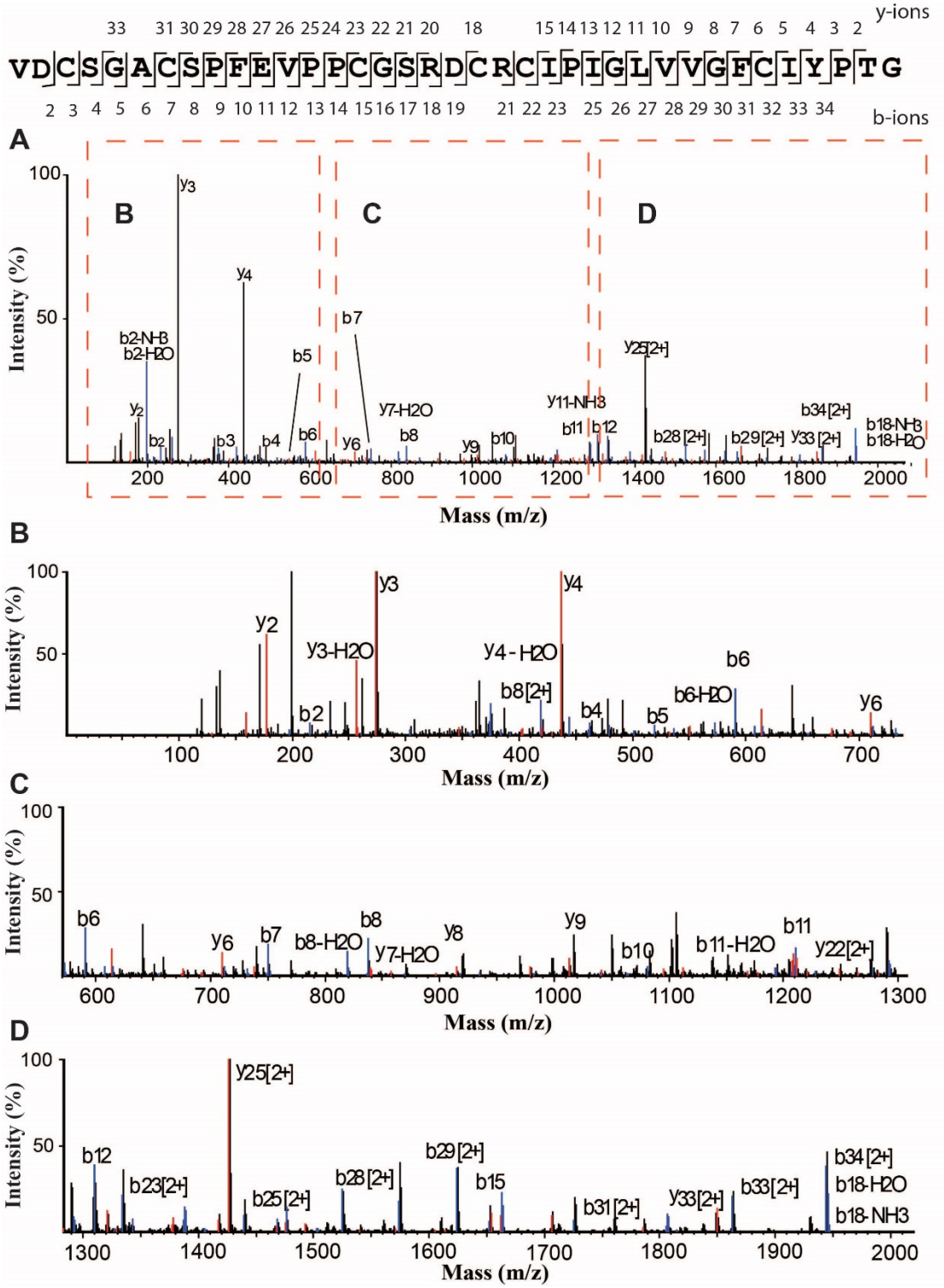



**Figure 4.5. Mass spectra of astratide aM1 from LC-ESI-LTQ-Orbitrap MS/MS in positive ion mode.** Mass ranges of (A) 100 and 600 m/z, (B) 600 and 1300 m/z and (C) 1300 and 2000 m/z were selected for scanning.

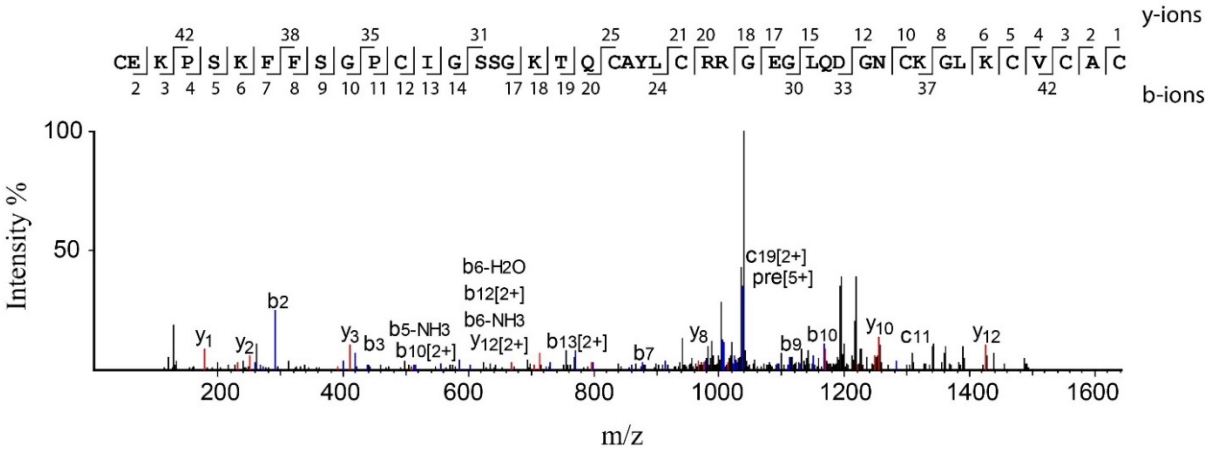



**Figure 4.6. Mass spectrum of astratide bM1 from LC-ESI-LTQ-Orbitrap MS/MS in positive ion mode.** The spectrum was scanned between mass ranges from 0 to 2000 m/z.

### 4.2.3. Sequence analysis of astratides

A BLAST search was performed using *α*-astratide aM1 sequence as a query, and the results showed that aM1 shares an 85.7% sequence similarity with PA1b, which is a hormone-like CRP identified from pea seeds [278]. A comparison between aM1 and PA1b-like peptides from *P. sativum* [292], *Glycine max* [293], *Glycine soja* [294], *Lens culinaris* [295] and *Medicago truncatula* [292] indicates a high sequence identity (62.2 − 86.1%) and similarity (75.7 − 94.4%) between aM1 and PA1b-like peptides, suggesting that aM1 is a PA1b-like peptide (Table 4.1). In addition, the high sequence similarity (94.4%) between aM1 and leginsulin 1 provides confidence that they share same disulfide linkages as Cys I-IV, Cys II-V and Cys III-VI (Table 4.1) forming a cystine-knot structure [283]. Although these peptides belong to the same subfamily of PA1b-like peptides, their biophysical properties varies. For example, aM1, gS1 and mT1 are acidic peptides whereas all the other PA1b-like peptides are basic or neutral peptides.

On the contrary, a high sequence similarity (43.2 – 61.7%) can be observed between *β*-astratide bM1 and plant defensins such as fabatin-2, RS-AFP1, NaD1, and VrD1 (Table 4.2). Results indicated that both bM1 and other reported plant defensins are cationic peptides with 45 to 51 amino acids in length. High sequence similarity of 55.3% between bM1 and AhPDF1 [296] suggests that they share a similar disulfide linkage as Cys I-VIII, Cys II-V, Cys III-VI and Cys IV-VII (Table 4.2) and structure fold with *cis*-oriented disulfides from the C-terminal β-strand [41].

**Table 4.1. Sequence comparison of *α*-astratide aM1 and other reported PA1b-like peptides.**

| Peptide | Species | Amino acid sequence | Mass (Da)[1] | Charge[2] | PI | Approach[3] | Identity% | Similarity% | Ref |
|---|---|---|---|---|---|---|---|---|---|
| Loop | | 1 2 3 4 5 6 | | | | | | | |
| aM1 | *A. membranaceus* | VDCSGACSPF EVPPCGSRDC RCIPIGLVVG FCIYPTG | 3811.5 | -1 | 4.56 | T, P | | | This work |
| PA1b | *P. sativum* | ASCNGVCSPF EMPPCGTSAC RCIPVGLVIG YCRNPSG | 3742.4 | +1 | 7.81 | T,P | 65.7 | 85.7 | [278] |
| PsaA1b012 | *P. sativum* | ASCNGVCSPF EMPPCGTSAC RCIPVGLFIG YCRNPSG | 3790.4 | +1 | 7.81 | T | 62.9 | 82.9 | [292] |
| PsaA1b014 | *P. sativum* | ASCNGVCSPF EMPPCGSSAC RCIPVGLLIG YCRNPSG | 3742.4 | +1 | 7.81 | T | 65.7 | 85.7 | [292] |
| PsaA1b015 | *P. sativum* | ISCNGVCSPF DIPPCGSPLC RCIPAGLVIG NCRNPYG | 3789.5 | +1 | 7.78 | T | 62.2 | 75.7 | [292] |
| Leginsulin1 | *G. max* | ADCNGACSPF EVPPCRSRDC RCVPIGLFVG FCIHPTG | 3920.5 | 0 | 6.74 | T | 86.1 | 94.4 | [293] |
| Leginsulin2 | *G. max* | ADCNGACSPF EMPPCRSRDC RCVPIGLVAG FCIHPTG | 3876.5 | 0 | 6.74 | T | 83.3 | 94.4 | [292] |
| gS1 | *G. soja* | ADCNGACSPF EVPPCRSSDC RCVPIGLFVG FCIHPTG | 3850.5 | -1 | 5.38 | P | 83.3 | 91.7 | [294] |
| lC1 | *L. culinaris* | ADCNGACSPF EMPPCRSSAC RCIPVGLVVG YCRHPSS | 3879.5 | +1 | 7.82 | T, P | 71.4 | 88.6 | [295] |
| mT1 | *M. truncatula* | TDCSGACSPF EMPPCRSSDC RCIPIGLVAG YCTYPSS | 3870.4 | -1 | 4.56 | T | 80.0 | 88.6 | [292] |

[1]Mass (Da) = reported mass. [2]Charge: the total charge is the sum of positive and negative charges present in the sequence. [3]Approach: The approach used to obtain the primary sequences by transcriptomic (T) and/or proteomic (P) analysis. The cysteine residues were highlighted in yellow.

**Table 4.2. Sequence comparison and physiochemical properties of *β*-astratide bM1 and reported plant defensins**

| Peptide | Species | Amino acid sequence | Mass (Da)[1] | Charge[2] | Similarity % | Ref |
|---|---|---|---|---|---|---|
| bM1 | *A. membranaceus* | ----CEKPSKFFSGPCIGSSGKTQCAYLCRRGEGLQDGNCKGLK----CVC--AC | 4724.2 | +4 | | This work |
| Fabatin-2 | *V. faba* | LLGRCKVKSNRFHGPCLT---DTHCSTVCRG-EGYKGGDCHGLR--RRCMC--LC | 5206.1 | +6 | 61.7 | [297] |
| RS-AFP1 | *R. sativus* | -QKLCERPSGTWSGVCGN---NNACKNQCINLEKARHGSCNYVFPAHKCICYFPC | 5682.5 | +4 | 52.1 | [48] |
| NaD1 | *N. alata* | --RECKTESNTFPGICIT---KPPCRKACIS-EKFTDGHCSKIL--RRCLCTKPC | 5292.5 | +6 | 44.4 | [298] |
| VrD1 | *V. radiata* | --RTCMIKKEGW-GKCLI---DTTCAHSCKN-RGYIGGNCKGMT--RTCYCLVNC | 5110.3 | +6 | 50.0 | [299] |
| Lc-def | *L. culinaris* | --KTCENLSDSFKGPCIP---DGNCNKHCKEKEHLLSGRCRDDFR---CWCTRNC | 5437.3 | +2 | 43.2 | [300] |
| NsD7 | *N. suaveolens* | --KDCKRESNTFPGICIT---KPPCRKACIR-EKFTDGHCSKIL--RRCLCTKPC | 5374.6 | +8 | 48.9 | [301] |
| AhPDF1 | *A. halleri* | --RLCEKPSGTWSGVCGN---NGACRNQCIRLEKARHGSCNYVFPAHKCICYFPC | 5567.5 | +5 | 55.3 | [296] |

[1]Mass (Da) = reported mass. [2]Charge: the total charge is the sum of positive and negative charges present in the sequence. [3]Approach: The approach used to obtain the primary sequences by transcriptomic (T) and/or proteomic (P) analysis. The cysteine residues were highlighted in yellow.

It can be observed that the sequence of aM1 can be divided into five intercysteinyl loops based on their cysteine spacing. A sequence logo analysis was generated using the aligned sequences of aM1 and other PA1b-like peptides to reveal the occurrence and similarity of amino acid in each position (Figure 4.7). It can be observed that five prolines (Pro9, Pro13, Pro14, Pro24, and Pro34), three glycines (Gly5, Gly26, and Gly30), Ser8, Phe10, Arg21, and Leu27 together with the six cysteine residues were absolutely conserved in all PA1b-like peptides.

α-Astratide aM1 is considered as proline-rich based on its high content of proline residues (14%). This feature can be observed in other families of CRPs, such as CKAIs and α-ginkgotides [26, 302]. The Pro-rich feature suggests that these CRPs may bind to growth factor receptor-bound protein 2 (Grb2), which contains two SH3 domains that can form a direct complex with the proline-rich regions of other peptides. Upon binding Grb2, these proline-rich peptides, such as aM1, may help in cell signal transduction and thus be advantageous for drug design.

Human prolyl oligopeptidase (POP) is a serine-type protease, which is involved in psychiatric and neurodegenerative diseases [303]. Hence, the inhibition of POP is useful for treating these related diseases [304]. Cyclotide kalata B1 and psysol 2 are POP inhibitors containing three prolines in loops 3, 5, and 6 [303]. It is interesting to notice that, similarly, aM1 possesses three prolines in loop 2 and two prolines in loop 5 and 6, which suggests that aM1 may act as a potential POP inhibitor like kalata B1 and psysol 2.

Another distinct characteristic for aM1 is that it is Ile/Leu/Val rich, with four valine, three isoleucine, and one leucine residues. A similar feature can also be found in cliotides [77] and jasmintides [25]. Interestingly, both peptides are shown to have insecticidal or feeding deterrent activity, in agreement with the finding that aM1 is cytotoxic to insect cells. The combination of Pro-and Leu/Ile-rich confers aM1 with hydrophobic patches. Indeed, loop 5 of aM1 and other PA1b-like peptides contain highly conserved hydrophobic residues, which are thought to confer the insecticidal activity of PA1b-like peptides.

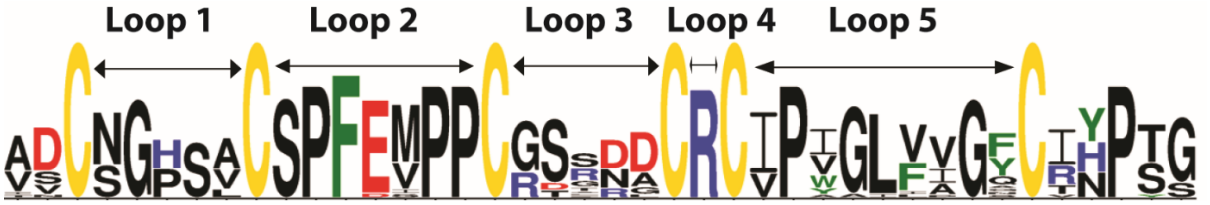



**Figure 4.7. Sequence logo of aligned aM1 and other PA1b-like peptides.** The overall height of the stack indicates the conservation of the amino acid in each position. Within each stack,

the relative occurrence frequency of the amino acid was represented by the height of the symbol. Cysteine residues are displayed in yellow whereas acidic (D and E), basic (H, K, R) and aromatic residues (F, W, Y) are highlighted in red, blue and green colors, respectively.

### 4.2.4. Biosynthesis pathway of astratides

A bioinformatic analysis using OneKP and GenBank database revealed the full-length precursor sequence of aM1 and another 19 PA1b-like peptides from different plant families (Figure 4.8). Both aM1 and other PA1b-like peptides comprise a five-domain precursor arrangement similar to PA1b, which includes a signal peptide domain, a mature peptide, a hinge region, a PA1a-like domain, and a C-terminal tail. The presence of a signal peptide suggests that these peptides may follow the conventional pathway for the secretory peptides [57]. An Ala residue is absolutely conserved at the cleavage site of their N-terminus, whereas a Gly residue is highly conserved at the C-terminal processing site for the release of the mature peptide. With 37 amino acid residues, the length of all PA1b-like peptides is highly conserved. However, a variation in length from 8 to 28 amino acid residues was observed in the C-terminal tails of all the PA1b-like peptides, and among which, aM1 possesses the shortest C-terminal domain.

The pre-protein precursor of PA1b, PA1 is bioprocessed to produce two peptides, PA1a (C-terminal fragment) and PA1b (N-terminal fragment) [278]. Theoretically, these two peptides are equimolar, but studies showed that PA1a was never detected at proteomic levels. Thus the PA1a fragment is postulated to only assist in the correct folding of PA1b [305]. Similar to other CRPs, aM1 undergoes the same secretory pathway to release the mature peptide, which involves the procedures of the cleavage of a signal peptide by signal peptidase, the removal of the hinge domain and *C*-terminal tail by endopeptidase. In the end, the mature aM1 is transported to the Golgi apparatus for post-translational modifications and packed into vesicles for secretion [26].

Cliotides, a group of cyclotides extracted from *Clitoria ternatea*, display a chimera characteristic in their precursors. Their precursors generally consist of a combination of cyclotide domain and albumin-1 domain, with the cyclotide domain replacing the PA1b mature domain. No such chimeric structure and cyclic characteristic are observed in aM1 and other PA1b-like peptides, which is due to the lack of C-terminal Asn. Together with asparaginyl endopeptidase, the C-terminal Asn can form a reactive thioester bond that leads head-to-tail ligation to afford the cyclized structure [306]. The differences between cyclotides and PA1b-

like peptides provide an understanding of the biosynthesis pathway of CRPs in the plant Fabaceae family.

CKAIs are a group of insecticidal plant-derived α-amylase inhibitors belonging to knottin family. Unlike aM1 and other PA1b-like peptides, CKAIs are proline-rich and adopt a three-domain precursor sequence [26]. They also undergo a secretory protein processing pathway, which usually contains the signal peptide removal and the pro-peptide cleavage before releasing the mature peptide [26].

Like 6C-HLPs and thionins, the pro-peptide domain was absent in the precursor sequence of aM1. However, with a five-domain architecture of precursor arrangement, aM1 is different from other 6C-CRPs, which generally contain two to four domains (Figure 4.9). With 128 aa, aM1 contains a relatively longer full precursor sequence when compared to CKAIs (93 aa), 6C-HLPs (84 aa), carboxypeptidase inhibitors (90 aa), jasmintides (100 aa) and $\beta$-ginkgotides (90 aa). The understanding of aM1 precursor arrangement provides clues for developing transgenetic crops as well as a bacterial display system.

In contrast, the transcriptomic data mining showed that the precursor sequence of bM1 comprised a two-domain arrangement, which is similar to other plant defensins, comprising a signal domain and a mature domain (Figure 4.10). Two types of precursor arrangements have been reported from plant defensins, which include a majority with two domains and a minor group with three domains [300]. The additional C-terminal tail present in the three-domain plant defensin precursor was postulated to be associated with the vacuolar sorting mechanism [45].

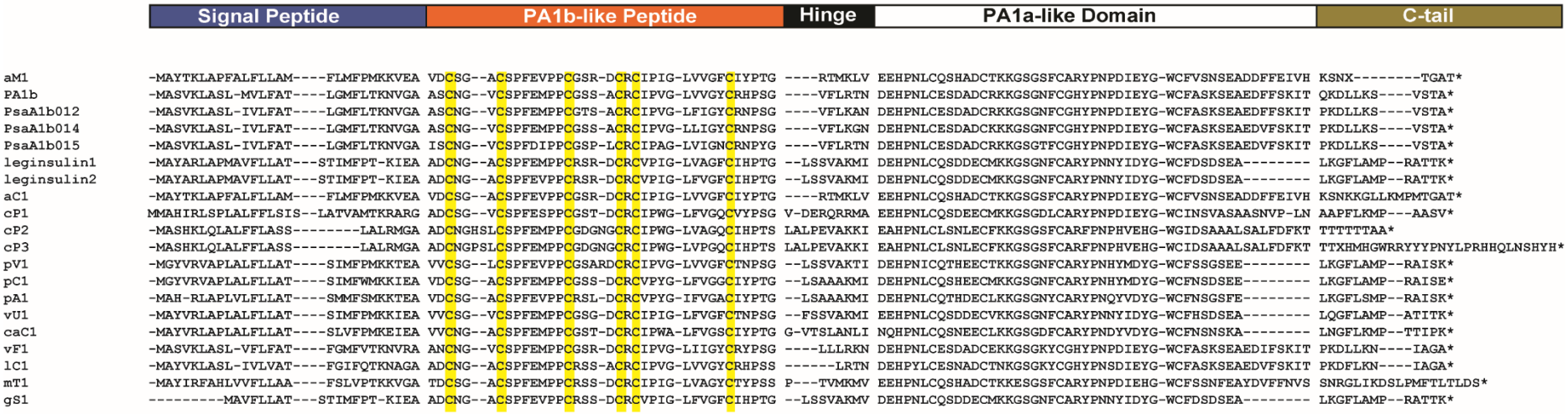



**Figure 4.8. Precursor arrangement and biosynthesis pathway.** (A) Gene alignment of aM1 and other PA1b-like peptides. Their precursor sequences contain five domains, including a signal domain, a mature domain, a hinge domain, a PA1a-like domain, and a C-terminal domain. To release the mature peptide, SPase cleaves the signal peptide before the hinge domain, and C-terminal tail is cleaved by endopeptidase.

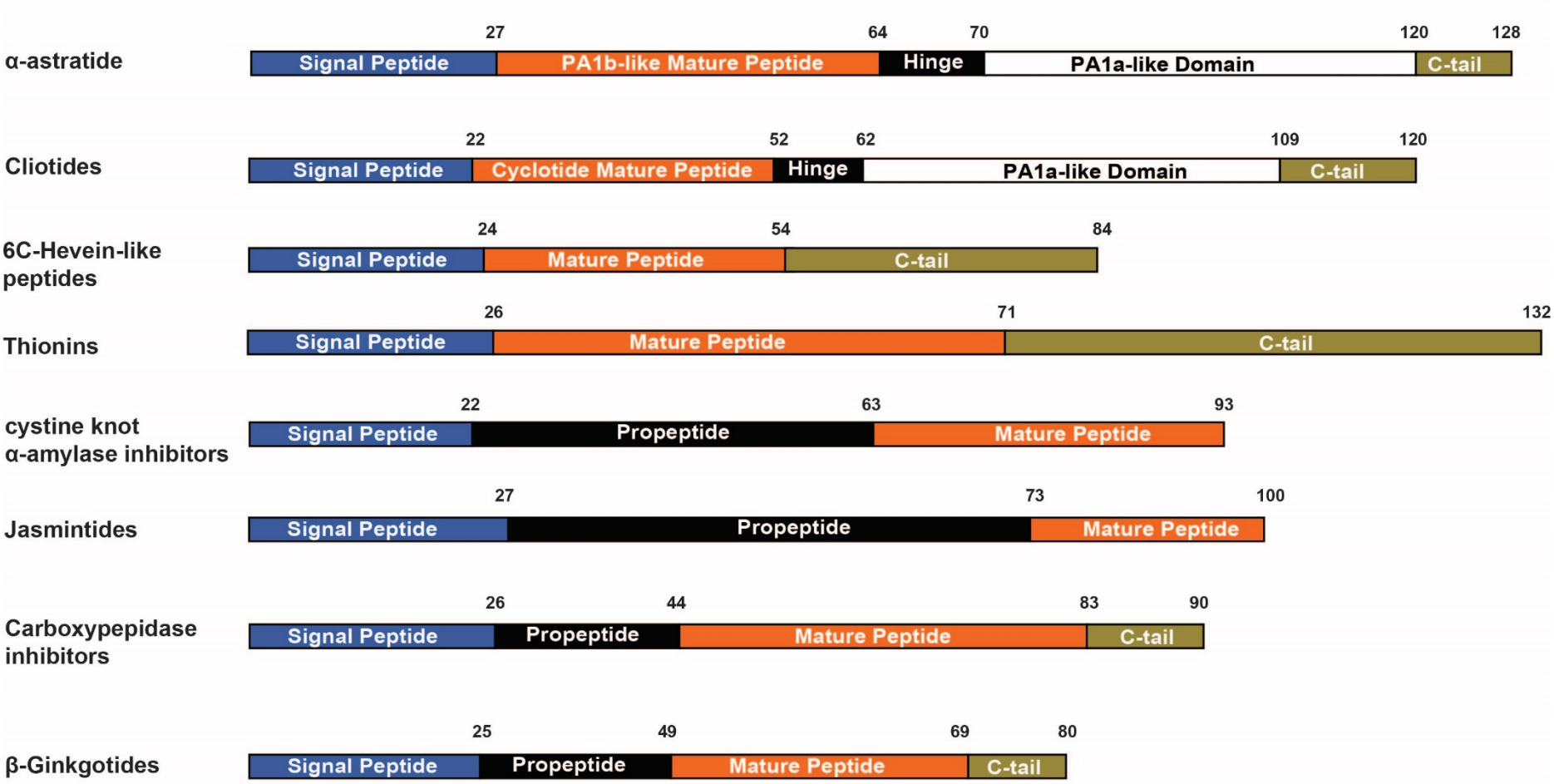



**Figure 4.9. Comparison of precursor sequences of *α*-astratide aM1 and other known 6C-CRPs.** The average numbers of each domain of each peptide were displayed on the top accordingly.

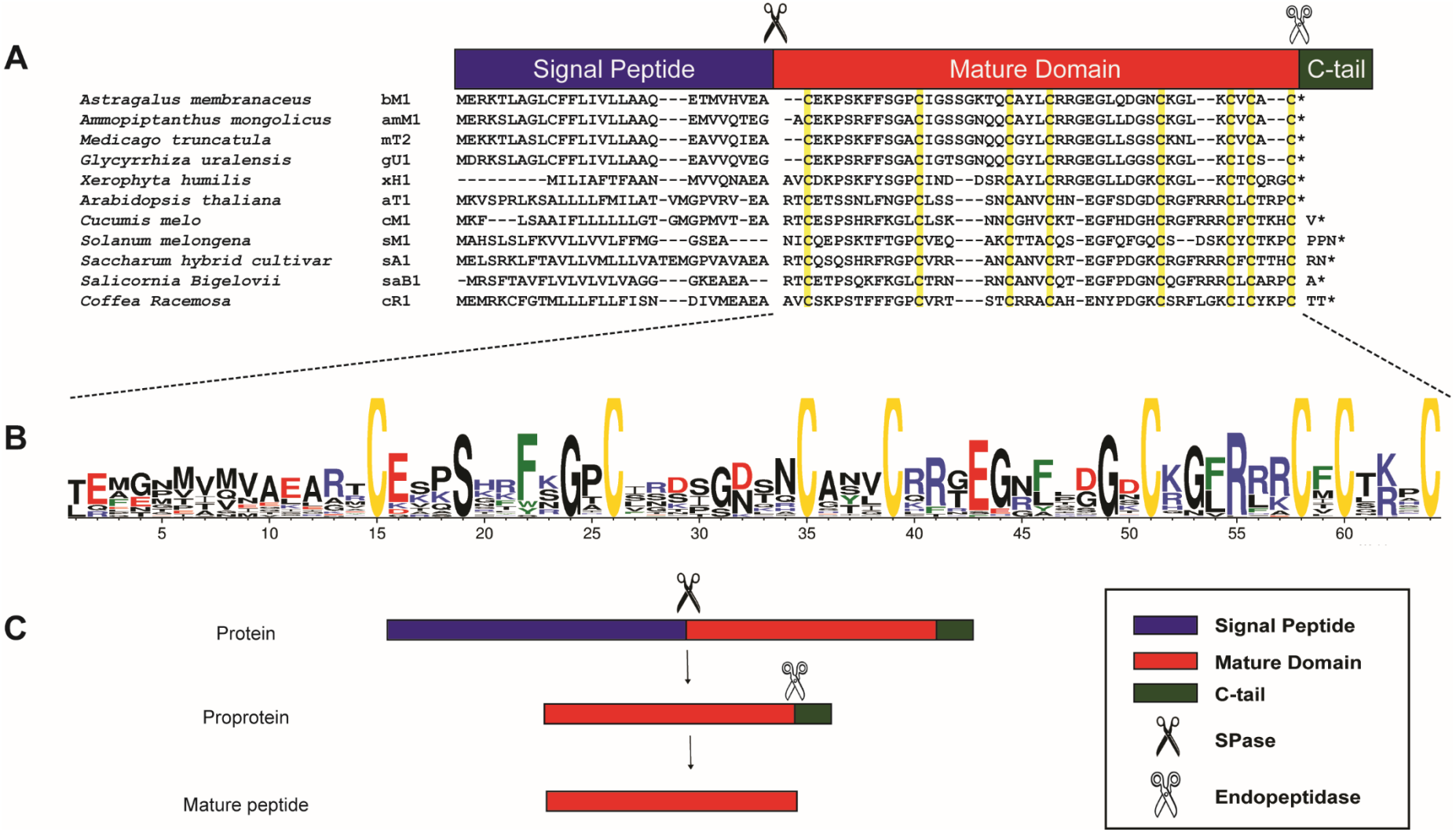



**Figure 4.10. Gene alignment and biosynthesis pathway.** (A) Two-domain precursor arrangement of *β*-astratide bM1 and other plant defensins. SPase is capable of removing the signal peptide from the full precursor sequences. (B) Sequence analysis of plant defensins. Accession numbers

of all the sequences are as follows: amm1, mt2, gu1, xh1, at1, cm1, sm1, sa1, sab1, cr1 (GenBank: JZ388864.1, BG452703.1, FS281873.1, JK691135.1, DR380981.1, JG547968.1, FS046114.1, CA259589.1, DY529894.1, GT663820.1) ∗ Represents the stop codon.

### 4.2.5. Evolution and origin of astratides

To reveal the evolutionary relationship between $\alpha$-astratide aM1, cyclotides, and other PA1b-like peptides, a phylogenetic analysis was performed based on the neighbor-joining clustering algorithm (Figure 4.11). Except for cP1, cP2, cP3, and aC1, which are PA1b-like peptides identified from the Convolvulaceae and Amaranthaceae families, the major PA1b-like peptides are distributed in the plant Fabaceae family. Genetic divergence within the plant phyla results from mutations in the mature peptide, which leads to functional diversification [21]. Although belonging to the same plant Fabaceae family, cliotides and other PA1b-like peptides are separated into two clusters based on the phylogenetic tree analysis. This result confirmed the hypothesis that the presence of chimeric structures in cliotides is a result of horizontal gene transfer between plant nuclear genomes or convergent evolution from PA1b-like peptides to cyclotides [306]. Furthermore, PA1b-like peptides in the Fabaceae family were separated based on different tribes, which were characterized by a limited range of floral characteristics, with a greater emphasis on petal morphology and stamen arrangement [307]. Importantly, most of the PA1b-like peptides are identified either from the tribe of pea (Fabeae) or from the tribe of soybean (Phaseoleae). In contrast, aM1 which belongs to a new tribe galegeae, is the first PA1b-like peptide to be discovered from medicinal plants. Such diversity of PA1b-like peptides may represent a new direction for discovering novel PA1b-like peptides [308].

Similarly, the evolutionary relationship between bM1 and other plant defensins was revealed by the phylogenetic analysis (Figure 4.12). Generally, plant defensins are classified into 8C- and 10C-plant defensins based on their cysteine numbers. Recently, Marilyn A. Anderson *et al*. further divided defensins into different subfamilies based on their cysteine motifs. In 8C-plant defensins, C-$X_{10}$-C-$X_5$-C-$X_3$-C-$X_{[9\text{-}10]}$-C-$X_{[6\text{-}8]}$-C-X-C-$X_3$-C is the most common cysteine spacing [41]. However, unlike the other reported two-domain plant defensins with a $CXCX_3C$ motif at the C-terminus, bM1 possesses a unique CXCXC motif. Additionally, the transcriptomic database search revealed that it is a unique motif that presents in plants of Fabaceae family. Previously, fabatin-2 from *Vicia faba* was the only reported plant defensin with this unique motif but lacked a detailed characterization [297]. Phylogenetic analysis revealed that two major clusters, Cluster 1 and 2, were formed based on their different C-terminal motif. Cluster 1 represents bM1 and the other plant defensins that contain the CXCXC motif at C-terminal from Fabaceae family, whereas Cluster 2 refers to plant defensins with the $CXCX_3C$ motif, suggesting that bM1 might belong to a new subfamily of plant defensins.

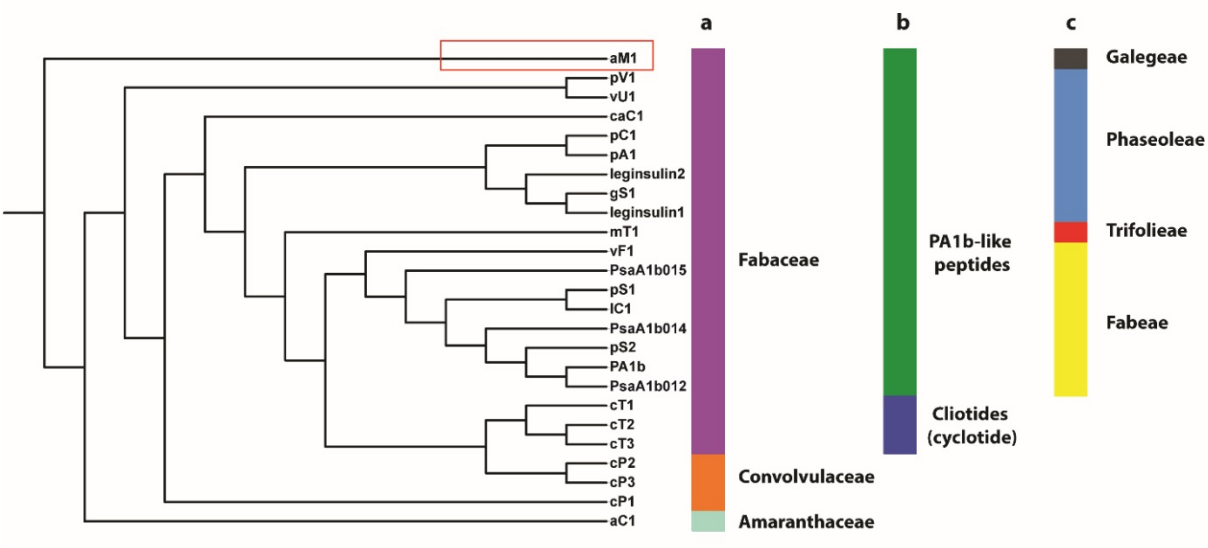



**Figure 4.11. Phylogenetic tree analysis of aM1, PA1b-like peptides, and cliotides.** The precursor sequences were aligned by MUSCLE. All the PA1b-like peptides are clustered separately based on (a) different plant families, (b) different CRP families, and (c) different tribes under the same family. A neighbor-joining clustering algorithm was used to construct a phylogenetic tree with a bootstrap test of 1000 replicates in MEGA6.

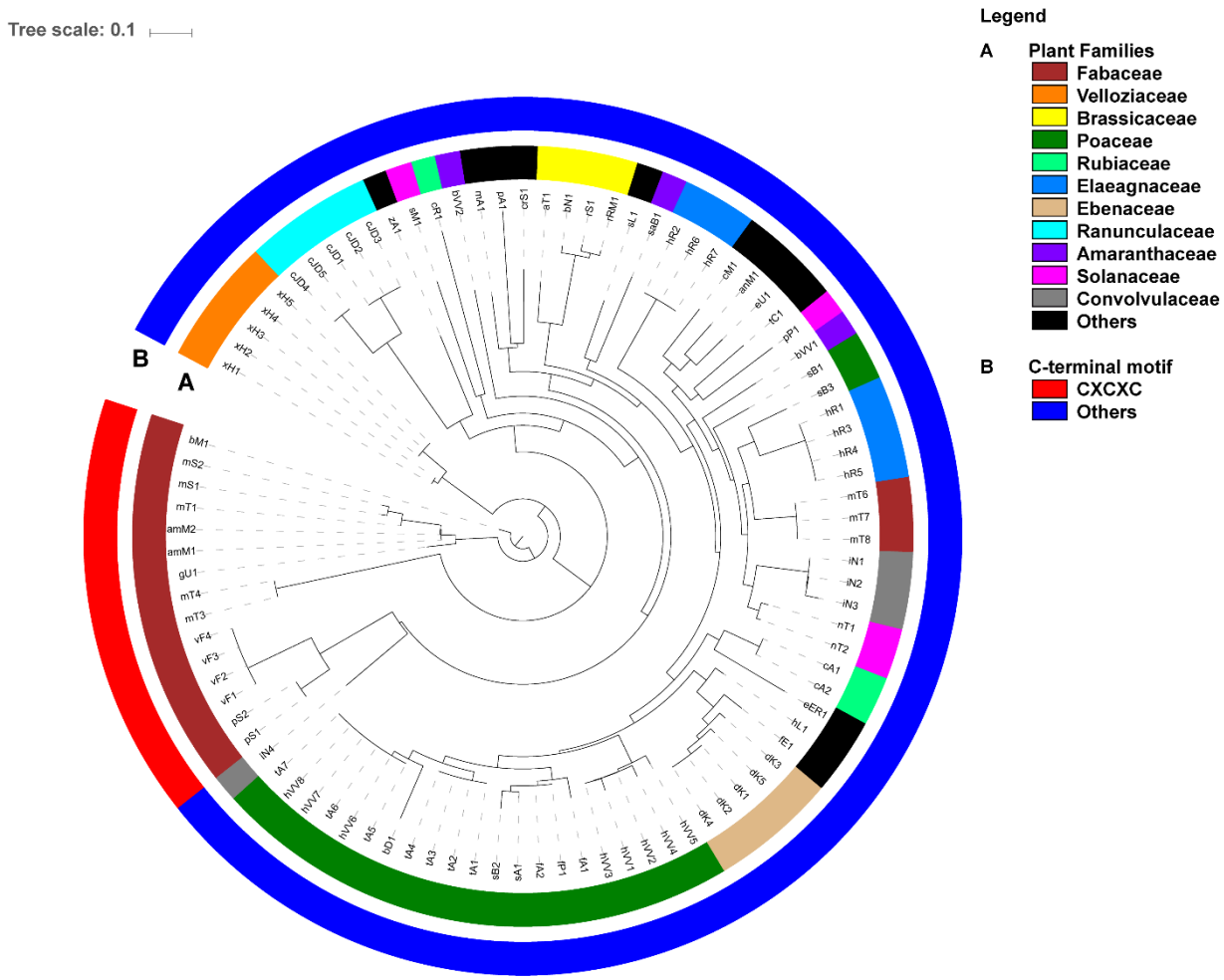



**Figure 4.12. Phylogenetic tree of bM1 and other defensin-like peptides.** A neighbor-joining clustering algorithm was employed to analyze the aligned mature domain of defensins. All the defensins are clustered separately due to the (A) different plant families and (B) different C-terminal motif. Uniprot accession number: LCR74 (Q9FFP8), LCR71 (P82781), LCR66 (Q9C947), J1-1 (Q43413), J1-2 (O65740), LCR72 (Q9ZUL8), VrD1 (Q6T418), LCR69 (Q39182), LCR68 (Q9ZUL7), LCR70(Q41914), PPT (Q40901), LCR75 (P82784), LCR76 (P82785), MtDef4 (G7L736), SD2 (P82659), NaD1 (Q8GTM0), NP-THN1 (O24115), Lc-def (B3F051), VrD2 (Q8W434), MsDef1 (Q9FPM3), LCR78 (P82787), RS-AFP1 (P69241), Rs-Apf2 (P30230), At-AFP1 (P30224), AhPDF1 (Q29SA6), AFP3 (Q39313), RS-AFP4 (O24331), LCR77 (Q9FI23) and Rs-AFP At2g26010 (O80995). GenBank accession number: pS1 (CAB82859.1), pS2 (AAA33638.1), PsDef 1 (EF455616.1) and Gamma-thionin 1(BAB19054.1).

### 4.2.6. Insecticidal activity of aM1

A cytotoxicity assay on insect Sf9 cells was used to evaluate the insecticidal activity of aM1, and an MTT assay was used to determine the cell viability. After a 24-hour incubation with different concentrations of aM1 (0.01-200 μM), the mortality rate of the Sf9 cells increased dose-dependently with an average lethal dose ($LD_{50}$) of 5.86 μM (Figure 4.13). In contrast, aM1showed non-toxic effect on the mammalian CHO-K1 cells at concentrations up to 200 $\mu$M, suggesting that the toxicity of aM1 is insect cell-specific.

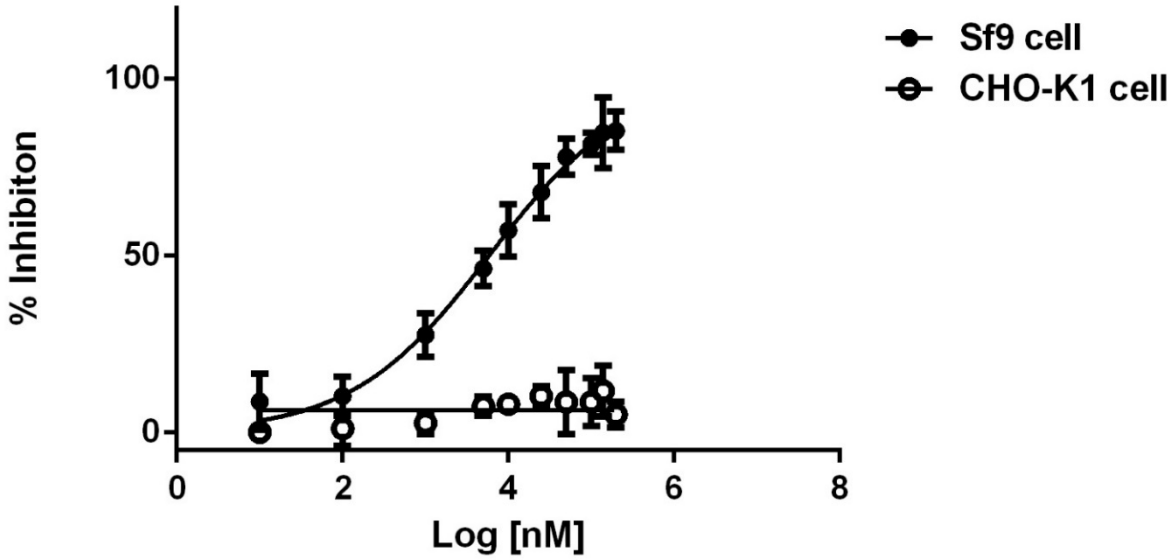



**Figure 4.13. The cytotoxicity of aM1 on Sf9 cells and CHO-K1 cells.** Different concentrations of aM1 were incubated with Sf9 cells and CHO-K1 cells for 24 h at 27 °C and 37 °C, respectively. To well display the graph, the x-axis was presented using log [nM] as unit, with the value of 1.00, 2.00, 3.00, 3.69, 4.00, 4.39, 4.69, 5, 5.15 and 5.30, respectively. The MTT assay was employed to determine the survival rate of the cells based on a colorimetric reaction.

To monitor the cytotoxic effect of aM1 on Sf9 cells, bright-field microscopy was employed. 5 μM of aM1 was incubated with Sf9 cells at 27 °C for 15 h, and the images were recorded at different time points. At 0 h, Sf9 cells showed a normal, spherical shape. At later time points, the cell membrane started to lose its integrity, leading to cell death (Figure 4.14A). Additionally, confocal microscopy showed that the incubation of aM1 destroyed the cell membrane of Sf9 cells and caused their death (Figure 4.14C). In contrast, control Sf9 cells showed no significant morphological changes during the incubation (Figure 4.14B and 4.14D).

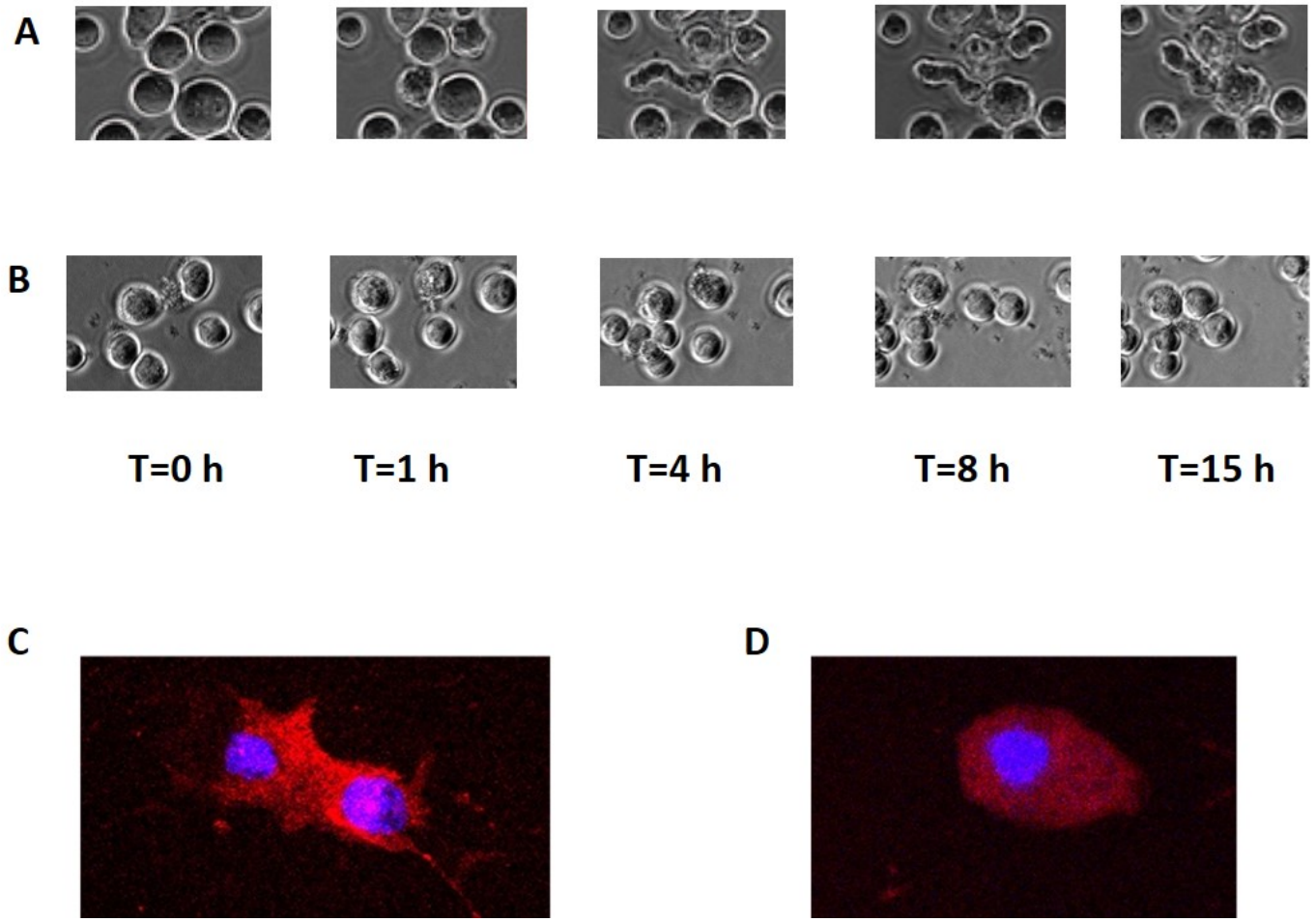



**Figure 4.14. Microscopy Phase-contrast image of Sf9 cells after incubating with (A) 5 µM aM1 and (B) 0.1% DMSO for 15 h, respectively.** The cell membrane of Sf9 cells was labeled with the membrane marker PKH26 (4 $\mu$M), and the nucleic acid was stained with Hoechst dye. Labeled cells were incubated with (C) control and (D) 5 $\mu$M aM1 labeled at free amines with NHS-AlexaFluor488 at 27 °C for 24 h.

Until now, only a few cystine-knot peptides were reported to possess insecticidal activity, such as cyclotides [309], the amaranthus $\alpha$-amylase inhibitor (AAI) and PA1b [87, 280]. However, the mechanisms of their insecticidal activities are different. AAI caused the death of insect larvae by inhibiting their $\alpha$-amylase whereas it is postulated that the activity of cyclotides is related to membrane binding [310]. Different from cyclotides and $\alpha$-amylase inhibitor, the insecticidal activity of PA1b is activated through a membrane protein-based receptor [305].

A high TM-align score (0.62) between the predicted structure of aM1 and the 3D structure of PA1b suggests that they share similar disulfide connectivity and secondary structure (Table 4.3). Thus, due to the high sequence and structural similarity, aM1 may exhibit its insecticidal activity through a mechanism similar to PA1b. Structural prediction showed that aM1 contains a cystine-knot structural fold with two anti-parallel β sheets but lacks the $\alpha$-helical turn compared to PA1b (Figure 4.15A). In PA1b, the hydrophobic loop 5 (Val25, Leu27, Val28, and Ile29) together with the opposite residue Phe10, formed a highly hydrophobic surface. On the contrary, the hydrophilic face, situated at the other pole of the molecule, is contributed by Ser2, Asn4, Thr17, Ser18, Asn34, and Ser36 [281]. Similarly, this amphipathic structural characteristic can be observed in aM1 (Figure 4.15B). A mutagenesis study on PA1b has shown that the four residues, Phe10, Arg21, Ile23, and Leu27 are important residues both for its binding activity and toxicity. Our results showed that aM1 contains the same Phe10, Arg21, Ile23, and Leu27 residues, suggesting that aM1 may exhibit the insecticidal activity through the same mechanism as PA1b, which is based on the interaction with a membrane-based receptor in insect cells [280].

**Table 4.3. TM align score between astratide aM1, PA1b, cystine knot $\alpha$-amylase inhibitors and cyclotides.**

| CRP family | Peptide | PDB | TM align score |
|---|---|---|---|
| Pea Albumin 1-b | PA1b | 1P8B | 0.6179 |
| Cystine knot $\alpha$-amylase inhibitors | AAI | 1QFD | 0.4631 |
| | Ac4 | 2MI9 | 0.3801 |
| Cyclotide | Kalata B1 | 2F2I | 0.3813 |
| | MCOTI-II | 1IB9 | 0.3321 |

*Pairwise alignment using astratide aM1 was performed. A TM-score>0.5 means that the two structures are in about the same fold, while value<0.3 means that they have a random similarity.

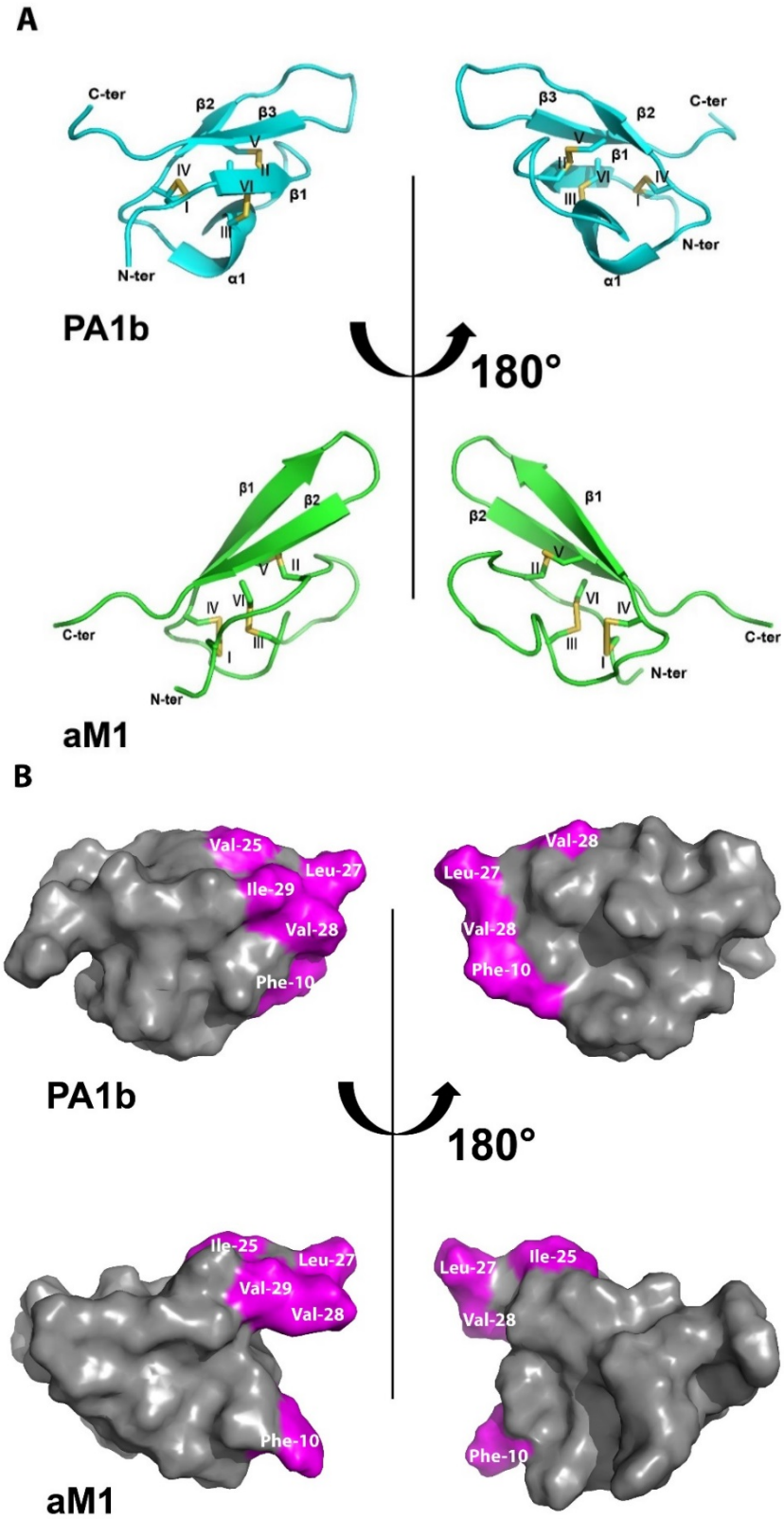



**Figure 4.15. 3D structure comparison of PA1b (PDB: 1P8B) and predicted aM1.** (A) The ribbon representation of the PA1b and predicted aM1 structure in two views. The disulfide bonds are formed between CysI-CysIV, CysII-CysV, and CysIII-CysVI. (B) Surface

representation showing the localization of residues of PA1b and the hydrophobic residues of PA1b and aM1, respectively. The hydrophobic residues are highlighted in pink.

### 4.2.7. Effect of aM1 on insulin secretion

To evaluated the effect of aM1 on insulin secretion, different concentrations of aM1 (0.01-100 μM) were co-incubated with mouse pancreatic β-TC cells for 24 h. 10 $\mu$M of GW9508, which is a compound that can potentiate insulin secretion in pancreatic $\beta$ cells, is used as a positive control in the study [311]. After incubation, the amount of insulin released by pancreatic $\beta$ cells was measured by ELISA. Figure 4.16 revealed that aM1 has no significant effect on insulin secretion of mouse pancreatic $\beta$ cells at low concentrations up to 1 $\mu$M. In contrast, aM1 was able to decrease the insulin secretion at high concentrations >1 $\mu$M in a dose-dependent manner. The insulin concentration decreased from 1225.20 $\mu$IU/mL to 811.14 $\mu$IU/mL with increasing concentrations of aM1 (1 $\mu$M to 100 $\mu$M).

Previous studies showed that >5 $\mu$g/g of PA1b was able to increase the blood glucose level of healthy C57BL/6 and type II diabetic mice. However, at low concentration (2.5 $\mu$g/g), no significant effect was observed [284]. In addition, PA1b was reported to have a hyperglycemic effect in mice upon binding to VDAC-1, an ion channel protein on the pancreatic $\beta$ cell membrane [285]. Thus, based on the sequence similarity between aM1 and PA1b, it can be postulated that aM1 could also bind to VDAC-1 and block the ion channel function to reduce the cation influx, leading to the reduction of insulin exocytosis. The results confirmed that aM1 could interfere with mammalian physiology in regulating glucose homeostasis.

To maintain glucose homeostasis, it is essential to keep a balance between the release and action of insulin since abnormal insulin secretion may lead to type 2 diabetes. Insufficient insulin secretion is the cause of hyperglycemia, whereas insulin hypersecretion would cause the beta cells to become exhausted, resulting in a reduced ability to respond to glucose stimuli and subsequent degeneration [312]. Hence the reduction of insulin secretion can restore the beta-cell function to increase the insulin sensitivity to beta cell stimulation. This proposed mechanism was supported by an effective anti-diabetic drug pioglitazone, which reduces insulin secretion. The decreased insulin secretion reduces the metabolism of beta cells, preventing it from exhausting, and reduces insulin resistance as well [313]. Therefore, it can be suspect that the function of aM1 on decreasing insulin secretion may follow a similar mechanism as pioglitazone. However, details of the putative bioactive mechanism of aM1 in regulating blood glucose remain unknown.

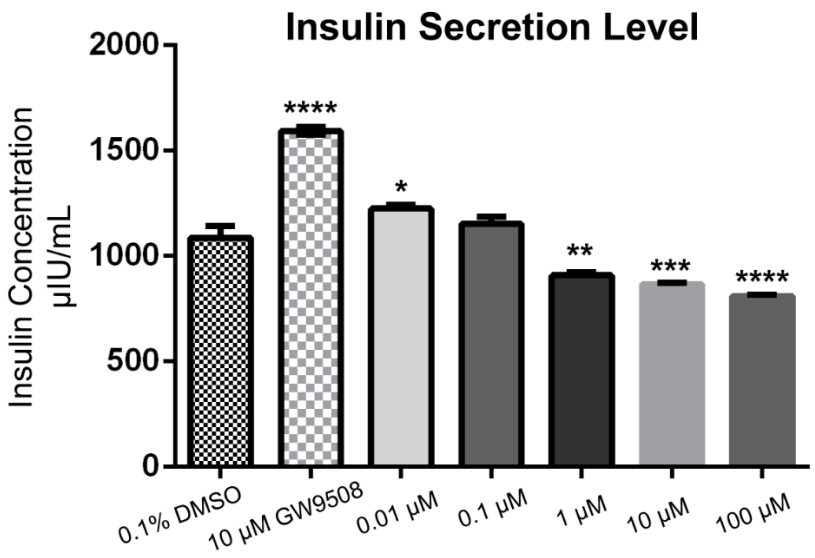



**Figure 4.16. Effect of aM1 on insulin secretion level in β-TC cells.** Different concentrations of GW9508 were incubated with β-TC cells for 24 h. 1, 10 and 100 $\mu$M of aM1 showed a significant decrease in the insulin secretion level in mouse pancreatic β-TC cells as compared with 0.1% DMSO. 0.01, 0.1 $\mu$M of aM1 and 10 $\mu$M of GW9508 increased the insulin secretion level as compared to control. (*p<0.05, **p < 0.01, ***p < 0.001, ****p<0.0001).

### 4.2.8. Predicted structure of bM1

Figure 4.17A showed the predicted 3D structure of bM1 by I-TASSER. Three disulfide bonds have been predicted as CysII-CysV, CysIII-CysVI, and CysIV-CysVII, together with two antiparallel β-strands and one α-helix. It can be elucidated that the fourth disulfide bond was linked by Cys I-CysVIII. The surface topology of bM1 was shown in Figure 4.17B. Hydrophobic residues (P, F, G, I, A, L, and V), acidic residues (E and D) and basic residues (K and R) are highlighted in pink, red and blue, respectively.

Depending on the primary, secondary, and tertiary structure, defensins have been classified into two distinct superfamilies, *cis*-defensin and *trans*-defensin. The dominant superfamily in plant defensins are the *cis*-defensins, which adopts a structure comprising two parallel disulfide bonds that bond final β-strand to an α-helix. In contrast, members from *trans*-defensin superfamily contain two analogous disulfide bonds that point in opposite directions from the final β-strand [287]. Based on the predicted structure of bM1, in which the disulfide bonds link the final β-strand to an α-helix, it is obvious that bM1 belongs to the *cis*-defensin superfamily.

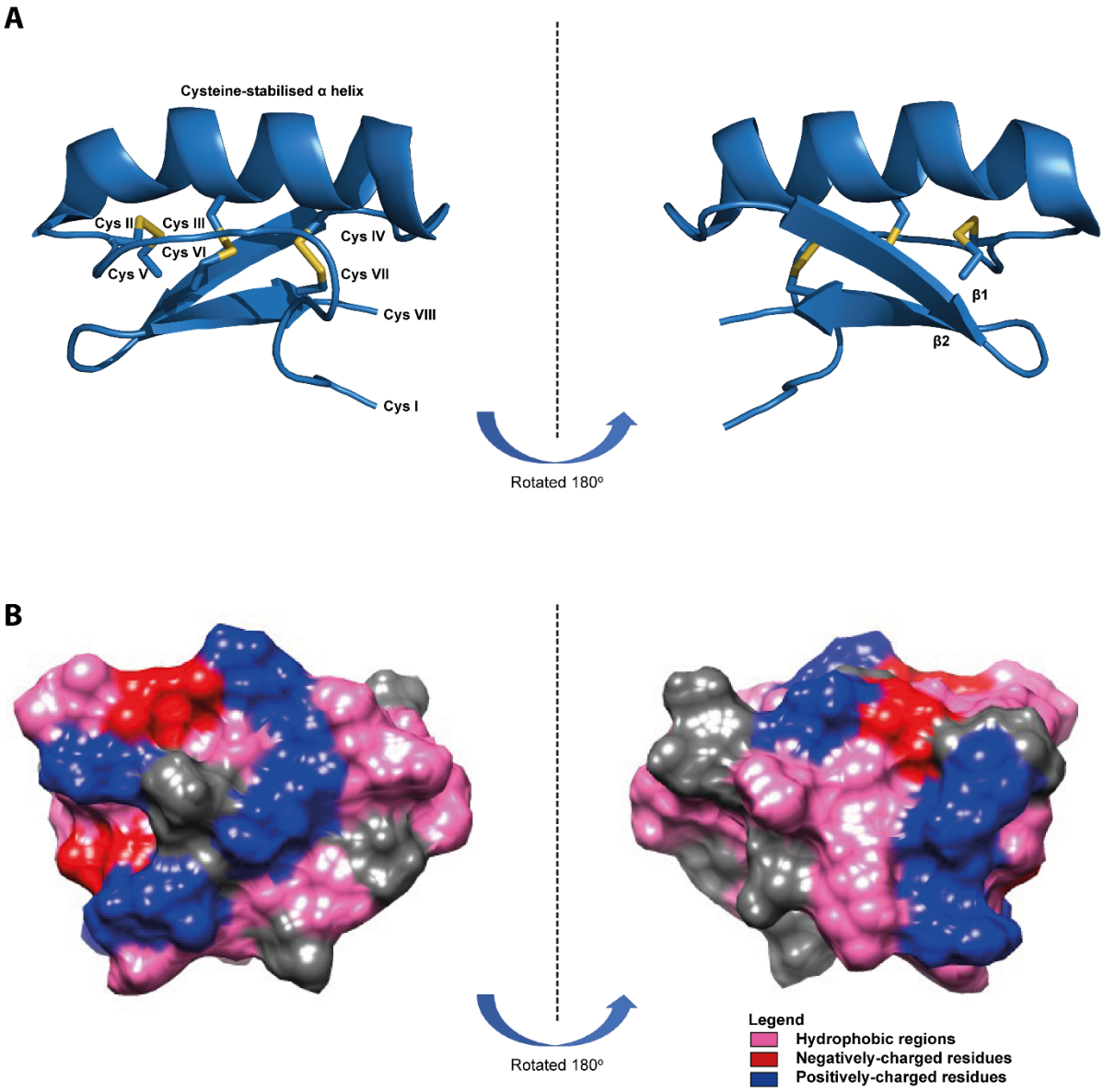



**Figure 4.17. Predicted structure of bM1.** (A) Cartoon representation of bM1 structure in two views. (B) Surface representation of the bM1 predicted structure. Hydrophobic, acidic and basic residues are indicated in pink, red, and blue respectively.

### 4.2.9. Anti-fungal activity of bM1

The inhibitory effect of bM1 on the growth of four phytopathogenic fungal strains was evaluated using a disc diffusion assay. Fungi strains were incubated at 30 °C for 24 to 72 h until a radical colony was formed. 20 µL of different concentrations of bM1 (1, 5 and 10 mg/mL) and milli-Q water were placed on discs at the end of growing mycelia. After 24 h incubation, the formation of crescent-shaped inhibition zones was observed, which demonstrated that bM1 exerted antifungal activity against all four strains (Figure 4.18A). The

half-maximal inhibitory concentration ($IC_{50}$) was calculated at the range from 2.7 $\mu$g/mL to 130.3 $\mu$g/mL after a 24 h incubation at 30 °C, for the four strains (Figure 4.18B). Figure 4.18C revealed the morphological changes of *A. alternate* fungal spores prior to and after treatment with different concentrations of bM1. The treatment resulted in shorter and highly branched hyphae compared to the control experiment. Vacuolar granulation, swollen hyphal tips and restarted budding hyphae were also observed, in a dose-dependent manner.

Generally, plant defensins are shown to have strong antifungal activities. For example, Ms-Def1 from *Medicago sativa* seeds, Rs-AFPs from *Raphanus sativus* radish seeds and Ct-AMP1 from *Clitoria ternatea* are reported to have a broad antifungal spectrum with an $IC_{50}$=0.3−100 µg/mL [12]. The mechanisms related to the inhibition activity of fungal growth can be explained in three ways, internalization mediated by receptors, membrane translocation, and membrane permeabilization [314]. Previous studies have reported that plant defensins are capable of interacting with host membrane components specifically and thus lead to cell death. Alternatively, plant defensins can interact with intracellular targets through the internalization with the fungal wall and exhibit the inhibitory effects [315]. However, the mechanism of action of bM1 remains unclear, and further experiments are needed to clarify this issue.

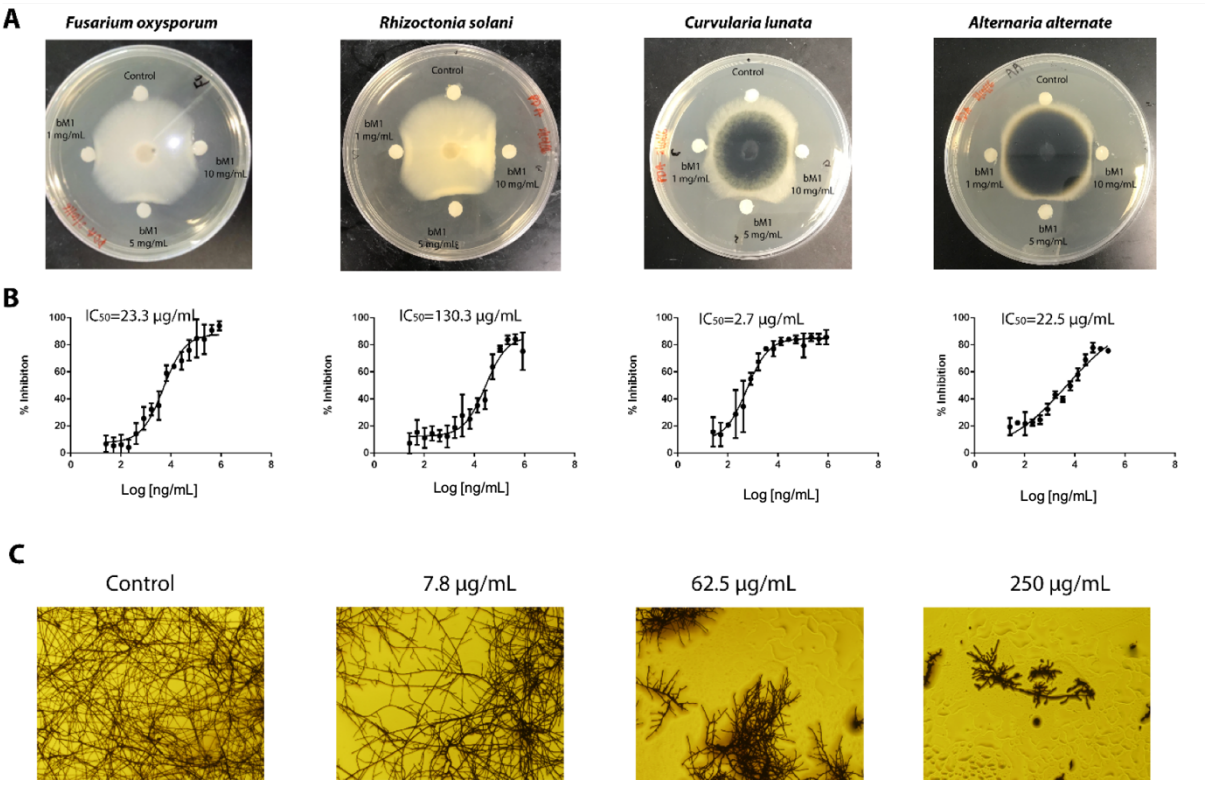


**Figure 4.18. Anti-fungal activity of bM1 towards four phytopathogenic fungal strains.** (A) Formation of arc-shaped inhibition zone of fungal mycelia indicates the susceptibility of *F. oxysporum, R. solani, C. lunata* and *A. alternate* to different concentrations of bM1. (B) Dose-response curves generated from the microbroth dilution assay were used to calculate the $IC_{50}$ values. (C) Microscopic view of the mycelium growth of *A. alternate* with different concentrations of bM1. The well without the treatment of bM1 was served as a control.

### 4.2.10. Metabolic stability of astratides

The metabolic stability of aM1 and bM1 against thermal, chemical, proteolytic degradation was illustrated by several stability tests. Figure 4.19A revealed that that >80% of aM1 remained stable after incubation 100 °C for 2 h. Figure 4.19B revealed that aM1 remained stable in an acidic condition for 2 h, with 98% of the peptide remaining (Figure 4.19B). Similarly, aM1 was resistant to tryptic degradation with >88% of the peptide remaining after a 6-h incubation with pepsin and aminopeptidase I (Figure 4.19C and D). In addition, > 90% of aM1 remained intact after incubation with human serum for 36 h (Figure 4.19E). A similar situation was observed in $\beta$-astratide bM1. It showed high stability against thermal, enzymatic, chemical, and human serum degradation with peptide remaining > 90% (Figure 4.20). Prolonging the thermal, acid and enzymatic treatment does not change the trend of degradation. This is in agreement with other plant CRPs that contain a compact structure, whose thermal, acidic stability were examined up to 2 h treatment and enzymatic stability examined up to 6 h treatment [20, 26, 27]. The results suggest that they may serve as the active components in *A. membranaceus* roots and have the potential to be orally active drug leads.

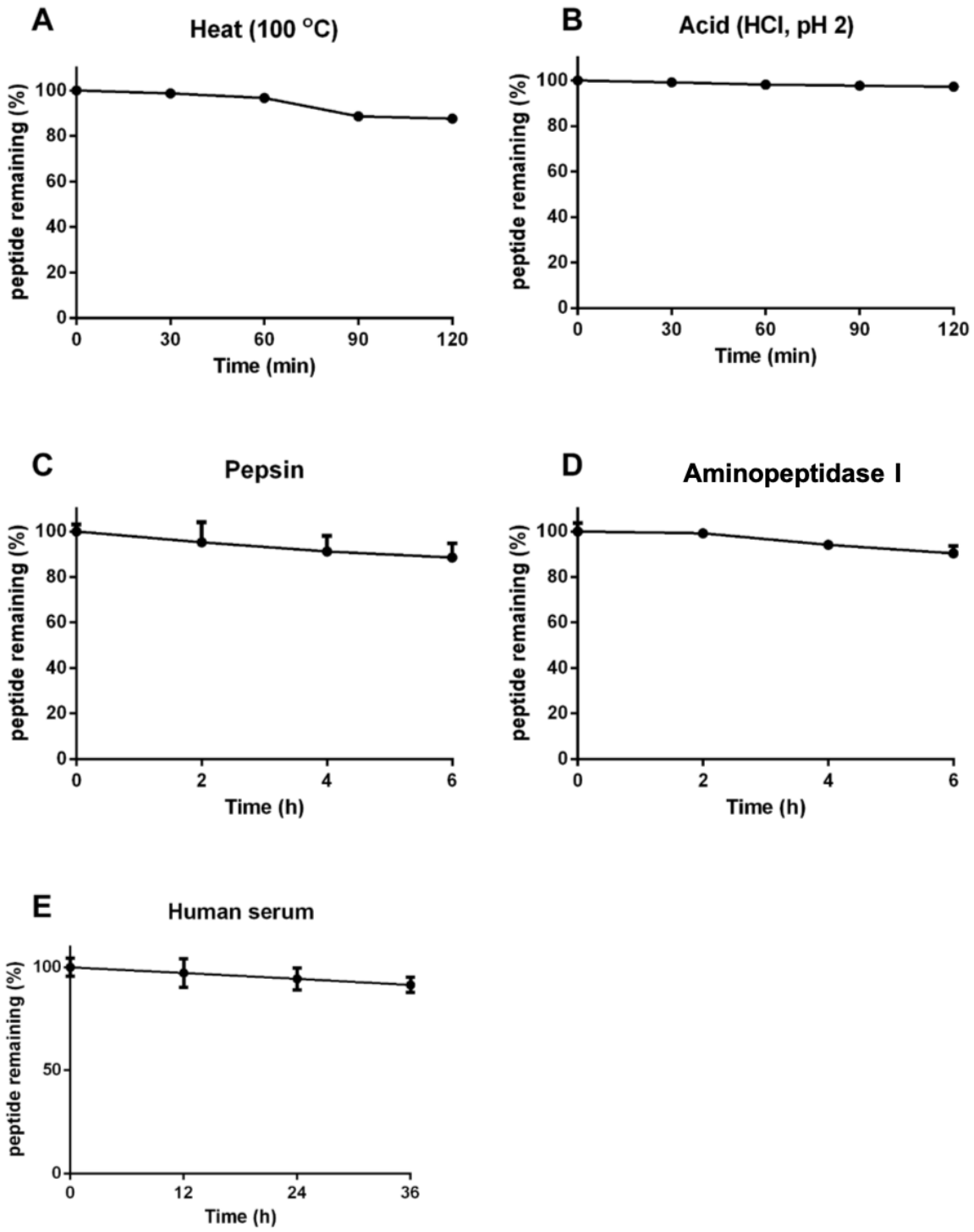



**Figure 4.19. Stability assays of *α*-astratide aM1.** (A) Thermal stability of aM1 against heat condition for 2 h. (B) Acidic condition stability of aM1 incubated in HCl (pH 2.0) for 2 h. (C) Enzymatic stability of aM1 against pepsin for 6 h. (D) Enzymatic stability of aM1 against aminopeptidase I for 6 h. (E) Human serum-mediated stability of aM1 at 37 °C for 36 h.

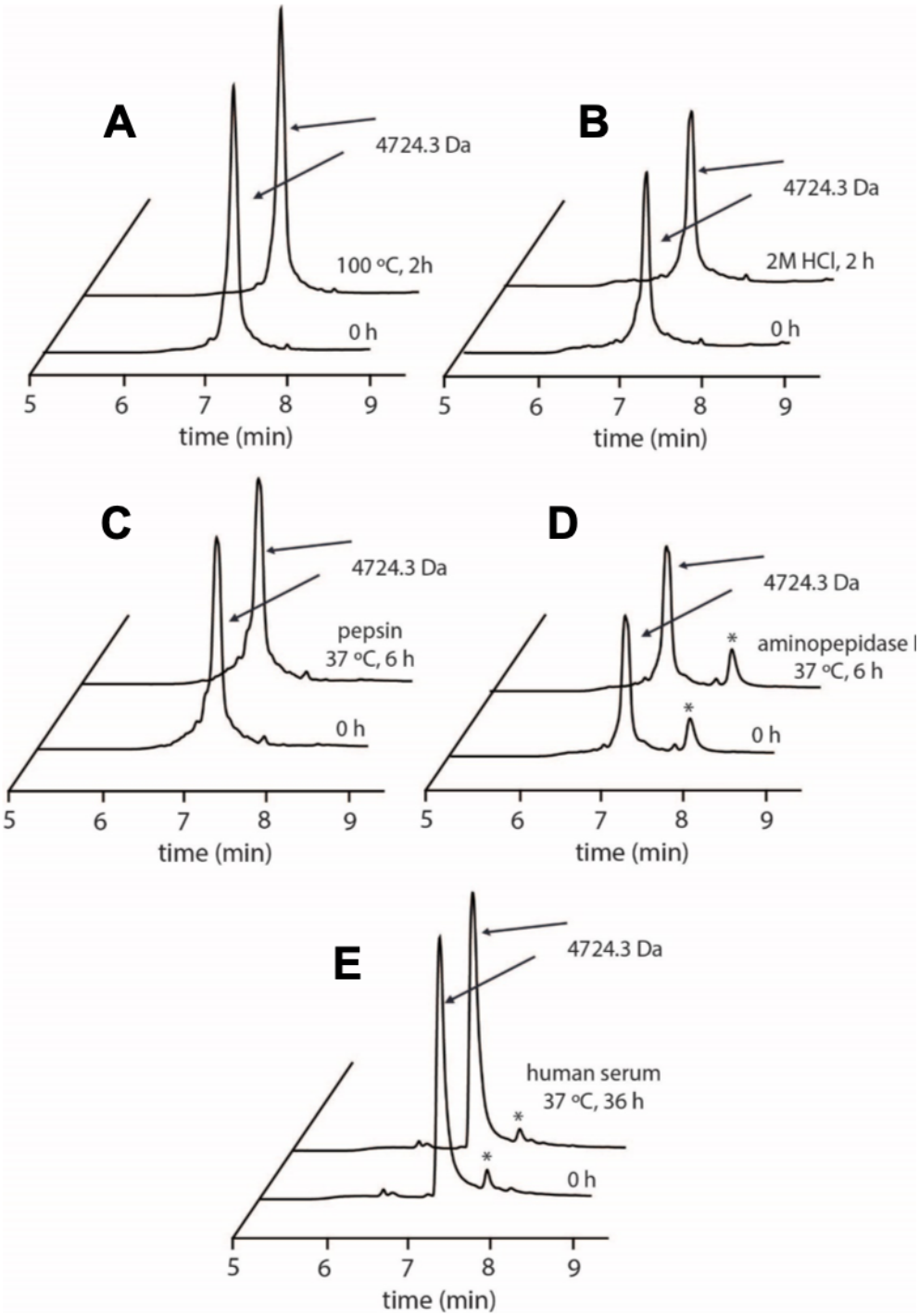



**Figure 4.20. Stability assays of *β*-astratide bM1.** (A) Stability of bM1 under heat condition at 100 °C for 2 h. (B) Stability of bM1 under an acidic condition in HCl (pH 2.0) for 2 h. (C) Stability of bM1 against endopeptidase pepsin for 6 h in buffer at 37 °C. (D) Stability of bM1 against exopeptidase aminopeptidase I for 6 h in the buffer as suggest by manufactures at 37 °C. (E) Stability of bM1 in human serum at 37 °C for 36 h.

## 4.3. Conclusion

This chapter described the discovery, isolation, and identification of two CRPs, namely $\alpha$-astratide aM1 and $\beta$-astratide bM1, from the *A. membranaceus* roots. Proteomic analysis revealed that $\alpha$-astratide aM1 is a 6C-CRP belonging to the family of PA1b-like peptides, whereas $\beta$-astratide bM1 is an 8C-CRP belonging to the family of plant defensins. Both astratides displayed high metabolic stability against temperature and chemical changes. Data mining revealed a five-domain precursor arrangement in astratide aM1, which consists of a signal domain, a mature peptide domain, a hinge domain, a PA1a-like domain, and a C-terminal domain. The five-domain architecture is different from other 6C-CRPs with a three- to four-domain precursor arrangement. Moreover, phylogenetic analysis showed that aM1 is a novel PA1b-like peptide identified from a new tribe of the Fabaceae family. Bioassays showed that aM1 is cytotoxic against insect Sf9 cells and capable of reducing insulin secretion in normal mouse pancreatic β cells at high concentrations. On the contrary, $\beta$-astratide bM1 shares similar cysteine spacing with plant defensins and contains a typical two-domain defensin-like precursor. The phylogenetic tree revealed that bM1 is a new plant defensin with unusual C-terminal motif and has strong anti-fungal activity against four phytopathogenic fungi strains. In conclusion, our study enlarges the existing library of PA1b-like peptides and plant defensins, exploring their sequence diversity, structure, biosynthesis pathway, and evolutionary relationship and suggesting the role of CRPs as an active compound in plants.

# Chapter 5 Coffeetides: conversion of coffee waste to value-added non-chitin-binding hevein-like peptides

## 5.1. Introduction

*Coffea* belongs to the Rubiaceae family and is a genus of flowering plants whose seeds (coffee beans) are used to produce a variety of coffee beverages and products [178]. There are more than 100 species of *Coffea*, among which, *C. canephora* (synonyms: *C. robusta*) is the second most cultivated species in the world, accounting for approximately 40% of coffee production [316]. Originated from tropical forests, it is more adapted to harsh conditions and hence can reach great crop yield by cultivating at low cost. Different parts of coffee plants have been used as medical treatment all around the world for a long time. The coffee seeds decoction has been used to treat influenza, to increase milk production for nursing mothers and used as a cardiotonic as well as a neurotonic [179]. Leaf decoction was used to treat anemia, edema, asthenia, and rage [179], while the leaf poultice can be used for fever. For the coffee fruit, its decoction can be used for hepatitis and used as a stimulant for sleepiness and drunkenness [179]. Coffee was reported to have antioxidant [317, 318], anti-diabetic activity[186-188] and have effects on cardiovascular diseases [184, 185].

Being the second-largest worldwide commodity, coffee serves as a source of income for millions of people and represents an important part as foreign currency [180]. According to ICO, approximately 130 million bags of coffee brew were produced in 2011-12 seasons [178]. However, nearly more than 50% of waste was generated during industrial processing, which includes solid residues like coffee pulps, husks, mucilage, and silver skins [178]. Based on the traditional medicinal value of coffee plants, it is promising to discover the biological functions of coffee waste, which can help to utilize coffee waste and reduce environmental damage.

Previous researches of active ingredients in coffee mainly focus on small molecules such as caffeine [190-192, 319], diterpene [193, 194], and chlorogenic acid [195, 196, 320]. To date, the best known are caffeine, which is an alkaloid widely consumed as a psychoactive drug. However, no disulfide-constrained peptides within 2-6 kDa have been reported for coffee species. Peptide and protein-driven compounds, unlike small metabolites, are normally not considered as bioactive components in medicinal plants [12, 277, 321]. This bias is due to the common perception that they are unstable during the decoction procedure in traditional medicine and their easy degradation in the gastrointestinal tract after ingestion. However, recent findings suggest otherwise. A great number of studies showed that cysteine-rich peptides (CRPs) are native compounds rich in cysteines, a feature that enables the formation of multiple

disulfide bonds. The presence of these disulfide bonds contributes to a compact structure and provides high resistance to chemical and proteolytic degradation [12, 113, 322]. CRPs in plants are involved defense against bacteria [323], insects [309] and fungi [324]. With a molecular weight ranging from 2-6 kDa, containing 3-5 disulfide bonds, CRPs represent a new class of underexplored biologics in medicinal plants due to their stability of thermal and proteolytic degradation.

Based on different cysteine motifs and distinct disulfide bond patterns, CRPs within 2-6 kDa can be divided into different families [12]. According to these criteria, the CRP families have been classified as thionins, defensins, hevein-like peptides (HLPs) and knottin-type peptides [12]. These families can be further divided based on cysteine numbers. Thus far, there are three major CRP families identified in 8C-CRPs, which include 8C-defensins, 8C-thionins, and 8C-HLPs [21, 132, 288, 309]. 8C-HLPs can be characterized by an evolutionary conserved –CC- motif and can be further classified into two subfamilies based on the presence or absence of a chitin-biding domain. The chitin-binding 8C-HLPs has been identified and reported from many plant species, such as morintides from *Moringa oleifera*, vaccatides from *Vaccaria hispanica* and ginkgotides from *Ginkgo biloba* [21, 27, 57]. The common feature of these 8C-HLPs is the presence of a chitin-binding domain, which confers the peptide ability to bind to chitin, a major constituent of fungal walls. The non-chitin-binding 8C-HLPs, in contrast, although containing the same tandemly-connecting cysteine motif, they lack a chitin-binding domain. Until now, only ginsentides that isolated from ginseng species have been reported from this subfamily, which has raised the question of whether non-chitin-binding 8C-HLPs are distributed in other plant families.

In this chapter, data mining revealed the presence of 89 putative non-chitin-binding HLPs from 12 plant families at the transcriptomic level. To validate this finding, I report the discovery and characterization of a new subfamily of non-chitin-binding 8C-HLPs, designated as coffeetides, from the husks (coffee waste) of *C. canephora* and *C. liberica* of Rubiaceae family. Proteomic, transcriptomic analyses and disulfide mapping revealed that coffeetides contain the same cysteine motif and disulfide connectivity as ginsentides. The phylogenetic tree showed coffeetides form a distinct cluster that is different from other non-chitin-binding 8C-HLPs, which may due to the presence of a shorter pro-peptide in the precursor sequences. Additionally, NMR structure determination showed that coffeetides contain a highly compact pseudocyclic structure that confers them stability against harsh conditions. Furthermore, biological studies showed that coffeetide cC1 is non-cytotoxic peptides that could increase cell migration and enhance the metabolism of human neuroblastoma SH-SY5Y cells. Taken together, our finding of coffeetides expands the existing library of CRPs. More importantly, these bioactive

compounds discovered from coffee waste could have therapeutic importance and provide a way to convert industrial waste into value-added products.

## 5.2. Results

### 5.2.1. Screening of putative non-chitin-binding 8C-hevein-like peptides

Previously our lab has reported a novel family of non-chitin-binding 8C-hevein-like peptides, ginsentides from three ginseng species: *Panax ginseng*, *Panax quinquefolius,* and *Panax notoginseng*. To investigate the presence of putative non-chitin binding 8C-HLPs in plant species other than ginseng, ginsentide sequences were used as queries for the tBLASTn transcriptomic database search. A total number of 89 putative non-chitin binding 8C-HLPs, containing 31-34 amino acids and a conserved cysteine motif of CXnCXnCCXnCXnCXnCXnC but lacking a chitin-binding domain, were identified from 24 plant species in 12 different plant families. A phylogenetic tree was established based on the aligned precursor sequences of 8C-CRPs using a neighbor-joining clustering algorithm (Figure 5.1). It can be observed that these 8C-CRPs are separated into three major clusters: 8C-HLPs (highlighted in green in Figure 5.1A), 8C-thionins and 8C-defensins. Figure 3.1B indicated that two different clusters were observed in 8C-HLPs: chitin-binding 8C-HLPs and non-chitin-binding 8C-HLPs, which includes ginsentides and ginsentides-like peptides. The 99 ginsentides and putative non-chitin-binding 8C-HLPs were separated into 13 clusters based on different plant families (Figure 5.1C).

To validate this finding that the putative non-chitin binding 8C-HLPs are distributed in other plant families, five important crops or medicinal plants (*Coffea canephora*, *Coffea liberica*, *Theobroma cacao*, *Triticum aestivum,* and *Eleutherococcus trifoliatus*) were selected for screening using MALDI-TOF MS. According to our previous protocol, a mass shift of 58 Da for each cysteine residues before and after the *S*-reduction with DTT and *S*-alkylation with IAA confirmed the presence of putative CRPs [38]. Our results showed that a mass shift of 464 Da was observed in the peptides from these five species after the *S*-alkylation (Figure 5.2), suggesting that they contained eight cysteine residues. Due to the abundance, the putative non-chitin binding 8C-HLPs present in *Coffea* species were selected for further characterization.

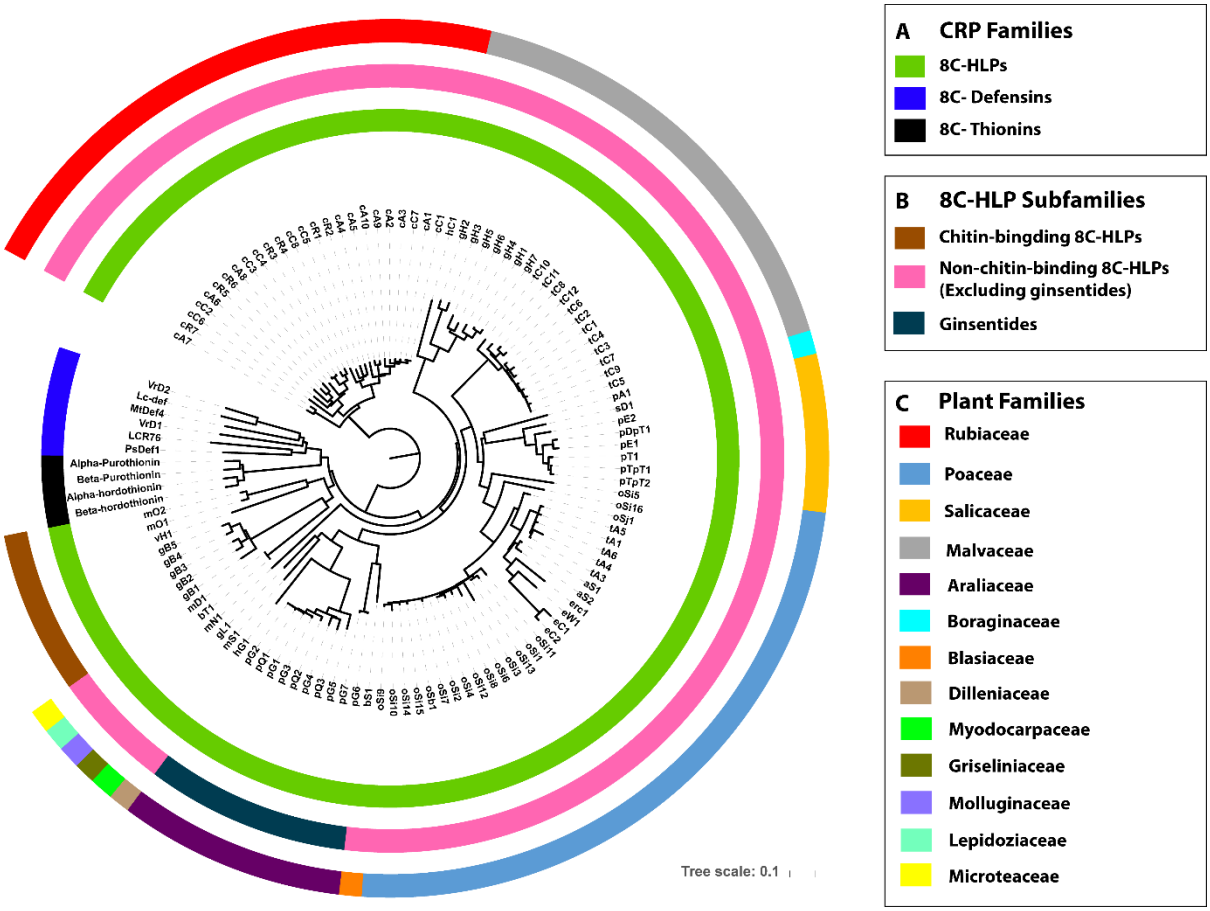



**Figure 5.1. Phylogenetic tree of 8C-CRPs.** The alignment of precursor sequences was accomplished by MUSCLE. The phylogenetic tree was displayed using iTOL v3. Classification is based on (A) CRP family, (B) HLP subfamily and (C) plant's family.

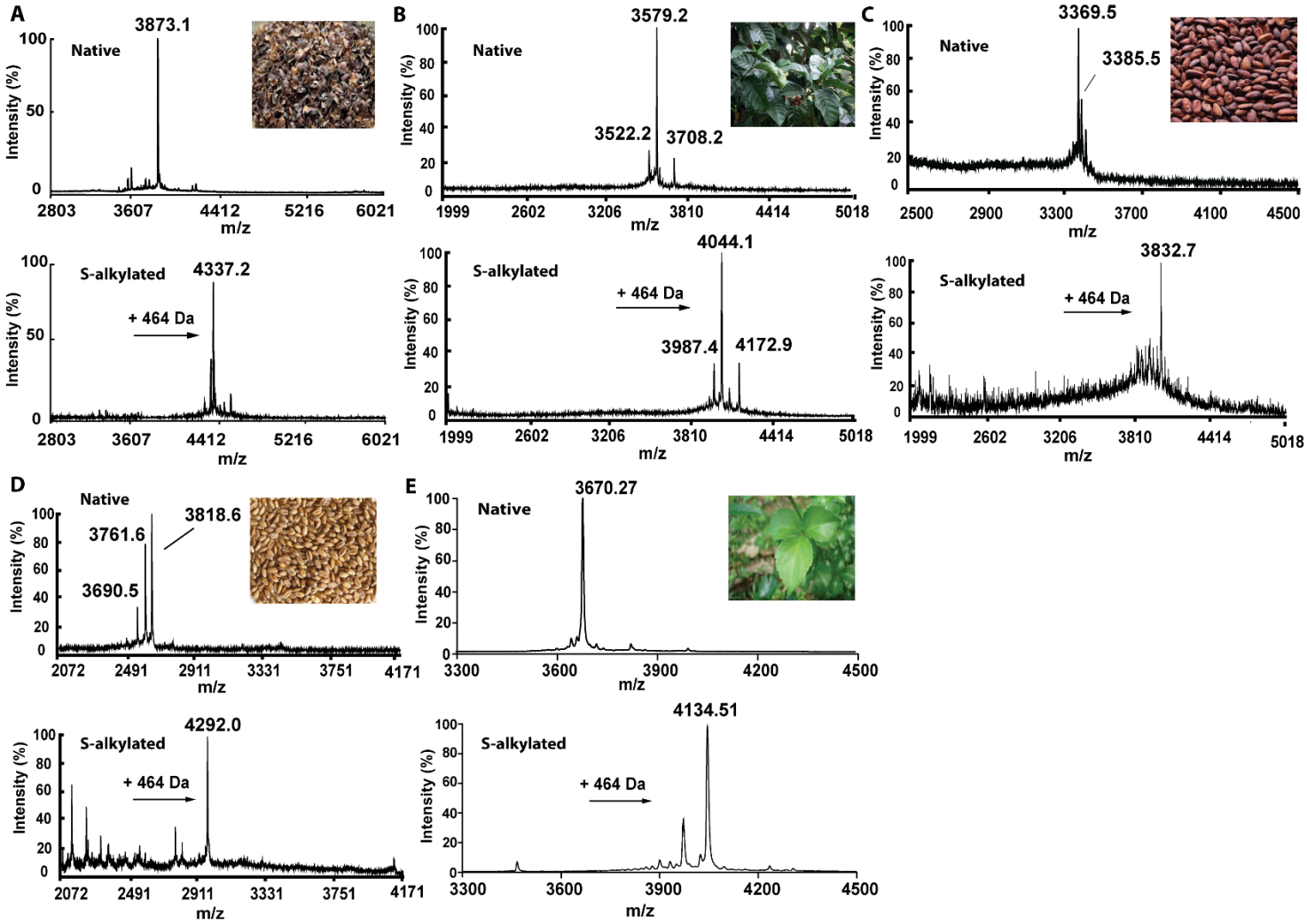



**Figure 5.2. MS profiles of five selected plants.** MS profiles of (A) *Coffea canephora*, (B) *Coffea liberica*, (C) *Theobroma cacao*, (D) *Triticum aestivum*, (E) *Eleutherococcus trifoliatus* and (F*) Eleutherococcus senticosus* before and after S-reduction and S-alkylation. A mass shift of 464 Da can be observed in all peptides after S-alkylation.

### 5.2.2. Isolation and sequence characterization of coffeetides from *C.canephora*

1 kg of *C.canephora* husks were used for a large-scale extraction using 10 L of milli-Q water. After multiple rounds of AEX- and RP-HPLC, three putative CRPs, designated as coffeetides cC1, cC1b and cC1c with relative monoisotopic molecular masses $[M + H]^+$ of 3873.1, 3764.3 and 3633.4 Da, respectively, were isolated (Figure 5.3).

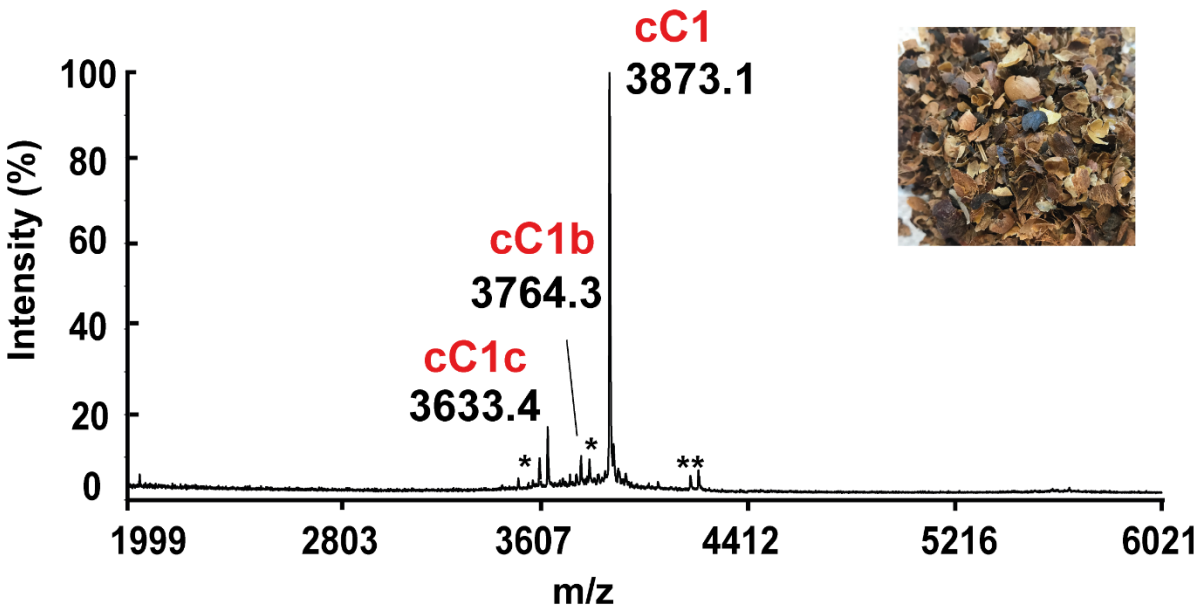



**Figure 5.3. MALDI-TOF MS profile of husks of C. canephora.** MS profile of *C. canephora* husks ranging from 2-6 kDa. Coffeetide cC1, cC1b, and cC1c are labeled at the top of the corresponding peaks. * Unknown compounds.

LTQ Orbitrap Elite MS/MS was employed to determine the primary sequences of coffeetides. The fully S-alkylated coffeetides were analyzed by nanospray tandem MS and their primary sequences were deduced by evaluating the mass differences between the *b*- and *y*-series ions (Figure 5.4). Subsequently, the coffeetide cC1 sequence was deduced as ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC, with an N-terminal pyroglutamic acid (Z; Figure 5.4A). The assignment of isobaric amino acids Leu/Ile or Lys/Gln was achieved by analyzing the cDNA obtained from GenBank. Coffeetide cC1 contains 37 amino acids in length and is rich in cysteine (eight residues). The primary sequence of cC1b and cC1c were determined in the same manner (Figure 5.4B and 5.4C). Coffeetide cC1b and cC1c were found to be the truncated version of cC1, with one or two deletion(s) of the amino acid at the N-terminus. Hence coffeetide cC1 was selected as the representative for further characterization.

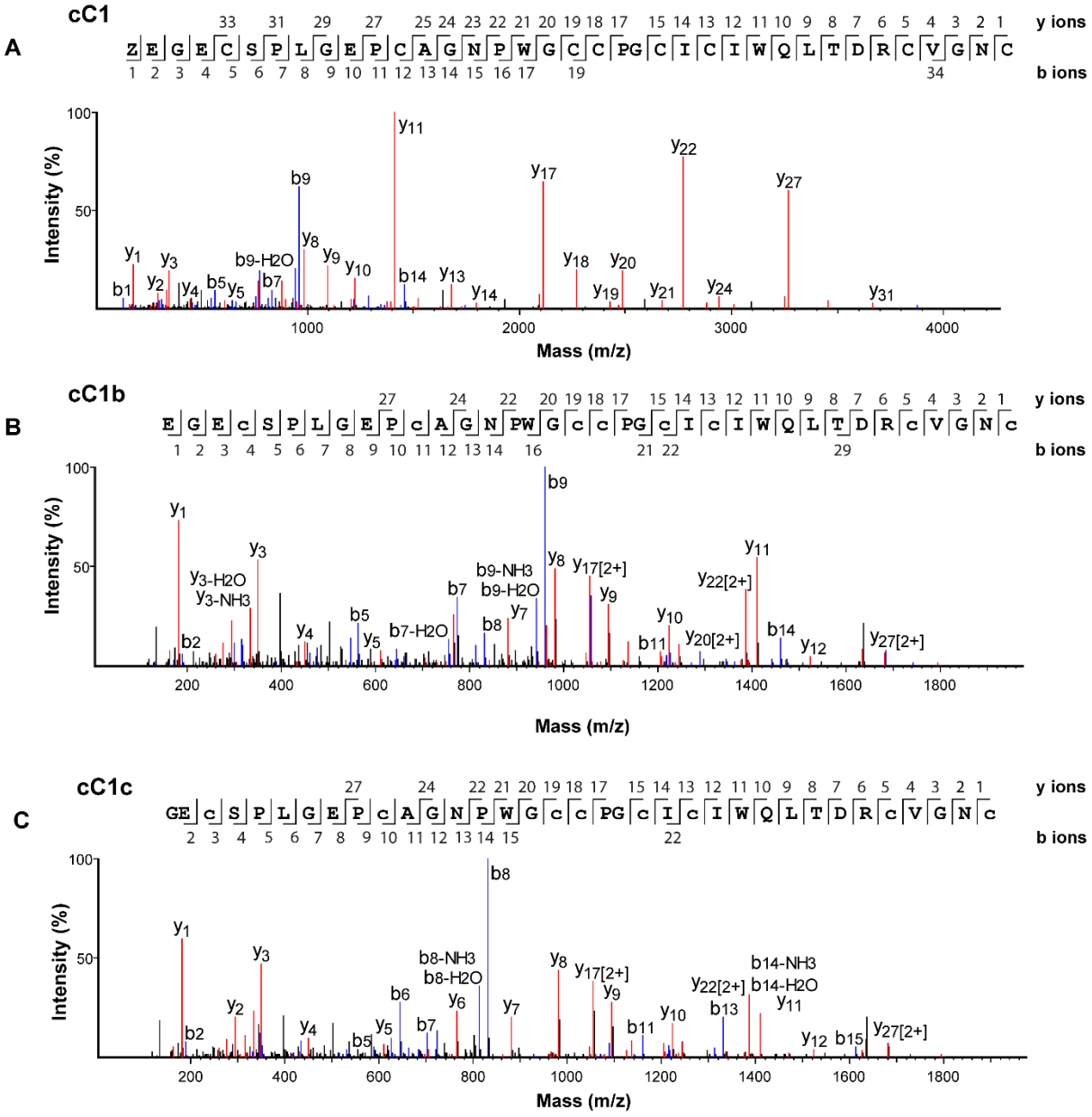



**Figure 5.4. Mass spectra of coffeetides from LC-ESI-LTQ-Orbitrap MS/MS in positive ion mode.** Mass spectra of (A) cC1, (B) cC1b and (c) cC1c were scanned between mass ranges from 0 to 4000, 1800 and 1800 m/z, respectively. Transcriptomic data was employed to confirm the assignment of isobaric amino acids such as Leu/IIe.

### 5.2.3. Isolation of coffeetides from *C. liberica*

To understand the presence of coffeetides in different *Coffea* species and different tissues, we performed a scale-up extraction of the husks (1 kg) of *C. canephora* and the leaves (1 kg) of *C. liberica* using 5 L of milli-Q water. Subsequently, several rounds of SCX- and RP-HPLC was performed to purify the peptides. Four putative CRPs were isolated from *C. liberica* leaves,

and designated cL1, cL1b, cL1c, and cL2, with relative monoisotopic molecular masses $[M + H]^+$ of 3819.2, 3708.2, 3579.1 and 3646.1 Da, respectively (Figure 5.5A). The same methods were used to extract the *C. liberica* husks, and two putative CRPs were observed in the MS profile (Figure 5.5B).

*S*-reduction and *S*-alkylation were employed to confirm the number of cysteines present in each peptide. Figure 5.6 showed that cL1, cL1b, cL1c, and cL2 are coffeetides containing eight cysteines while two peptides with the molecular mass of 3722.1 and 3736.1 Da are 6C-CRPs (Figure 5.7), which are not coffeetides and hence were not studied in the thesis.

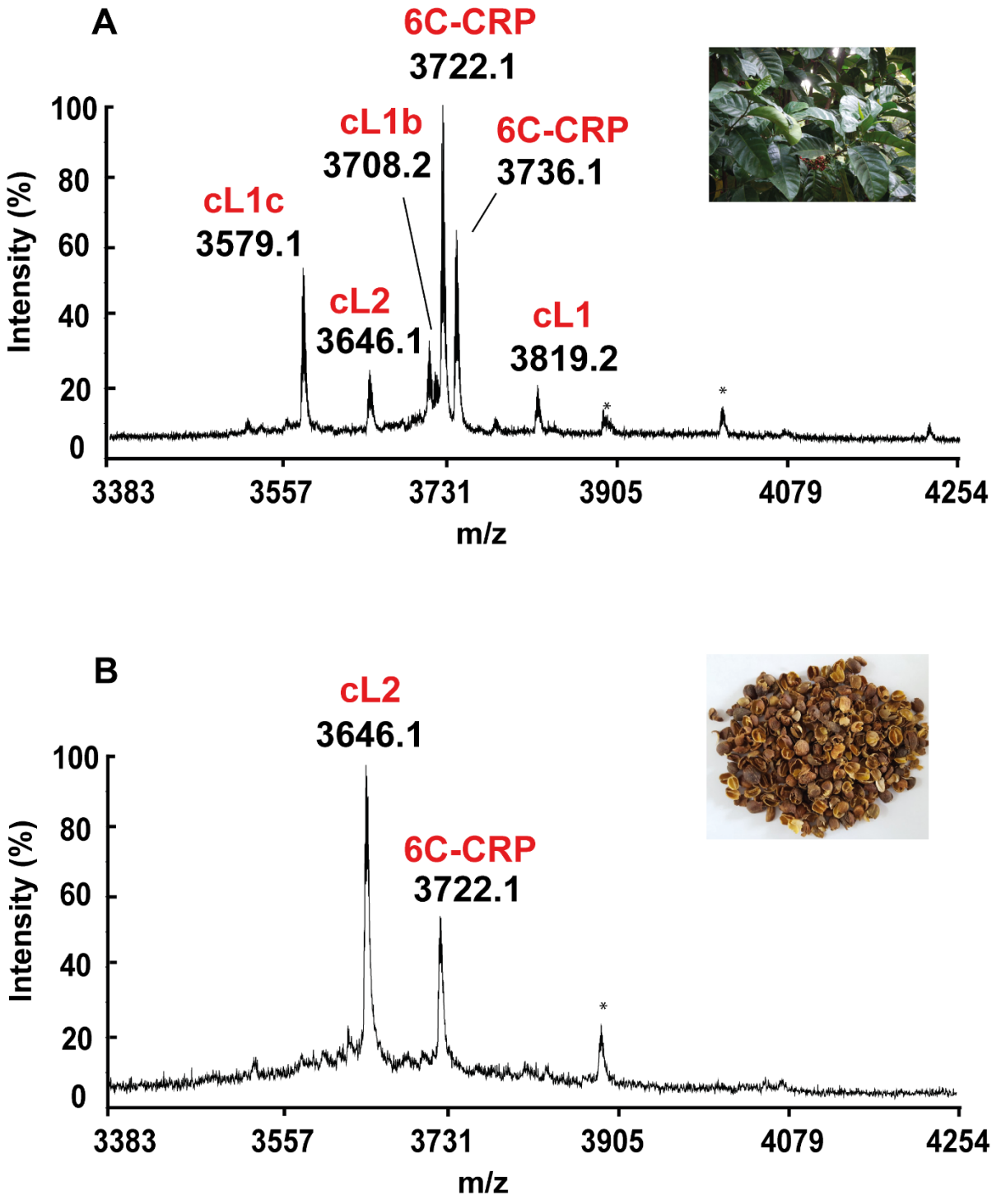



**Figure 5.5. Tissue-specific expression of putative coffeetides from *C. liberica*.** (A) MS profile of coffeetides from leaves. (B) MS profile of coffeetides from husks.

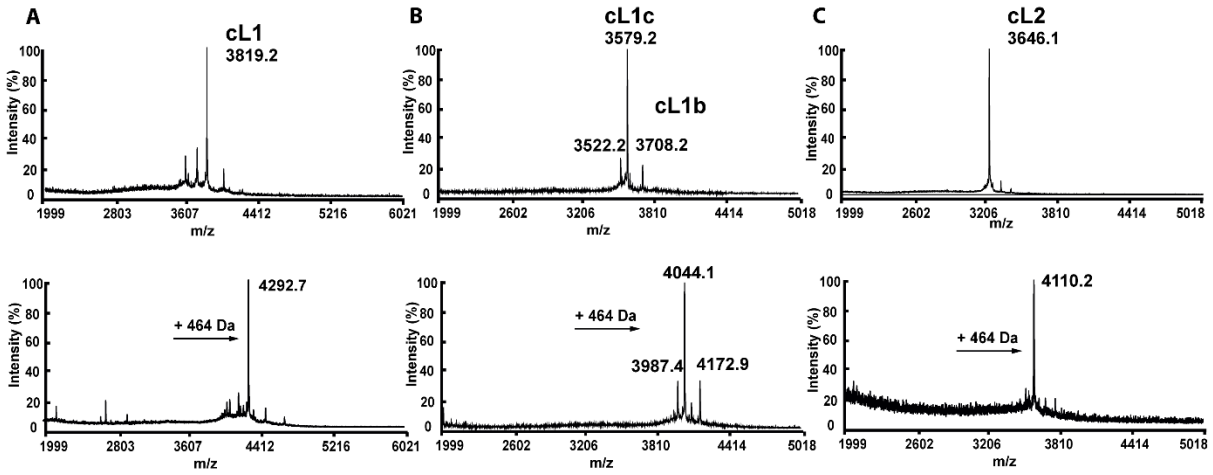



**Figure 5.6. MS profiles of (A) cL1, (B) cL1b, cL1c and (c) cL2 before and after *S*-reduction and *S*-alkylation.** A mass shift of 464 Da can be observed in all peptides after *S*-alkylation.

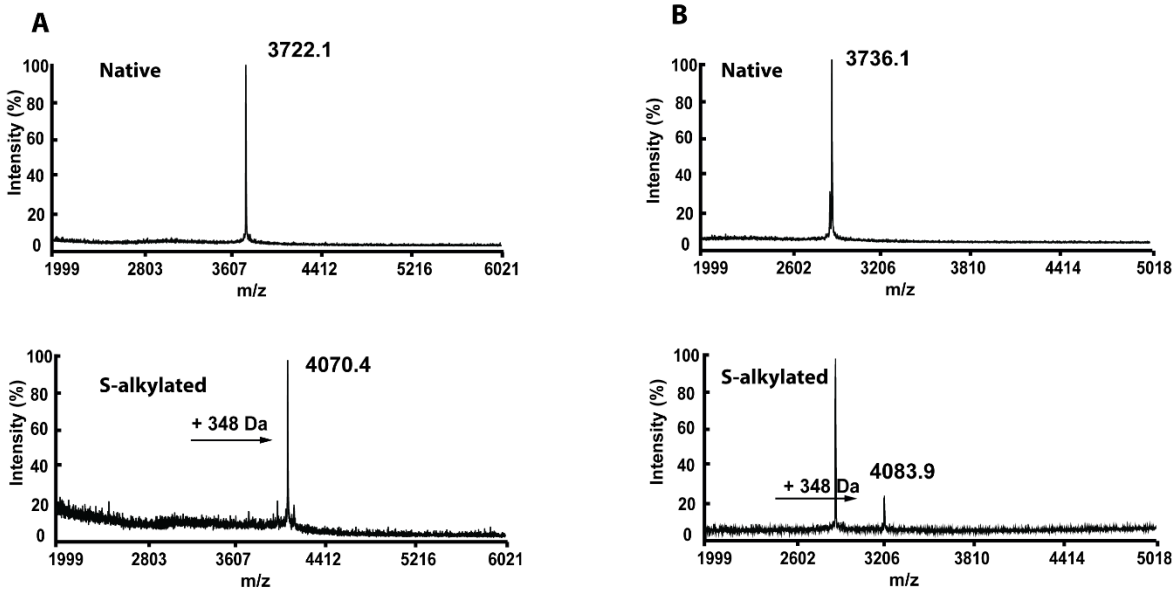



**Figure 5.7. MS profiles of two CRPs isolated from *C. liberica* leaves before and after *S*-reduction and *S*-alkylation.** A mass shift of 348 Da can be observed in all peptides after S-alkylation.

### 5.2.4. RNA extraction and transcriptomic database construction of *C. liberica*

Total RNA of the fresh leaves of *C. liberica* was extracted with Trizol, and the quality was determined by the value of OD260/280, OD260/230 and TAE-based agarose gel (Table 5.1, Figure 5.8). OD260, OD280, and OD230 represent the absorbance value of RNA, protein, and carbohydrate or residual phenol, respectively. The quality of RNA was assessed by OD260/280

and OD260/230. If the value of OD260/280 cannot reach 1.8, it means that the sample contains protein contaminants. For OD260/230, if it is lower than 1.8, it means that there might be carbohydrate carryover or residual phenol from nucleic acid extraction. The RNA was considered to be pure for both the value of OD260/280 and OD260/230 have exceeded 1.8. The RNA was sent to BGI (Shenzhen, China) for *de novo* assembly of the transcriptome.

**Table 5.1. RNA analysis of leaves from *C. liberica*.**

| Sample | Total Mass (µg) | OD260/280 | OD260/230 | 28S/18S |
|---|---|---|---|---|
| *C. liberica* | 16.8 | 1.89 | 1.86 | 1.3 |

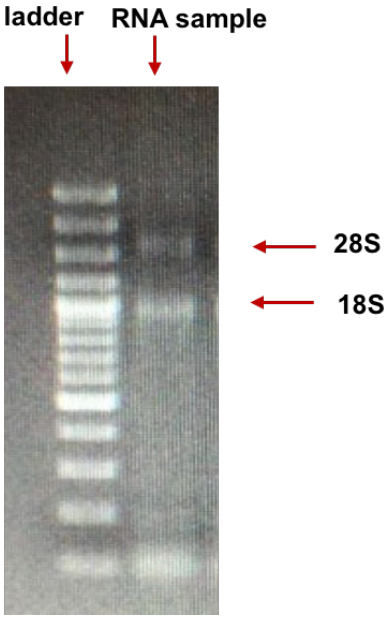



**Figure 5.8. TAE-based agarose gel running result.** Sharp 28S and 18S rRNA bands can be seen from the gel, and the 28S/18S result of 1.3 (Table 5.1) indicated that the RNA is intact.

### 5.2.5. Primary sequence determination of coffeetides from *C. liberica*

The primary sequence determination of coffeetide cL1, cL1b, cL1c, and cL2 followed the same manner as described in *C. canephora* species. Coffeetide cL1 sequence was determined as ZEGECSPAGKPCRPLGCCGACLCIVDHPTHEGTCVGNC (Figure 5.9A), while cL1b and cL1c were found to be the truncated version of cL1, with one or two deletions of amino acids at N-terminus (Figure 5.9B and 5.9C). Isobaric amino acids Leu/Ile or Lys/Gln was determined using transcriptomic data to obtain the full sequence. In addition, the full sequence of cL2 was confirmed as GECSPLGEPCAGNPWGCCPGCICIWQLTDRCIGNC (Figure 5.9D).

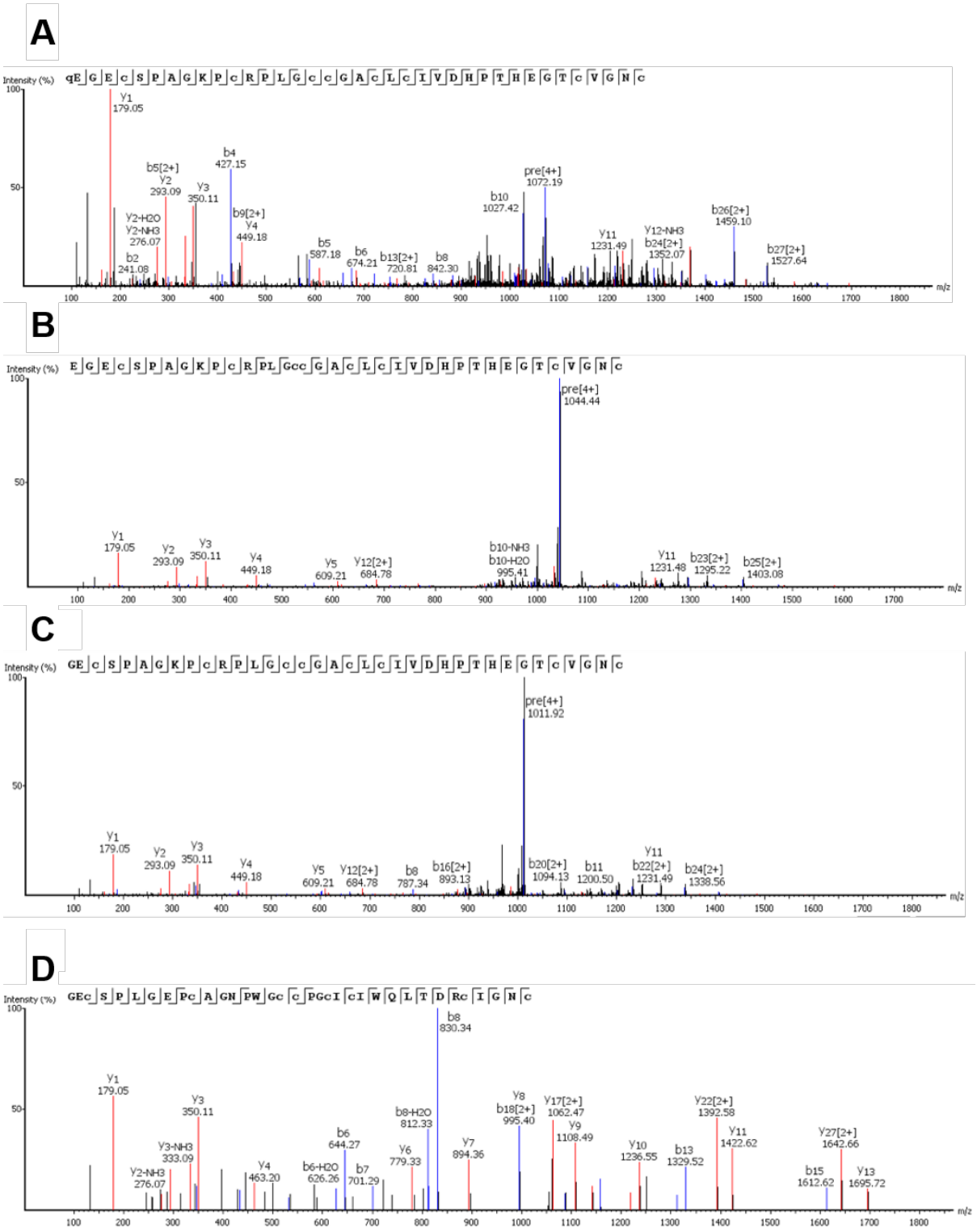



**Figure 5.9. Mass spectra of coffeetides from LC-ESI-LTQ-Orbitrap MS/MS in positive ion mode.** Mass spectra of (A) cL1, (B) cL1b, (C) cL1c and (D) cL2 were scanned between mass ranges from 0 to 2000 m/z. Transcriptomic data was employed to confirm the assignment of isobaric amino acids such as Leu/IIe.

### 5.2.6. Sequence comparison of coffeetides

To identify coffeetides in all *Coffea* species, tBLASTn was performed using coffeetide cC1 sequence as a query on EST database from GenBank. The sequences identified in EST means that they are being transcript and therefore are possible active genes. Our results revealed another 24 putative coffeetide-encoding gene sequences from the *Coffea* family of *C. arabica*, *C. canephora* and *C. racemose*, summarized in Table 5.2. Sequence alignment revealed that the coffeetide-encoding genes contained a conserved cysteine spacing of CXnCXnCCXnCXnCXnCXnC, which is characterized by an adjunct –CC- motif at the third and fourth positions. The sequence logo displayed in Figure 5.10 was obtained from the aligned sequences of the coffeetides and ginsentides. Sequence conservation was indicated by the overall height of a stack. Within each stack, the relative occurrence frequency of the amino acid was represented by the height of the symbol. Loop 5 containing highly conserved acidic residues of Asp32, Glu33, and Glu34, is the longest loop. Additionally, Pro16 and Gly39 are completely conserved and located in loop 2 and loop 6, respectively. Interestingly, all the coffeetide-encoding genes contain a very acidic N-terminus, in which Glu2 and Glu4 are highly conserved. Compared to ginsentides, the highly negatively charged coffeetides possess longer loop 5 and total length. In addition, the glycine-rich feature of ginsentides was absent in coffeetides.

**Table 5.2. Coffeetide sequences identified from *Coffea* species.**

| Peptide | Species | Amino acid sequence | Mass (Da)[1] | Charge[2] | PI | Approach[3] | Reference |
|---|---|---|---|---|---|---|---|
| Loop | | 1 2 3 4 5 6 | | | | | |
| cC1 | *C. canephora* | ZEGECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CVGNC | 3873.1 | -3 | 4.00 | T, P | This work |
| cC1b | *C. canephora* | -EGECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CVGNC | 3764.3 | -3 | 4.00 | T,P | This work |
| cC1c | *C. canephora* | --GECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CVGNC | 3633.4 | -2 | 4.14 | T,P | This work |
| cC2 | *C. canephora* | QEGECSPFGKPCRYNPWGCCDSCVCVAT-PADE-GRCLGNC | 4145.6 | -2 | 4.51 | T | [325] |
| cC3 | *C. canephora* | QEGECSPLGKPCKYNPWGCCGSCLCIVDQP-THEGTCVGNC | 4206.7 | -2 | 4.83 | T | [325] |
| cC4 | *C. canephora* | QEGECSPLGKPCRYNPWGCCGSCLCIVDQP-THEGTCVGNC | 4234.7 | -2 | 4.83 | T | [325] |
| cC5 | *C. canephora* | QEGECSALGKPCRYNPSGCCGLCVCVIPDPTDE-GSCIGIC | 4067.7 | -3 | 4.18 | T | [325] |
| cC6 | *C. canephora* | QEVECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CVGNC | 3931.6 | -3 | 4.00 | T | [325] |
| cC7 | *C. canephora* | QEPSCIPVLGSCVGNPWGCCPGCMCIRQ-LTDR---CHGYC | 3976.6 | 0 | 6.69 | T | [325] |
| cL1 | *C. liberica* | ZEGECSPAGKPCR--PLGCCGACLCIVDHP-THEGTCVGNC | 3819.2 | -2 | 5.36 | T,P | This work |
| cL1b | *C. liberica* | -EGECSPAGKPCR--PLGCCGACLCIVDHP-THEGTCVGNC | 3708.2 | -2 | 5.36 | T.P | This work |
| cL1c | *C. liberica* | --GECSPAGKPCR--PLGCCGACLCIVDHP-THEGTCVGNC | 3579.1 | -1 | 6.01 | T,P | This work |
| cL2 | *C. liberica* | --GECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CIGNC | 3649.2 | -2 | 4.14 | T,P | This work |
| cL3 | *C. liberica* | GKDTCIGLLESCKDDPWGCCFGCVCLWP--GDL---CRGSC | 3834.4 | -2 | 4.36 | T | This work |
| cL4 | *C. liberica* | QEGECSPAGKSCR--PVRCCDFCVCVVDYP-THVGTCRGNC | 4069.7 | -2 | 4.36 | T | This work |
| cL5 | *C. liberica* | QEGECSPAGKPCR--PVRCCDFCVCVVDYP-THVGTCRGNC | 4082.7 | 0 | 6.70 | T | This work |
| cL6 | *C. liberica* | QEPSCIPVLGSCVGNPWGCCPGCMCIRQ-LTDR---CHGYC | 3976.7 | 0 | 6.69 | T | This work |
| cA1 | *C. arabica* | QEGECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CIGNC | 3903.5 | -3 | 4.00 | T | [326] |
| cA2 | *C. arabica* | QEGECSPLGEACAGNPWGCCPGCICIWQ-LTDR---CVGNC | 3863.5 | -3 | 4.00 | T | [326] |
| cA3 | *C. arabica* | QEPSCLPAGESCTGNPWGCCPGCICIWQ-LTER---CVGNC | 3905.6 | -2 | 4.25 | T | [327] |
| cA4 | *C. arabica* | QEPSCIPVGEPCAGNPGGCCDGCICIWQ-LTDR---CAGSC | 3733.5 | -3 | 3.92 | T | [326] |
| cA5 | *C. arabica* | QEGECSPLGKPCRYNPRGCCDFCVCVVADVTDEEGSCRGNC | 4389.7 | -3 | 4.44 | T | [326] |
| cA6 | *C. arabica* | RKDTCIGLLESCKDDPYGCCPGCVCLWP--GDL---CRGDC | 3884.6 | -2 | 4.46 | T | [328] |
| cA7 | *C. arabica* | QEGECSPAGKPCR--PVRCCDSCLCIVDYP-THVGTCRGNC | 4047.7 | 0 | 6.70 | T | [328] |
| cA8 | *C. arabica* | QEGECSPLGKPCAGNPWGCCPGCICIWQ-LTDR---CIGNC | 3902.5 | -1 | 4.68 | T | [327] |
| cA9 | *C. arabica* | QEPSCIPVGEPCAGNPGGCCDGCICIWQ-LTDR---CAGSC | 3733.3 | -3 | 3.92 | T | [326] |
| cA10 | *C. arabica* | QEGECSPLGEPCAGNPWGCCPGCICIWQ-LTDR---CVGNC | 3890.1 | -3 | 4.00 | T | [327] |
| cR1 | *C. racemosa* | QEPRCIPALGSCVGNPWGCCFGCMCIRQ-LTDR---CLGYC | 4043.7 | +1 | 7.70 | T | [328] |
| cR2 | *C. racemosa* | QEPRCIPVFGSCVGNPWGCCFGCMCIRQ-HTNR---CLGYC | 4128.7 | +2 | 8.22 | T | [328] |
| cR3 | *C. racemosa* | QEGECSPFGKPCRYNPWGCCGSCLCVVDHP-THEGTCVGNC | 4263.7 | -2 | 5.36 | T | [328] |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| cR4 | *C. racemosa* | QEGECSPFGKPCRYNPWGCCGSCVCVVDHP-THEGTCLGNC | 4263.7 | -2 | 5.36 | T | [328] |
| cR5 | *C. racemosa* | QEGKCSPAGKPC--DPWGCCDFCVCVVDFPGGE-GRCAGNC | 3885.5 | -2 | 4.44 | T | [328] |
| cR6 | *C. racemosa* | QEEKCSPAGKPCRYNPRGCCDFCVCVVDFPGGE-GSCLGNC | 4218.7 | -1 | 4.94 | T | [328] |
| cR7 | *C. racemosa* | GKDTCIGLLESCKDDPWGCCPGCVCLWP--GDL---CRGSC | 3780.5 | -2 | 4.36 | T | [328] |

[1]Mass (Da) = reported mass. [2]Charge: the total charge is the sum of basic (Lys, Arg, and His residues) and acidic (Glu and Asp residues) residues present in the sequence. [3]Approach: The approach used to obtain the primary sequences by transcriptomic (T) and/or proteomic (P) analysis. The cysteine residues were highlighted in yellow.

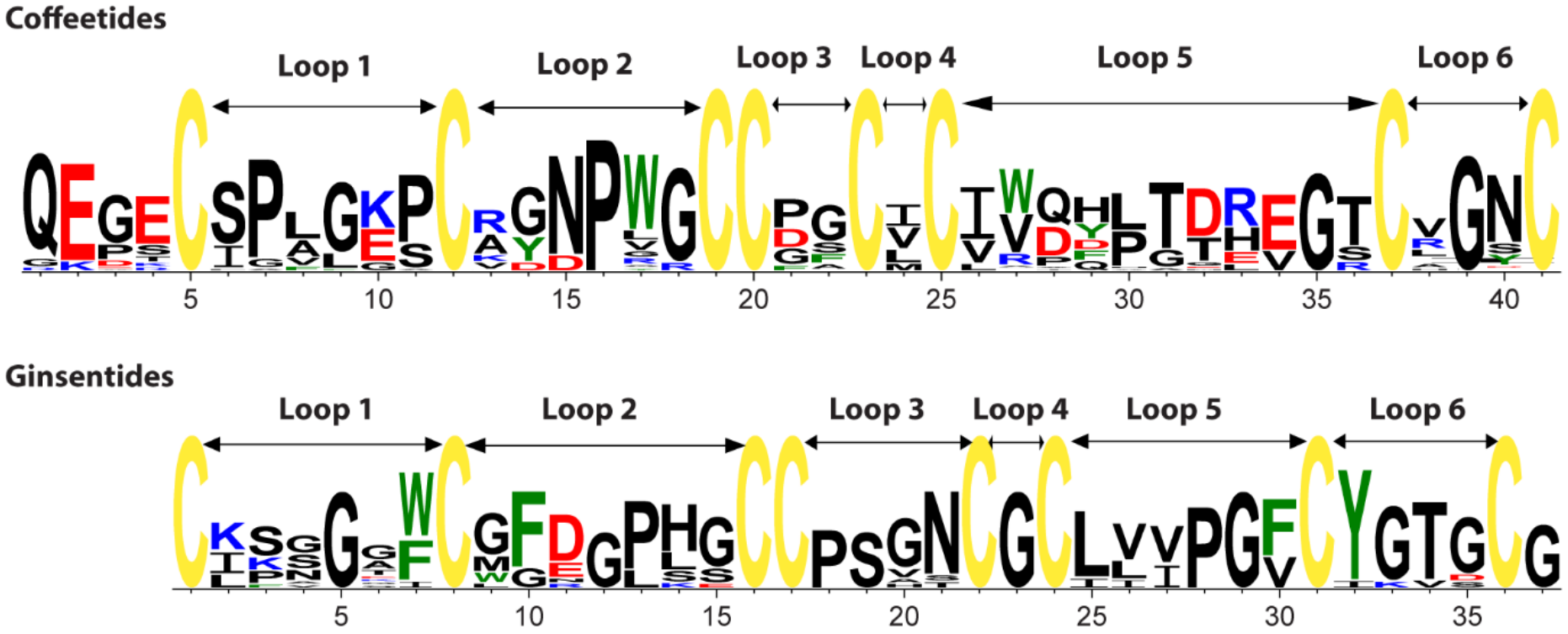



**Figure 5.10. Sequence logo of coffeetides and ginsentides.** The overall height of the stacks represents the conservation while the occurrence of each amino acids in positions is revealed by the height of the symbol within the stack. Yellow, red, blue and green color were used to indicate cysteine, acidic (D and E), basic (H, R, and K) and aromatic residues (F, W, and Y), respectively.

### 5.2.7. Biosynthesis pathway of coffeetides

The cDNA library obtained from RNA of *C. liberica* leaves was subjected to 5′-RACE PCR to acquire the full-length gene of cL1, cL1b, cL1c, and cL2. Primers designed to recognize 3′- and 5′-UTRs of this gene amplified the DNA sequences of cL1, cL1b, cL1c, and cL2. The full precursor sequences of coffeetides from *C. canephora*, *C. arabica,* and *C. Rosmosa* were acquired from GenBank database.

The precursor sequences of coffeetides, ginsentides and other coffeetide-like peptides from important crops including cacao (*Theobroma cacao*), cotton (*Gossypium raimondii*), rice (*Oryza sativa*) and wheat (*Triticum aestivum*) were compared in Figure 5.11 A. All coffeetides share the same three-domain precursor sequence arrangement as ginsentides, which includes an endoplasmic reticulum (ER) signal peptide which contains 22-24 aa, a pro-peptide (35-39 aa) and a mature peptide domain (35-41 aa). Sequence comparison revealed that between the signal peptide and pro-peptide in coffeetides, the cleavage site is highly conserved, which are between Gly and Gln or Lys. Figure 5.11B summarized all the precursor architecture of coffeetides and other known 8C-CRPs, which include 8C-defensins, chitin-binding 8C-HLPs, ginsentides, and 8C-thionins. 8C-defensins, chitin-binding 8C-HLPs, and 8C-thionins contain a two to three domain architecture containing a signal peptide followed by a mature peptide and an additional C-terminal tail. With a pro-peptide in between the signal peptide and mature peptide, it is obvious that coffeetides do not share the same precursor arrangement as 8C-defensins, chitin-binding 8C-HLPs, and 8C-thionins. Belonging to non-chitin-binding 8C-HLPs, coffeetides and ginsentides share the same precursor arrangement. The length of the mature domain in coffeetides and ginsentides is shorter than the other 8C-CRPs. However, coffeetides contain a shorter pro-peptide (mean: 38.7 aa) than those of ginsentides (mean: 61 aa).

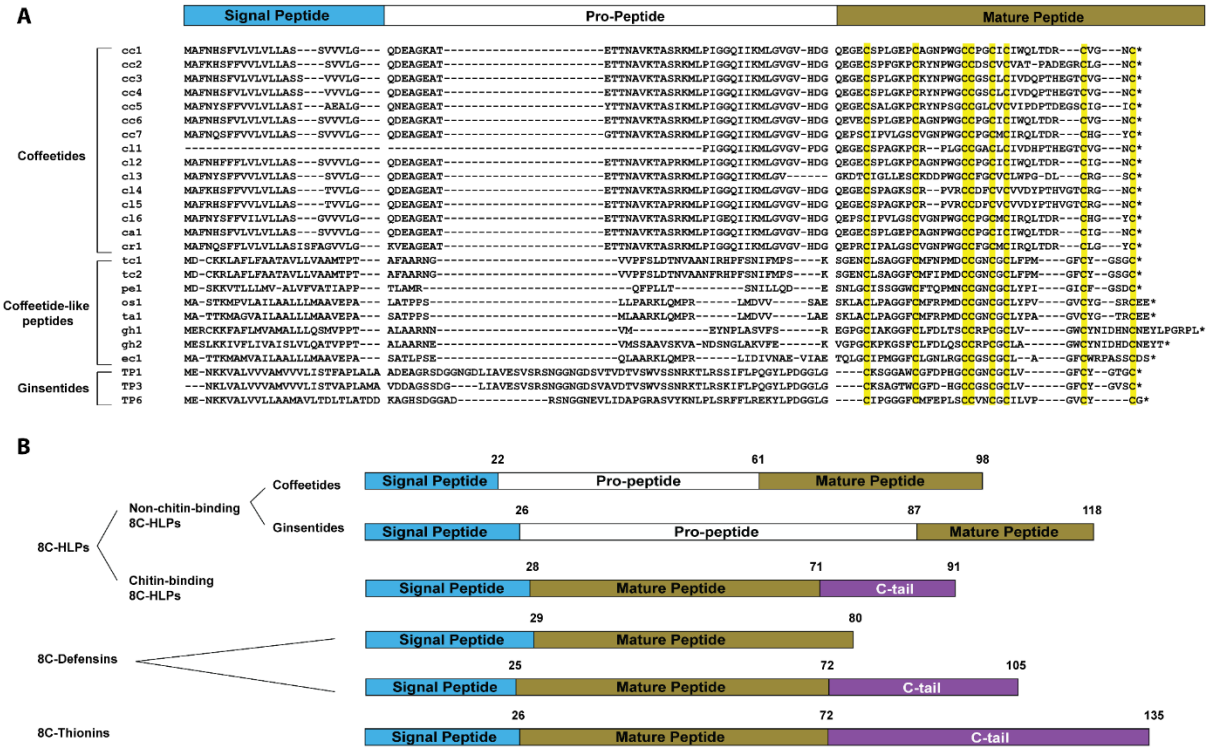



**Figure 5.11. Gene alignment of coffeetides, coffeetide-like peptides, and ginsentides.** The precursor sequences of coffeetides, coffeetide-like peptides, and ginsentides were aligned using MUSCLE. The precursors consist of a three-domain architecture, which includes a signal domain, a pro-domain, and a mature domain. ∗Represents the stop codon. (B) A comparison of precursor arrangement of coffeetides and other reported 8C-CRPs. The number on the top of each domain represents the average number of amino acids. GenBank accession number: cc1 (DV676066.1), cc2 (DV678117.1), cc3 (DV688598.1), cc4 (DV704915.1), cc5 (DV687022.1), cc6 (DV678112.1), cc7 (DV672477.1), cc8 (DV674842.1), ca1 (GT701156.1), ca2 (GT020922.1), ca3 (GR983294.1), ca4 (GT673445.1), ca5 (GT021473.1), ca6 (GT021132.1), ca7 (GW468514.1), ca8 (GW486522.1), ca9 (GT692623.1), ca10 (GT021252.1), cr1 (GT669255.1), cr2 (GT665740.1), cr3 (GT664847.1), cr4 (GT665189.1), cr5 (GT666921.1), cr6 (GT666030.1), cr7 (GT668698.1), tc2 (CU471503.1), pd1(CA821362.1), pe1 (AJ780051.1), ptt1 (BU827525.1), pt1 (CA925073.1), as1 (DY543302.1), gh1 (CO491697.1), osj1 (CI251708.1) and ta1(CK209254.1). OneKP accession number: hg1 (OKEF-2088352), ms1 (zAJFN-2096758) and eco1 (CXSJ-2102568).

### 5.2.8. Disulfide connectivity of coffeetide cC1

The disulfide linkage of coffeetides was determined by a stepwise partial reduction and alkylation. Coffeetide cC1 was first partially *S*-reduced with tris(2-carboxyethyl)phosphine

(TCEP) and *S*-alkylated with N-Ethylmaleimide (NEM) to obtain 1-SS, 2-SS, and 3-SS species (Figure 5.12A). Subsequently, these intermediates were subjected to RP-HPLC for purification. The NEM-alkylation causes a 126.15 Da increment for each cysteine residue and thus the intermediates which showed molecular weights of 4125.05, 4380.13 and 4629.68 Da contain one, two, and three reduced disulfide bonds, respectively. Fully reduction and alkylation with DTT and IAA were followed, and each IAA-alkylated cysteine will give rise to a mass shift of 57 Da. MALDI-TOF MS/MS was employed to determine the positions of NEM- and IAA-alkylated cysteine residues and to deduce the disulfide connectivity of cC1. Figure 5.12B–H summarized the procedure of elucidating the disulfide linkages. The 3S-S revealed the CysI–IV disulfide bind while the 2S-S showed the linkage of CysIII–VII and the 1S-S intermediate revealed CysII–VI. The fourth disulfide bond, from CysV–VIII, was obtained by deduction.

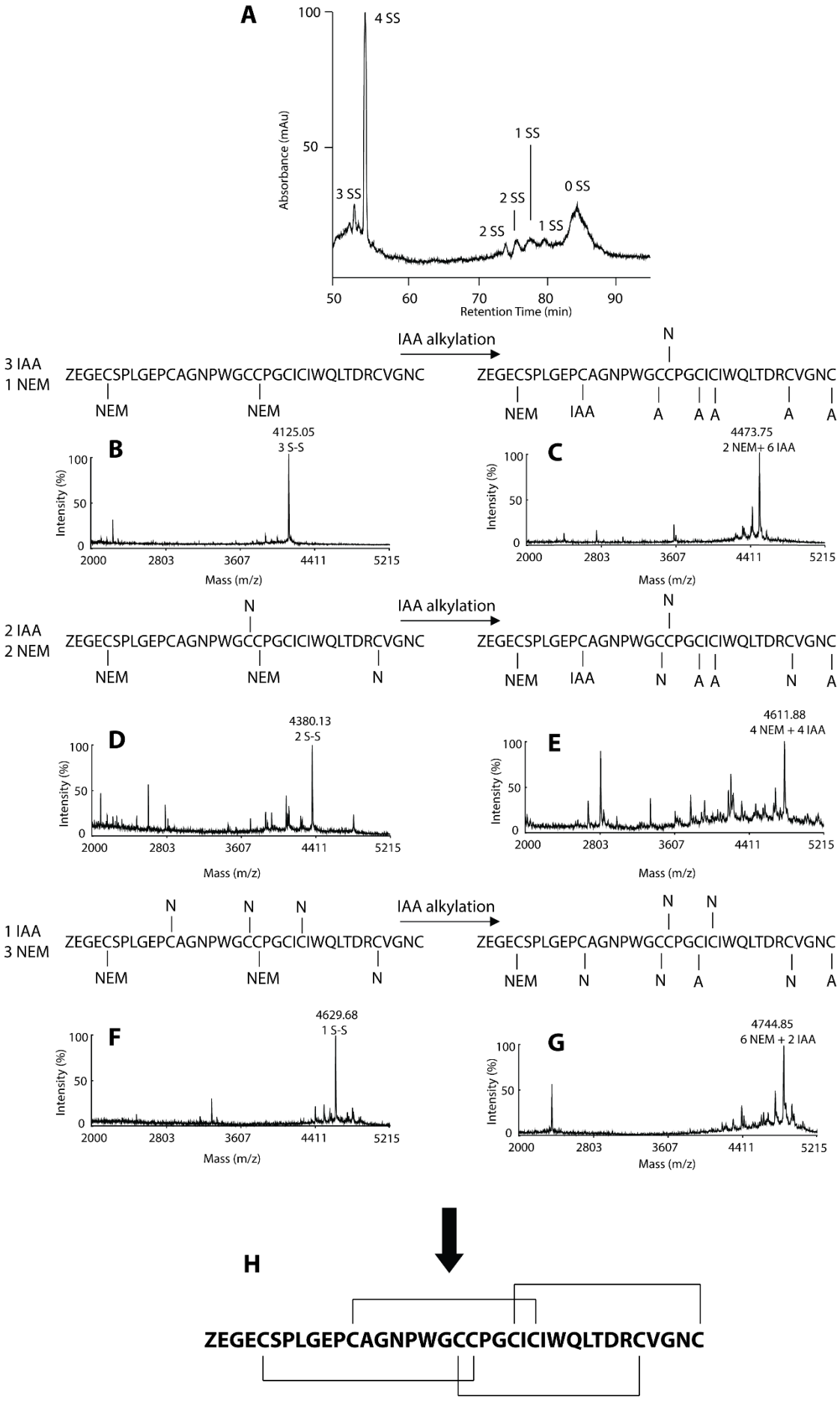

A
100
50
Absorbance (mAu)
4 SS
3 SS
2 SS
2 SS
1 SS
1 SS
0 SS
50
60
70
80
90
Retention Time (min)
IAA alkylation
3 IAA
1 NEM
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC
NEM
NEM
N
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC
NEM
IAA
A
A A
A
A
B
4125.05
3 S-S
Intensity (%)
100
50
0
2000
2803
3607
4411
5215
Mass (m/z)
C
4473.75
2 NEM+ 6 IAA
Intensity (%)
Mass (m/z)
IAA alkylation
2 IAA
2 NEM
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC
NEM
NEM
N
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC
NEM
IAA
N
A A
N
A
D
4380.13
2 S-S
Intensity (%)
Mass (m/z)
E
4611.88
4 NEM + 4 IAA
Intensity (%)
Mass (m/z)
IAA alkylation
1 IAA
3 NEM
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC
NEM
NEM
N
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC
NEM
N
N
A
N
A
F
4629.68
1 S-S
Intensity (%)
Mass (m/z)
G
4744.85
6 NEM + 2 IAA
Intensity (%)
Mass (m/z)
H
ZEGECSPLGEPCAGNPWGCCPGCICIWQLTDRCVGNC

**Figure 5.12. Disulfide connectivity illustration of coffeetide cC1.** (A) HPLC profile of partially alkylated cC1. The partially reduced mixture was subjected to HPLC for a separation, and the major peaks were collected. (B) MS spectrum of intermediate with two NEM-alkylated cysteines and (C) samples later fully reduced and alkylated with DTT and IAA. (D) MS spectrum of intermediate with four NEM-alkylated cysteines and (E) samples later fully reduced with DTT and *S*-alkylated IAA. (F) MS spectrum of intermediate with six NEM-alkylated cysteines and (G) samples later fully reduced and alkylated with DTT and IAA. The fully alkylated cC1 was desalted and sequenced by MALDI-TOF MS/MS. (H) Summary of the disulfide connectivity of coffeetide cC1.

### 5.2.9. Chemical synthesis and oxidative folding of cC1

The chemical synthesis of coffeetide cC1 was accomplished by the Fmoc solid-phase peptide synthesis (SPPS). TIS/EDT/$H_2O$/TFA (2.5:2.5:2.5:92.5) were used to cleave the synthesized peptide from the resin. MALDI-TOF MS and RP-HPLC were employed to access the mass and purity of the cleaved peptide. Once being cleaved from the resin, cC1 was in the reduced form with an m/z value of 3881 Da.

18 conditions were used to perform the oxidative folding of cC1 using different redox reagents, duration, and multiple concentrations of co-solvent. Ammonium bicarbonate buffer (pH 8.0) was used as a general buffer for all the conditions [101] (Table 5.3). Results showed that the use of a pair of redox with lower price, consisting of cysteamine and cystamine (10:100 mM), generated a higher folding yield of 18.03% (Run 1) compared to the standard redox pair GSSG and GSH (10:100 mM), which gave a yield of 9.25% (Run 2). The effect of adding DMSO to the folding reaction was compared in run 3-5. The folding yield increased to 59.12% (Run 3), 84.76% (Run 4) with adding 10 and 20% (v/v) of DMSO, respectively. However, when continued to increase the DMSO concentration to 30%, no significant change in folding yield was observed (Run 5). Run 6 and 7 tested the effect of adding isopropanol, an organic solvent that may enhance the conformational stability. However, results revealed that by adding 20 and 30% isopropanol, folding yields decreased to 29.41% (Run 6) and 11.41% (Run 7), respectively. It suggested that the addition of organic solvent will hinder the folding process. The influences of different folding times and different ratios of redox reagents were evaluated in run 8-18. The increase in the concentration of cystamine to 20 mM decreased the folding yield to 77.12% (Run 9). On the contrary, when 100 mM of cysteamine was added to the folding reaction, the increasing concentration of cysteamine to 200 mM has led to a higher folding yield of 81.69% (Run 12). However, a higher concentration of cysteamine (300 and

400 mM) did not increase the folding yield (Run 15 and 18). For incubation time (1, 3 and 24 h), it did not show significant improvement in folding yields after prolonging the folding reaction beyond 3 h. Figure 5.13 summarized the folding process of five selected folding conditions. Together, the optimal condition for oxidative folding of cC1 is determined as follows: 0.38 mg of synthetic, reduced cC1 was dissolved in 100 µL of 0.1 M of ammonium bicarbonate buffer (pH 8.0) containing 10 mM cystamine, 200 mM cysteamine and 20% (v/v) DMSO and incubated for 3 h. The optimal folding yield was 82.48%. The synthetic cC1 was then compared with native cC1 extracted from *C. canephora.* The UPLC profiles (Figure 5.14) indicated that native cC1 and synthetic cC1 co-eluted at the same time, suggesting that they are in a similar structural fold. Additionally, a structural characterization based on 2D NMR was performed to confirm the correcting folding of synthetic cC1. (Figure 5.15)

**Table 5.3. Parameters of the oxidative folding condition.**

| Run | GSSG (mM) | GSH (mM) | Cystamine (mM) | Cysteamine (mM) | DMSO (%) | iPrOH (%) | Time (h) | Yield (%) |
|---|---|---|---|---|---|---|---|---|
| **1** | | | 10 | 100 | 0 | 0 | 24 | 18.03 |
| **2** | 10 | 100 | | | 0 | 0 | 24 | 9.25 |
| **3** | | | 10 | 100 | 10 | 0 | 24 | 59.12 |
| **4** | | | 10 | 100 | 20 | 0 | 24 | 81.17 |
| **5** | | | 10 | 100 | 30 | 0 | 24 | 80.86 |
| **6** | | | 10 | 100 | 20 | 20 | 24 | 29.41 |
| **7** | | | 10 | 100 | 20 | 30 | 24 | 11.41 |
| **8** | | | 20 | 100 | 20 | 0 | 1 | 67.37 |
| **9** | | | 20 | 100 | 20 | 0 | 24 | 77.12 |
| **10** | | | 10 | 200 | 20 | 0 | 1 | 74.73 |
| **11** | | | 10 | 200 | 20 | 0 | 3 | 82.48 |
| **12** | | | 10 | 200 | 20 | 0 | 24 | 81.69 |
| **13** | | | 10 | 300 | 20 | 0 | 1 | 69.63 |
| **14** | | | 10 | 300 | 20 | 0 | 3 | 76.13 |
| **15** | | | 10 | 300 | 20 | 0 | 24 | 77.54 |
| **16** | | | 10 | 400 | 20 | 0 | 1 | 45.85 |
| **17** | | | 10 | 400 | 20 | 0 | 3 | 43.44 |
| **18** | | | 10 | 400 | 20 | 0 | 24 | 41.97 |

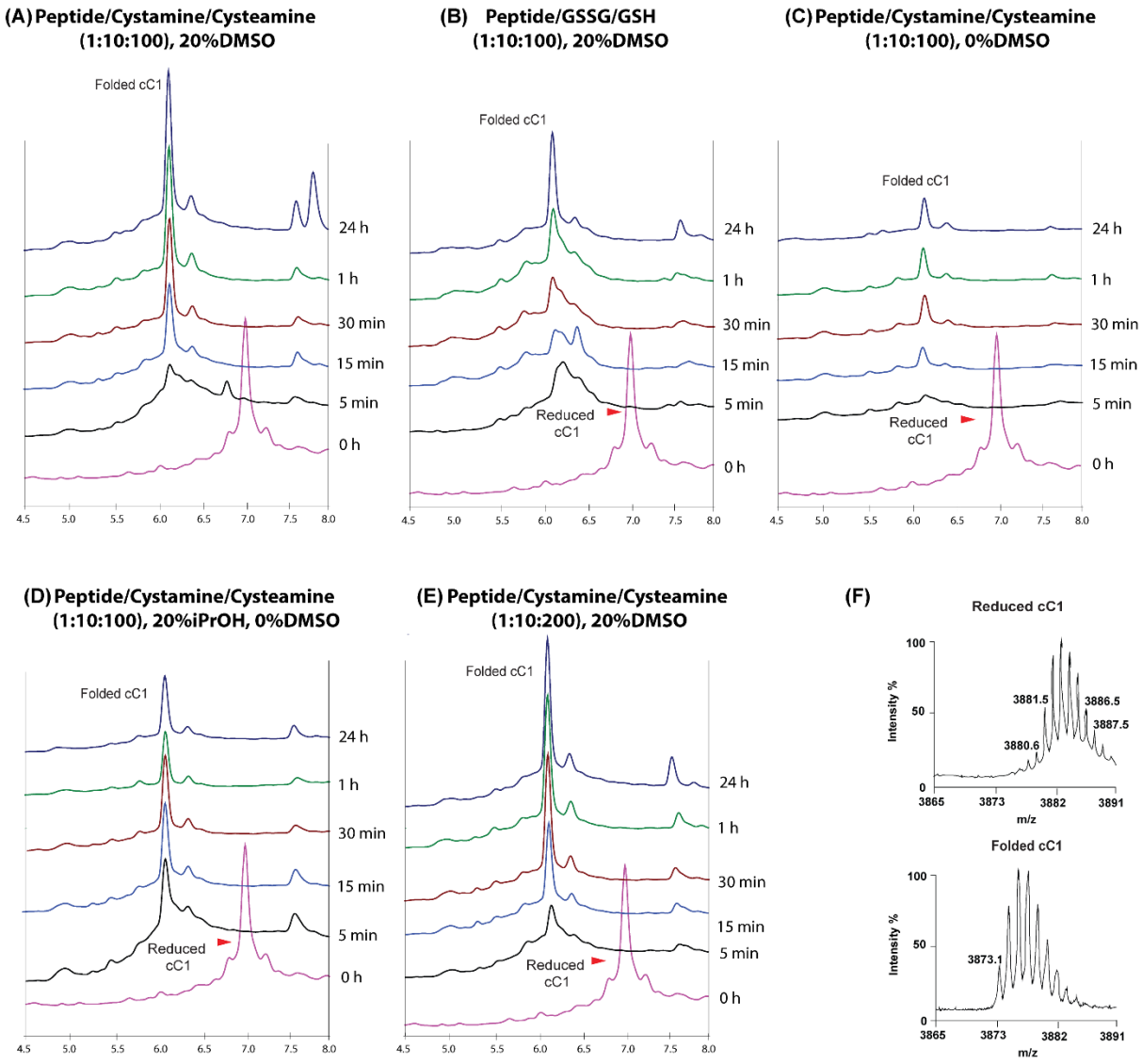



**Figure 5.13. Selected different oxidative folding conditions.** The synthetic cC1 was folded with the general folding condition (0.1 M $NH_4HCO_3$, pH 8). (A) Folding condition: 1 mM peptide+ 10 mM Cystamine/100 mM Cysteamine, with 20% DMSO. (B) Folding condition: 1 mM peptide+ 10 mM GSSG/100mM GSH, with 20% DMSO. (C) 1 mM peptide+ 10 mM Cystamine/100 mM Cysteamine. (D) 1 mM peptide+ 10 mM Cystamine/100 mM Cysteamine, with 20% iPrOH. (E) 1 mM peptide+ 10 mM Cystamine/200 mM Cysteamine, with 20% DMSO. (F) The MS profile of the peaks shown in the HPLC results, indicating reduced cC1 3881 Da, folded cC1 3873 Da.

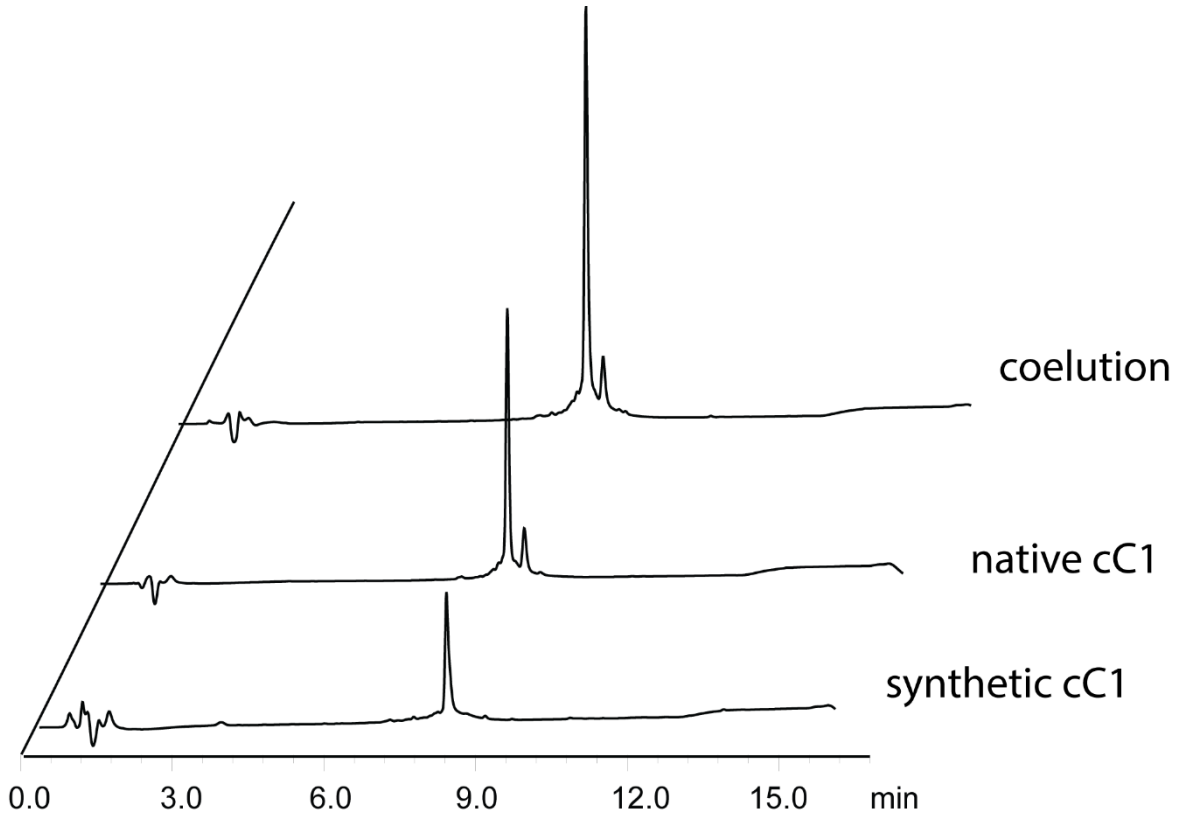



**Figure 5.14. HPLC profile comparison of native cC1 and synthetic cC1.** Folded synthetic cC1, native cC1 and the mixture of them were subjected to HPLC.

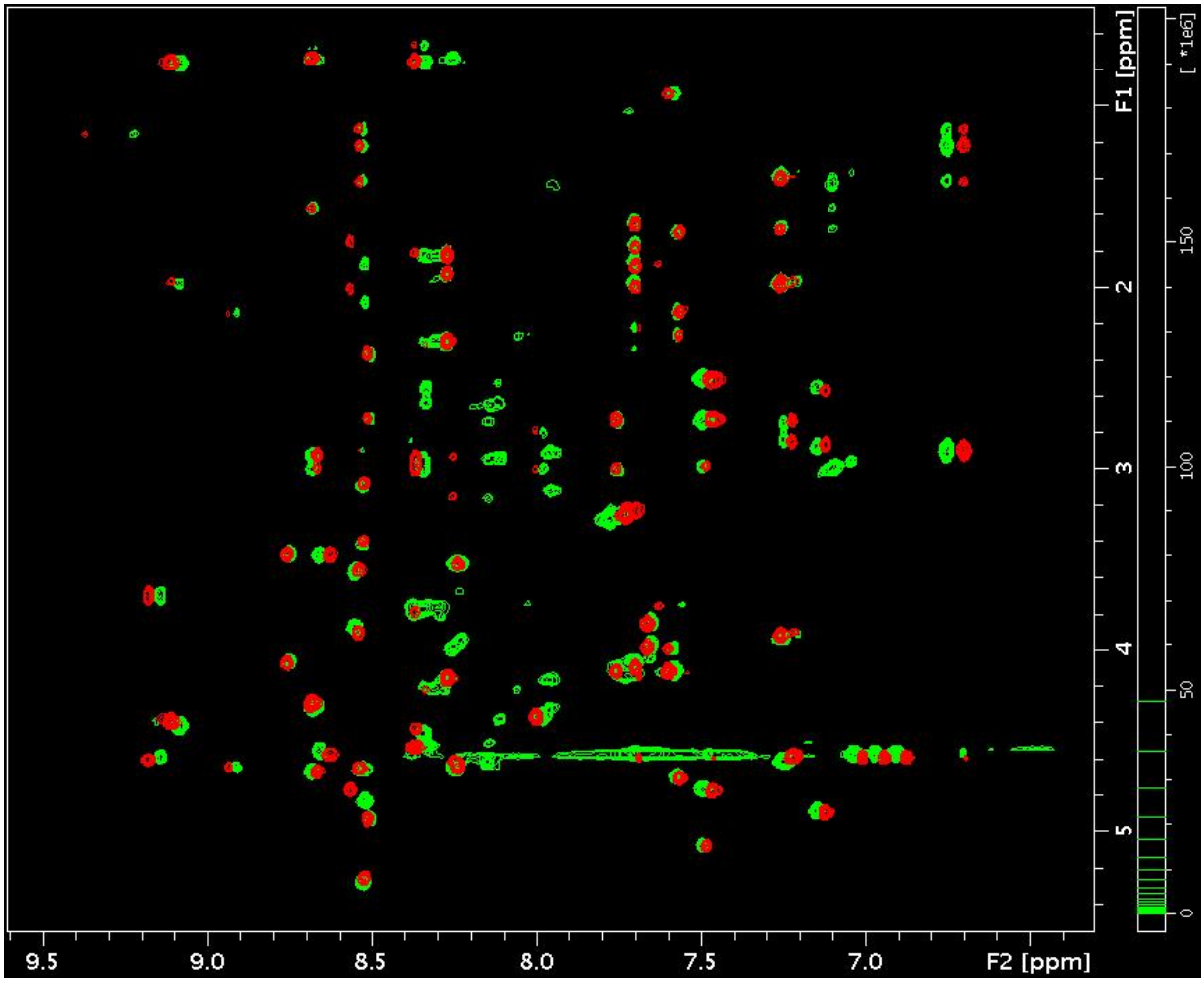



**Figure 5.15. Overlapped 2D NOESY spectra of native (red) and synthetic cC1 (green) displayed by Sparky 3.115.**

### 5.2.10. Solution Structure of cC1

The NMR solution structure of cC1 (Figure 5.16) was determined using the distance restraints obtained from 2D 1H-1H-TOCSY and NOESY, as well as the hydrogen bond restraints based on the H/D exchange NMR experiment (Table 5.4). All spin-spin systems of cC1 were identified except the first residue pyroglutamate, and ~98% of proton resonances were unambiguously assigned. The solution structure of cC1 was determined based on a total of 216 NMR derived distance restraints and four hydrogen bonds. Figure 5.16A shows the NMR ensemble of the 10 lowest-energy cC1 structures. The root-mean-square deviation (RMSD) value of the 10 best structures for residues Gly3-Glu10 and Gly18-Cys37 was 1.10±0.21 Å and that for all heavy atoms was 1.60 ±0.25 Å (Table 5.5). The cC1 structure consisted of two anti-parallel small β-strands (β1: Cys25-Trp27 and β2: Cys33-Gly35) and several tight turns and loops. The NMR structure of coffeetide cC1 revealed a similar structural fold and disulfide

connectivity as ginsentides. The three disulfide bonds Cys I-IV, II-VI and III-VII in cC1 has formed a cystine-knot fold whereas the additional penetrating disulfide bond Cys V-Cys VIII links the C-terminus to the β1 sheet. Its molecular shape was well-defined by several medium and long-range NOEs (Figure 5.16B). The 3D structure of coffeetide cC1 was deposited on Protein Data Bank with an accession number of 6JI7. Figure 5.16C showed the topology comparison of the electrostatic surface between cC1 and ginsentide TP1 (PDB: 2ml7) in two views. Compared to TP1, coffeetide cC1 is highly charged with four acidic residues (Glu2, Glu4, Glu10, and Asp31) and one basic residue (Arg32).

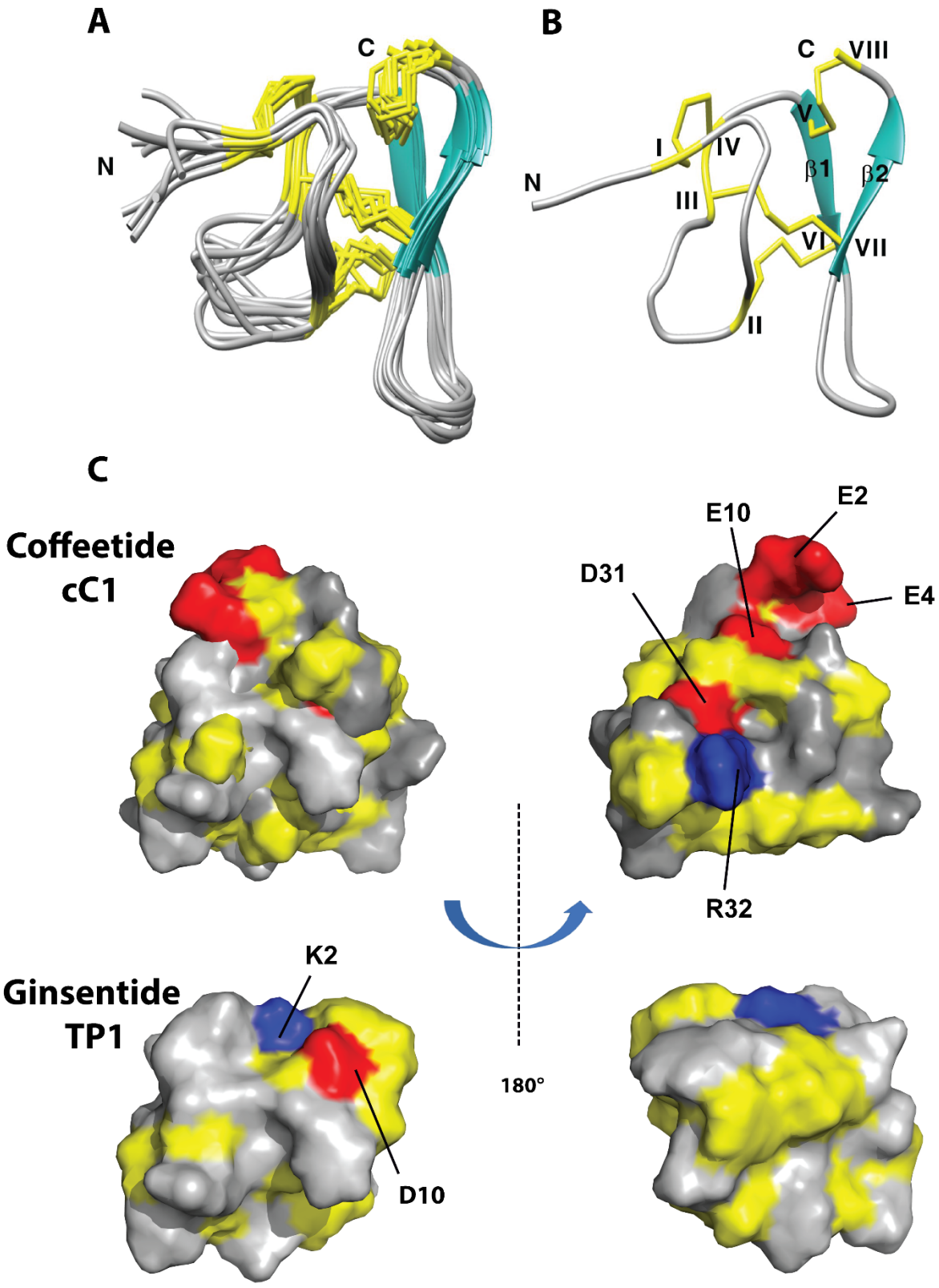



**Figure 5.16. The solution NMR structure of cC1.** (A) Superposition of the cC1 backbone traces from the final 10 ensembles solution structures and restrained energy minimized structure. (B) Ribbon representation of cC1 structure with disulfide bonds formed between CysI-CysIV, CysII-CysVI, CysIII-CysVII, and CysV-CysVIII. (C) Electrostatic surface

comparison of cC1 (PDB: 6JI7) and ginsentide TP1 (PDB: 2ml7) in two views. The acidic residues (D, E), basic residues (K, R) and hydrophobic residues are highlighted in red, blue, and yellow, respectively.

**Table 5.4. Structural statistics for the final 10 conformers of cC1[a]**

| | |
|---|---|
| Distance restraints | |
| Intra-residue ($i$-$j$ = 0) | 89 |
| Sequential ($\|i$-$j\|$ = 1) | 80 |
| Medium range ($2 \leq \|i$-$j\| \leq 4$) | 19 |
| Long range ($\|i$-$j\| \geq 5$) | 28 |
| Hydrogen bond | 4 |
| Total | 220 |
| Average rmsd to the mean structure (Å)[b] | |
| Backbone atoms | 1.10 ± 0.21 |
| Heavy atoms | 1.60 ± 0.25 |
| ϕ/ψ space[c] | |
| Most favored region (%) | 56.2 |
| Additionally allowed region (%) | 32.5 |
| Generously allowed region (%) | 6.7 |
| Disallowed region (%) | 4.6 |
| rmsd from covalent geometry | |
| Bonds (Å) | 0.006 ± 0.035 |
| Angles (deg.) | 0.210 ± 0.056 |
| Impropers (deg.) | 0.030 ± 0.019 |
| rmsd from experimental restraints | |
| NOEs (Å) | 0.0418 ± 0.0089 |

[a] Selected from 100 calculated conformers according to overall energy.
[b] Calculated with MOLMOL using range 3-13, 18-37.
[c] Calculated with PROCHECK-NMR.

**Table 5.5. Proton chemical shift assignments for each amino acid residues of coffeetide cC1.**

| | HN (ppm) | Hα (ppm) | Hβ (ppm) | | Others (ppm) |
|---|---|---|---|---|---|
| Z1 | | | | | |
| E2 | 8.382 | 4.297 | 2.074 | 1.970 | Hγ, 2.441 |
| G3 | 7.851 | 3.400 | | | |
| E4 | 7.383 | 4.082 | 2.125 | 1.544 | Hγ, 1.825 |
| C5 | 8.102 | 4.521 | 3.144 | 2.942 | |
| S6 | 9.301 | 4.753 | 3.870 | 3.827 | |
| P7 | | 4.351 | 2.214 | 1.963 | |
| L8 | 7.460 | 3.502 | 1.635 | 1.576 | Hγ, 1.417; Hδ, 0.960 |
| G9 | 8.900 | 4.226, 3.620 | | | |
| E10 | 7.708 | 4.845 | 2.278 | 1.847 | Hγ, 2.410 |
| P11 | | 4.580 | | | |

| | | | | | |
|---|---|---|---|---|---|
| C12 | 8.386 | 4.749 | 3.309 | 3.083 | |
| A13 | 9.478 | 3.997 | 1.301 | | |
| G14 | 8.663 | 4.043, 3.709 | | | |
| N15 | 7.266 | 4.134 | 3.007 | 2.716 | |
| P16 | | 4.134 | 1.858 | 1.619 | Hγ, 2.056, 1.236 |
| W17 | 7.888 | 4.281 | 3.149 | 2.881 | Hδ1,7.269, Hε1, 10.096 |
| G18 | 7.800 | 4.131, 4.003 | | | |
| C19 | 8.625 | 5.085 | 2.873 | 2.515 | |
| C20 | 9.071 | 4.790 | 3.387 | 2.296 | |
| P21 | | 4.346 | 2.345 | | |
| G22 | 8.748 | 4.731, 3.619 | | | |
| C23 | 8.666 | 5.398 | 3.548 | 3.228 | |
| I24 | 8.820 | 4.446 | 1.712 | | Hγ: 0.890, 1.400, 1.031; Hδ, 0.813 |
| C25 | 8.781 | 4.820 | 3.143 | 3.075 | |
| I26 | 8.504 | 4.671 | 1.960 | | Hγ: 0.902, 1.440, 1.265; Hδ, 0.813 |
| W27 | 8.473 | 4.588 | 3.147 | 3.087 | Hδ1,7.128, Hε1, 10.076 |
| Q28 | 7.840 | 4.242 | 1.927 | 1.800 | Hγ, 2.143, 2.026 |
| L29 | 7.742 | 3.904 | 2.015 | 1.689 | Hδ, 0.915, 0.892 |
| T30 | 7.739 | 4.266 | 4.143 | | Hγ2, 1.078 |
| D31 | 8.689 | 4.926 | 2.171 | 1.922 | |
| R32 | 8.659 | 4.790 | 1.565 | 1.273 | Hγ, 1.363 |
| C33 | 7.363 | 5.220 | 3.127 | 2.634 | |
| V34 | 9.238 | 4.549 | 2.119 | | Hγ, 0.905, 0.919 |
| G35 | 8.343 | 4.781, 3.669 | | | |
| N36 | 7.601 | 4.923 | 2.876 | 2.659 | |
| C37 | 7.357 | 4.734 | 2.998 | 2.880 | |

### 5.2.11. Thermal, enzymatic and serum stability of coffeetides

The stability of coffeetides against heat, acid, enzyme and human serum was evaluated by RP-UPLC. Figure 5.17 indicates that after treating with boiling water, cC1 remained more than 88% intact while 90% of the peptide was stable after it was treated with acid for 2 h. Coffeetide cC1 also displays high tolerance to endopeptidase pepsin and exopeptidase aminopeptidase I for 6 h, with more than 90% peptide remaining. Furthermore, more than 85% of peptide survived after the incubation with human serum for 36 h. Prolonging the thermal, acid and enzymatic treatment does not change the trend of degradation. This is in agreement with other plant CRPs that contain a compact structure, whose thermal, acidic stability were examined up to 2 h treatment and enzymatic stability examined up to 6 h treatment .”

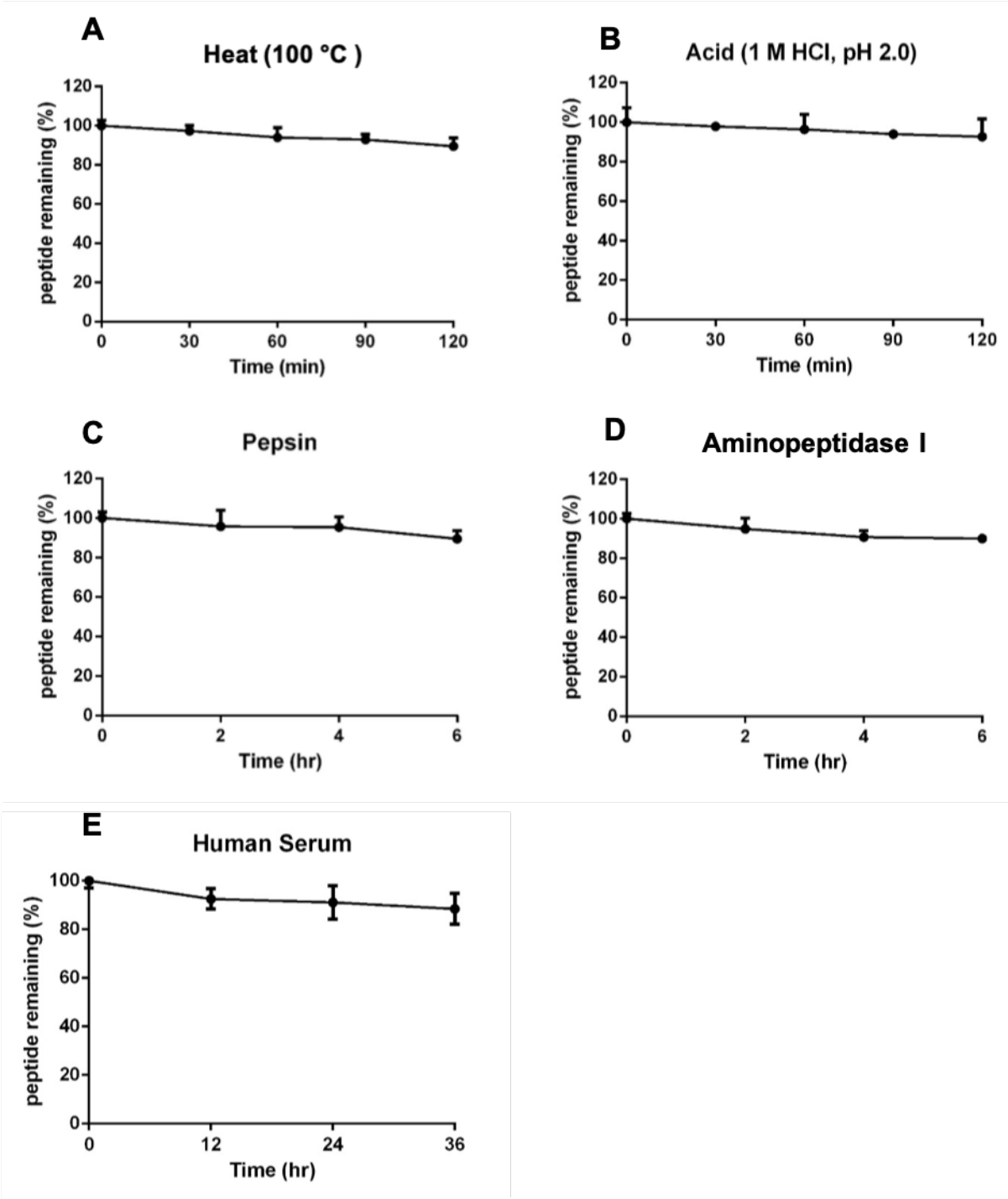



**Figure 5.17. Stability assays of coffeetide cC1.** Peptide remaining after treated with (A) 100 ◦C for 2 h, (B) acidic condition in HCl (pH 2.0) for 2 h, (C) exopeptidase enzyme pepsin for 6 h in buffer at 37 ◦C, (D) endopeptidase enzyme aminopeptidase I for 6 h in buffer at 37 ◦C and (E) human serum at 37 ◦C for 36 h.

### 5.2.12. Thermodynamics of $Fe^{3+}$- and $Mg^{2+}$- cC1 Binding Determined by ITC

Generally, ITC assay was employed to estimate the possible binding between ions and proteins [329]. Hence, it was employed to investigate the binding activity of cC1 with four metal ions. 200 µM of $K^+$, $Mg^{2+}$, $Ca^{2+,}$ and $Fe^{3+}$ were titrated into 20 µM of cC1, and the thermodynamic parameters for binding reactions were obtained at 20 °C (Figure 5.18). Results showed that cC1 did not have any measurable binding affinity to $Ca^{2+}$, while it binds to $K^+$, $Mg^{2+}$, and $Fe^{3+}$ with binding affinity $K_D$ of 1.18 ± 0.81, 4.19 ± 4.79 and 1.67 ± 1.80 µM, respectively. The ITC data of $K^+$, $Mg^{2+}$, and $Fe^{3+}$ fit well to a model of approximately one binding site per monomer. To further validate the binding activity, the same experiment was conducted using 400 µM of $K^+$, $Mg^{2+,}$ and $Fe^{3+}$ with 40 µM of cC1. The binding of $K^+$ could not be determined with confidence because after eliminating the unusual points, the differential power (DP) value did not change. $Mg^{2+}$ and $Fe^{3+}$ were confirmed binding to cC1 with a doubled DP value.

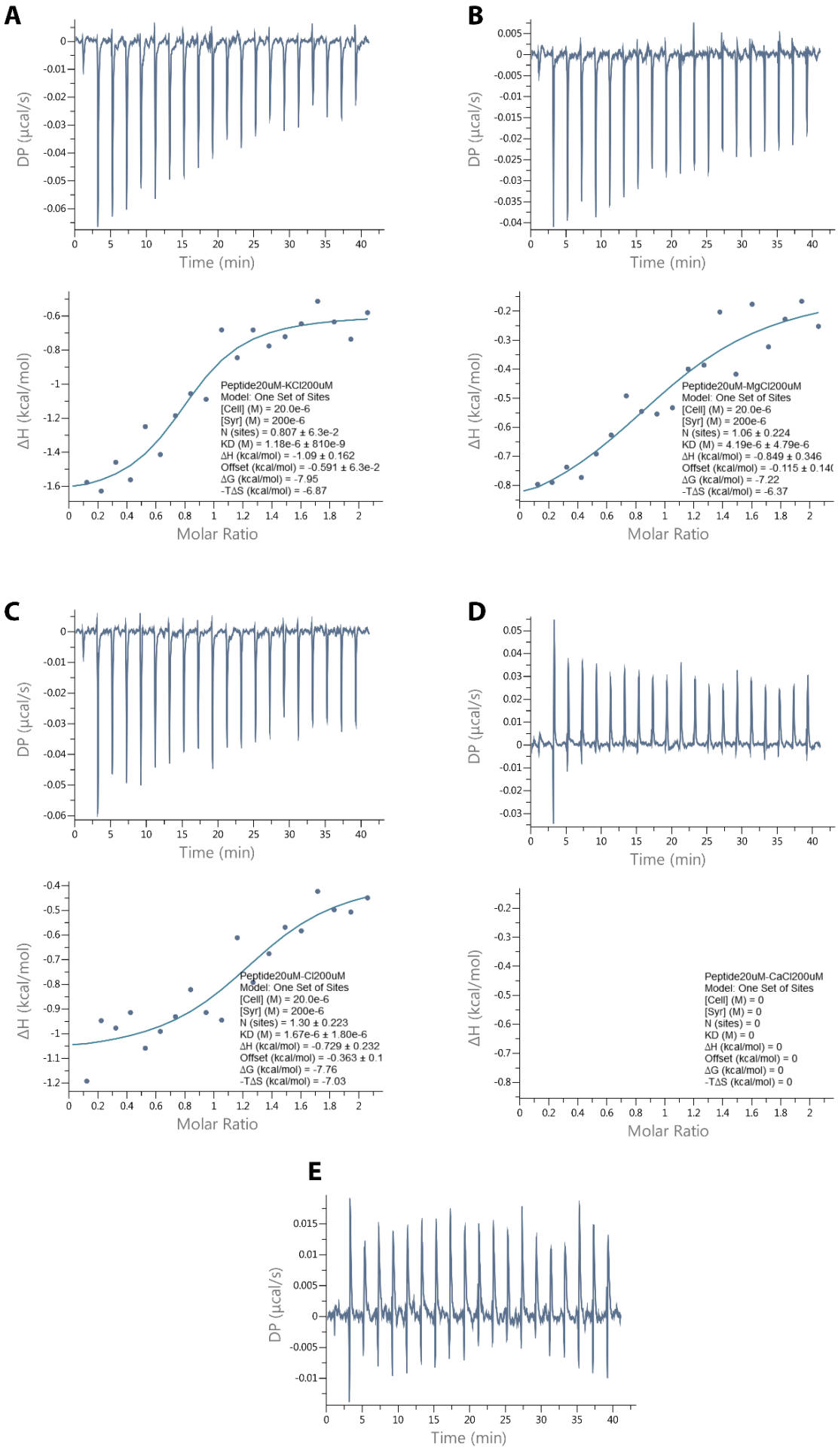

A
DP (µcal/s)
Time (min)
ΔH (kcal/mol)
Molar Ratio
Peptide20uM-KCl200uM
Model: One Set of Sites
[Cell] (M) = 20.0e-6
[Syr] (M) = 200e-6
N (sites) = 0.807 ± 6.3e-2
KD (M) = 1.18e-6 ± 810e-9
ΔH (kcal/mol) = -1.09 ± 0.162
Offset (kcal/mol) = -0.591 ± 6.3e-2
ΔG (kcal/mol) = -7.95
-TΔS (kcal/mol) = -6.87
B
DP (µcal/s)
Time (min)
ΔH (kcal/mol)
Molar Ratio
Peptide20uM-MgCl200uM
Model: One Set of Sites
[Cell] (M) = 20.0e-6
[Syr] (M) = 200e-6
N (sites) = 1.06 ± 0.224
KD (M) = 4.19e-6 ± 4.79e-6
ΔH (kcal/mol) = -0.849 ± 0.346
Offset (kcal/mol) = -0.115 ± 0.14(
ΔG (kcal/mol) = -7.22
-TΔS (kcal/mol) = -6.37
C
DP (µcal/s)
Time (min)
ΔH (kcal/mol)
Molar Ratio
Peptide20uM-Cl200uM
Model: One Set of Sites
[Cell] (M) = 20.0e-6
[Syr] (M) = 200e-6
N (sites) = 1.30 ± 0.223
KD (M) = 1.67e-6 ± 1.80e-6
ΔH (kcal/mol) = -0.729 ± 0.232
Offset (kcal/mol) = -0.363 ± 0.1
ΔG (kcal/mol) = -7.76
-TΔS (kcal/mol) = -7.03
D
DP (µcal/s)
Time (min)
ΔH (kcal/mol)
Molar Ratio
Peptide20uM-CaCl200uM
Model: One Set of Sites
[Cell] (M) = 0
[Syr] (M) = 0
N (sites) = 0
KD (M) = 0
ΔH (kcal/mol) = 0
Offset (kcal/mol) = 0
ΔG (kcal/mol) = 0
-TΔS (kcal/mol) = 0
E
DP (µcal/s)
Time (min)

**Figure 5.18. ITC binding assays.** The calorimetric titration isotherms of the binding interaction between coffeetide cC1 and (A) $K^+$, (B) $Mg^{2+}$, (C) $Fe3^+$, (D) $Ca^{2+}$ in 10 mM Tris buffer with 100mM NaCl (pH 6.3) at 298.15 K. (D) Binding interaction between coffeetide cC1 and buffer.

### 5.2.13. Biological activity of coffeetide cC1

To evaluate the potential functions of coffeetide cC1, a few preliminary biological assays were performed. To show that cC1 has the potential to be orally active biotherapeutics, the cytotoxicity effect of cC1 has been evaluated. Figure 5.19 showed that after the incubation of cC1 with HeLa cells and HUVEC-CS cells, no changes in cell viability were observed in both cell lines at concentrations up to 100 μM. Additionally, both LDH and MTT assay showed that coffeetide cC1 is non-cytotoxic to the cardiomyocyte H9c2 cells at concentrations up to 80 μM (Figure 5.20).

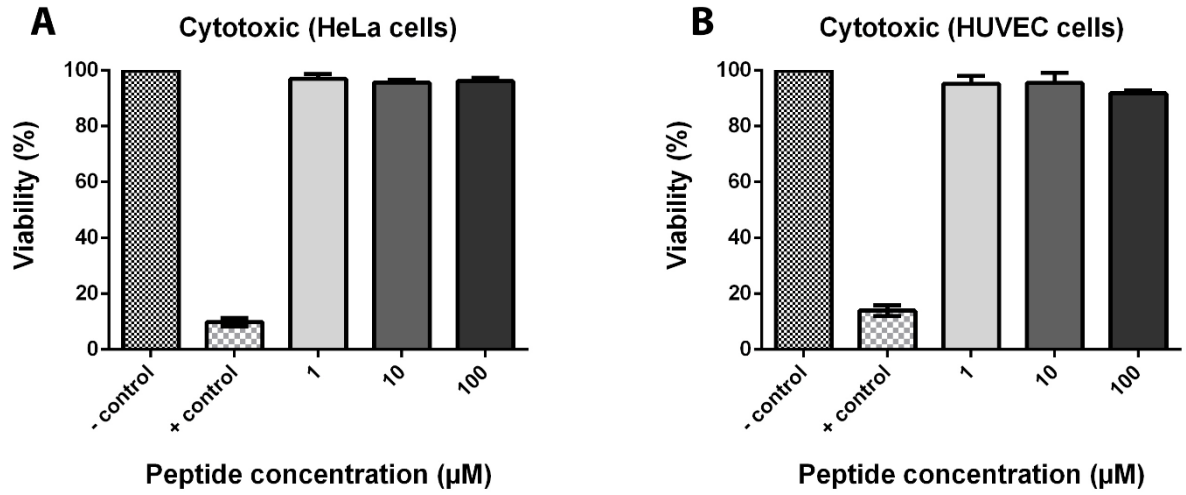



**Figure 5.19. Cytotoxic activity of coffeetide cC1 on HeLa cells and HUVEC-CS cells.** Different concentrations (1, 10 and 100 μM) of cC1 were added to incubated with (A) HeLa and (B) HUVEC-CS cells in a 96-well plate at 37 °C. After 24 h incubation, MTT assay was performed and the measured absorbance was used to determine the cell viability. 0.1% DMSO and 1% Triton-X were employed as the negative and positive control, respectively. All results are expressed as mean ± S.E.M. (n=3).

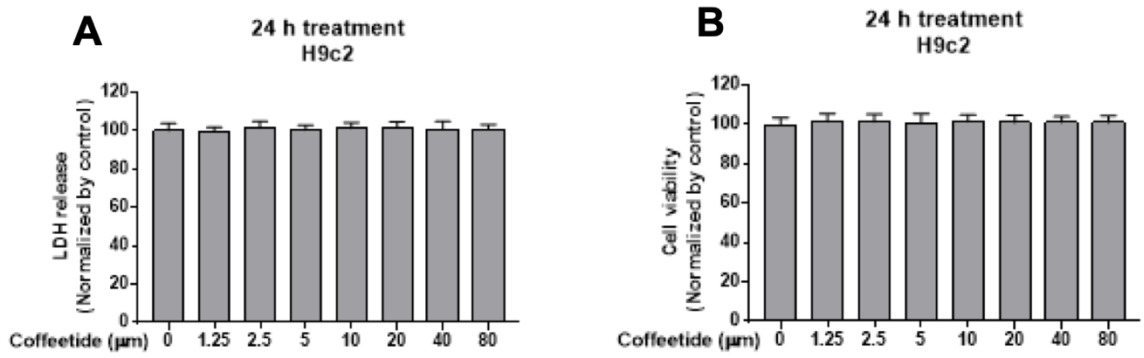



**Figure 5.20. Cytotoxic activity of coffeetide cC1 on H9c2 cells.** (A) LDH release in culture medium after incubation with different concentrations of cC1 (0 – 80 μM) for 48 h at 37 °C. The culture medium was mixed with the LDH reagent at a ratio of 9:1 and incubated at 37 °C for 2 h. The colorimetric quantification was performed at 450 nm wavelength in experimental triplicates. (B) Cell viability measured by MTT assay after incubation with cC1 at concentrations up to 80 μM. All results are normalized by the results of control, which are the cells that have not been treated with cC1.

Furthermore, a study was employed to investigate the possible bioactivity of cC1 using human-derived neuroblastoma SH-SY5Y cells as a model. Figure 5.21 revealed that cC1 was non-cytotoxic to SH-SY5Y cells at concentrations up to 80 μM, which was evaluated by LDH release assay. However, MTT assay showed that by incubating with coffeetide cC1, the cell viability of SH-SY5Y cells has increased in a dose-dependent manner and reached a 20% increase at a concentration of 80 μM. The results showed that cC1 could enhance SH-SY5Y cell metabolism and hence possesses potential effects on the nervous system.

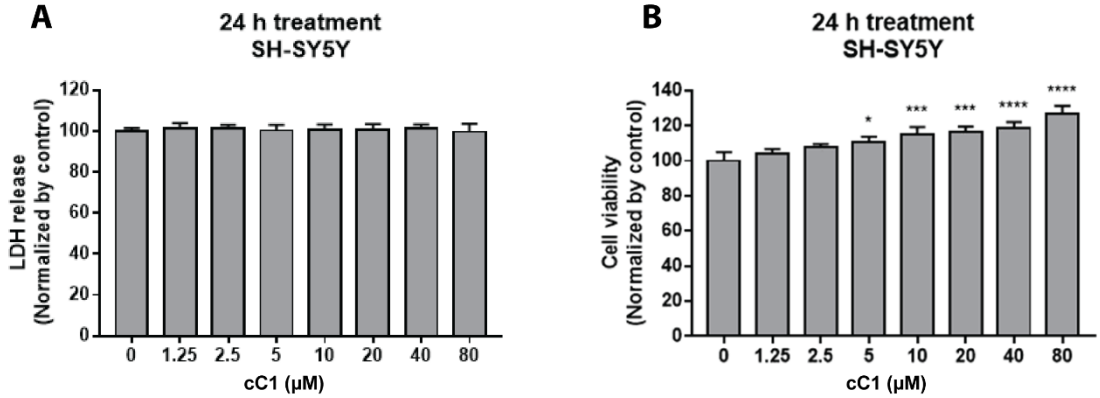



**Figure 5.21. Cytotoxic activity of coffeetide cC1 on SH-SY5Y cells. (A)** LDH release in culture medium after incubation with different concentrations of cC1 (0 – 80 μM) for 48 h at 37 °C. The culture medium was mixed with the LDH reagent at a ratio of 9:1 and incubated at 37 °C for 2 h. The colorimetric quantification was performed at 450 nm wavelength in

experimental triplicates. (B) Cell viability measured by MTT assay after incubation with cC1 at concentrations up to 80 μM. All results are normalized by the results of control, which are the cells that have not been treated with cC1. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001, compared to the control.

Moreover, a scratch assay has been performed on A431 cells to evaluate the potential effect of coffeetide cC1 on cell migration under serum-free conditions. Figure 5.22 and Figure 5.23 showed that after co-incubation with cC1 for 6 h, the extent of cell migration (150 ± 20 μm) was higher than the control groups (120 ± 20 μm), suggesting that cC1 may have a mild wound-healing effect.

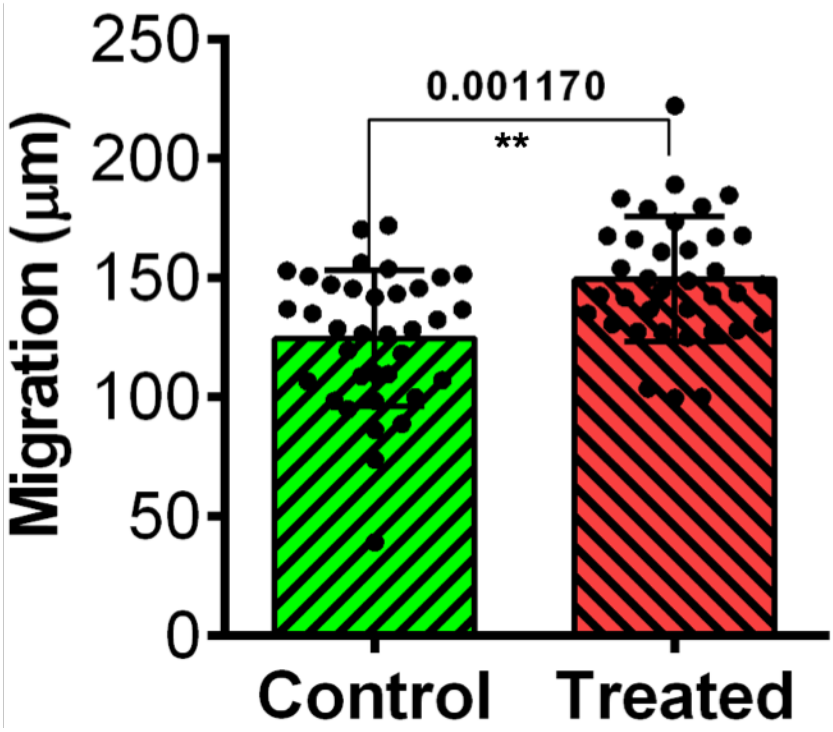



**Figure 5.22. Coffeetide cC1 enhanced the rate of cell migration of A431 cells.** There was a significant increase in the extent of cell migration in coffeetide cC1-treated cells compared with the control medium at 6 h. Data shown are expressed as mean ± S.E.M. (n=3), *p<0.05, 09ikm **p<0.01.

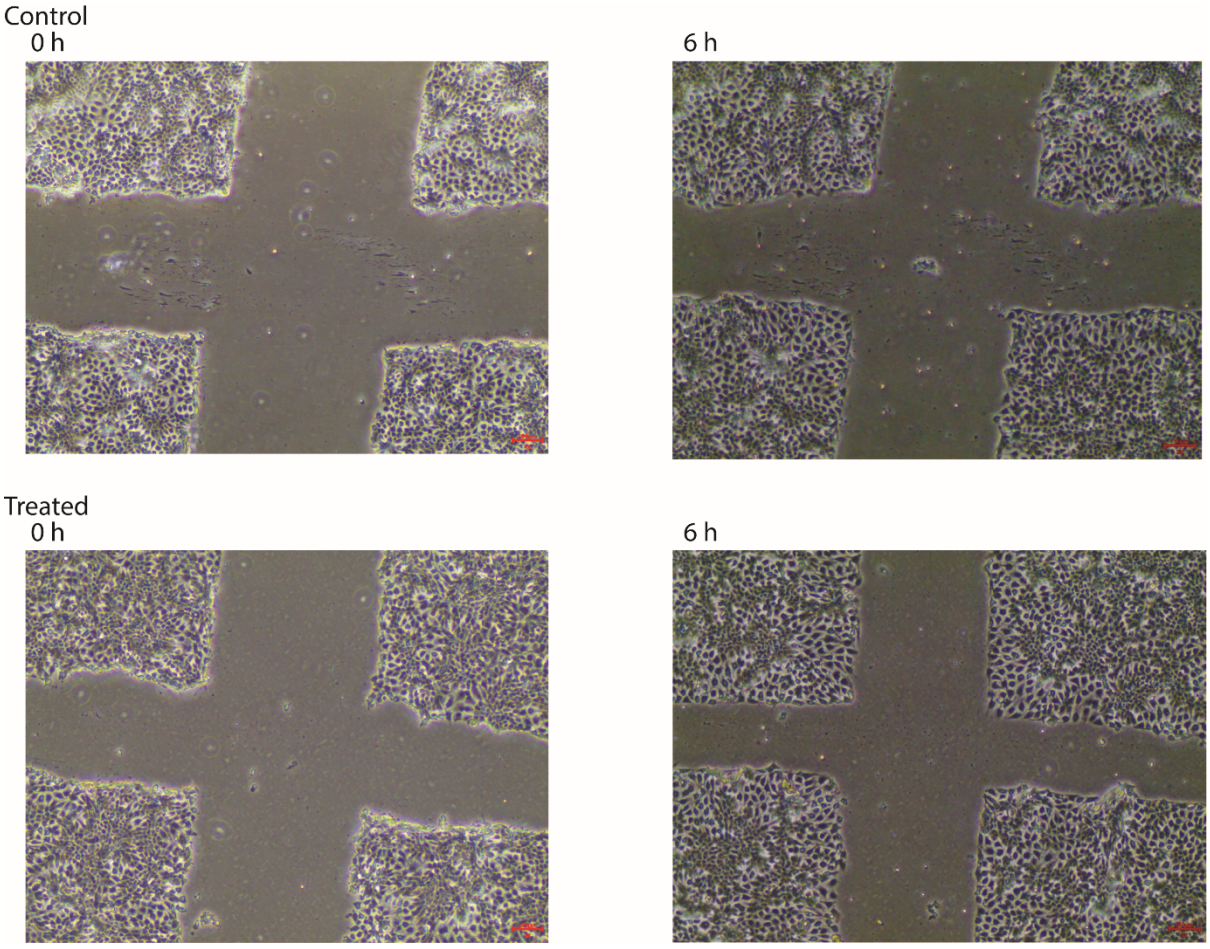



**Figure 5.23. Representative images of A431 cell scratch assays.** Images of cells immediately after the scratches have been made and then incubated with and without the presence of coffeetide cC1 for 6 h.

## 5.3. Discussion

### 5.3.1. The occurrence of non-chitin-binding 8C-HLPs in Plant kingdom

Genetic divergence within plant phyla results from mutations in the mature peptide, which also leads to functional diversification [21]. The phylogenetic tree constructed to discover the genetic relationship between 8C-defensins, 8C-thionins, and 8C-HLPs showed that they formed different clusters due to sequence variance. Further classification divided 8C-HLPs into two subfamilies, chitin-binding, and non-chitin-biding 8C-HLPs. Until now, ginsentides are the only families of non-chitin-biding 8C-HLPs that have been identified from plants. Our study showed that coffeetides and the other ginsentides-like peptides are grouped under the same cluster as ginsentides, suggesting that they may belong to the non-chitin-binding 8C-HLPs subfamily. The presence of 89 putative non-chitin-biding 8C-HLPs from 12 plant families revealed their high distribution in Plant kingdom. The non-chitin-biding 8C-HLPs

were further grouped into clusters based on different plant families. Interestingly, coffeetides are the only non-chitin-binding 8C-HLPs from the Rubiaceae family and formed a separate branch from the other putative non-chitin-binding 8C-HLPs. It suggested that although coffeetides belong to the same chitin-binding 8C-HLPs subfamily as ginsentides, they may possess evolutionary and functional differences.

### 5.3.2. Well-conserved features in coffeetides

In this study, seven coffeetides (cC1, cC1b, cC1c, cL1, cL1b, cL1c, and cL2) were isolated and characterized from *C. canephora* husks, *C. liberica* leaves, and husks, using transcriptomic and proteomic methods. Together with another 27 coffeetides identified at the transcriptomic level from four Coffea species, all coffeetides were shown to possess eight cysteines and six inter cysteinyl loops with a diagnostic –CC- motif at the third and fourth positions. Currently, such cysteine motif can be only found in HLPs and CKAIs [32, 56, 57].

In addition to eight cysteine residues, two other residues Pro-71 and Gly-94, are also absolutely conserved among all the coffeetides. The size of inter-cysteine loops 1, 2, 3, 4 and 6 are quite conserved, with six residues in loop 1, four to six in loop 2, two in loop 3, one in loop 4 and three in loop 6. In contrast, the inter-cysteine size of loop 5 varies from six to eleven residues. In addition, there is no conserved residue in loop 5. Such variability suggests the functional plasticity of the coffeetide subfamily is likely contributed by loop 5 for plant defense and adaptation. Most of the coffeetides are negatively charged, and this feature is attributed to the highly negative charged residues at the N-terminus and loop 5.

### 5.3.3. Sequence comparison with 8C-HLPs

Based on the different cysteine motifs and disulfide connectivity, plant CRPs are divided into different families [27]. HLPs are family of CRPs that possess a conserved cysteine motif of CXnCXnCCXnCXnCXnCXnC and can be classified into two subfamilies, chitin-binding HLPs and non-chitin-binding HLPs [59]. Chitin-binding HLPs such as morintides, ginkgotides, and vaccatides, are HLPs that containing a conserved chitin-binding domain $SX\Phi X\Phi GGX_4Y$ in intercysteinyl loop 3, where X and Φ represent small and aromatic amino acid, respectively [21, 27, 57]. In contrast, non-chitin-binding HLPs are HLPs that lack the chitin-binding domain, such as ginsentides. It can be observed that with a consecutive –CC- motif at Cys III and Cys IV, coffeetides do not contain a chitin-binding domain. The pairwise alignment showed that coffeetide cC1 showed a low sequence identity (25.0-36.8%) and similarity (28.6-38.5%) with the reported chitin-binding 8C-HLPs (Table 5.6). In addition, chitin-binding 8C-HLPs are generally basic and hydrophilic, whereas coffeetides are acidic and hydrophobic. When compared to ginsentides, coffeetide cC1 showed higher sequence similarities (39.4-59.1%).

The similarities are attributed to the presence of a $CX_6CX_{4-6}CCX_2CXC_{5-11}CX_3C$ motif in coffeetides, which contains a –CC- and –CXC- motif like ginsentides. However, coffeetides are different from ginsentides in terms of the loop size and the lack of glycine-rich features. The size of inter-cysteine loops 1, 2, 3, 4 and 6 are quite conserved in coffeetides, with six residues in loop 1, four to six in loop 2, two in loop 3, one in loop 4 and three in loop 6. In contrast, the inter-cysteine size of loop 5 varies from six to eleven residues. In addition, there is no conserved residue in loop 5. Such variability suggests the functional plasticity of the coffeetide subfamily is likely contributed by loop 5 for plant defense and adaptation. The second common feature in coffeetides is that most of the coffeetides are negatively charged, which can be a clue for identifying the functions of coffeetides, such as ion channel interaction.

**Table 5.6. Pairwise alignment of coffeetide cC1 with ginsentides and chitin-binding 8C-HLPs**

| Peptide | Identity (%) | Similarity (%) | Charge | PI |
|---|---|---|---|---|
| **Ginsentides** | | | | |
| TP1 | 36.4 | 42.4 | 0 | 6.68 |
| TP2 | 36.4 | 42.4 | 0 | 6.68 |
| TP3 | 36.4 | 42.4 | 0 | 6.68 |
| TP4 | 50.0 | 59.1 | -1 | 4.37 |
| TP5 | 36.4 | 42.4 | -1 | 5.21 |
| TP6 | 45.5 | 45.5 | -1 | 4.00 |
| TP7 | 36.4 | 42.4 | 0 | 6.68 |
| TP8 | 33.3 | 39.4 | -1 | 5.07 |
| TP9 | 36.4 | 42.4 | 0 | 6.68 |
| TP10 | 33.3 | 39.4 | -1 | 5.07 |
| TP11 | 50.0 | 59.1 | +1 | 7.70 |
| TP12 | 40.9 | 45.5 | -1 | 5.24 |
| TP13 | 45.5 | 45.5 | -1 | 4.00 |
| TP14 | 32.4 | 44.1 | +4 | 8.74 |
| **8C-chitin-binding hevein-like peptides** | | | | |
| vH1 | 27.3 | 30.3 | +2 | 8.22 |
| vH2 | 27.3 | 30.3 | +2 | 8.22 |
| Fa-AMP1 | 29.7 | 35.1 | +2 | 8.22 |
| Fa-AMP2 | 29.7 | 35.1 | +2 | 8.22 |
| Pn-AMP1 | 36.8 | 36.8 | +4 | 8.76 |
| Pn-AMP2 | 36.8 | 36.8 | +4 | 8.76 |
| mO1 | 25.0 | 30.6 | +3 | 8.53 |
| mO2 | 25.7 | 28.6 | +3 | 8.53 |
| Hevein | 27.5 | 30.0 | -1 | 4.83 |
| gB1 | 30.8 | 38.5 | +3 | 8.50 |
| gB5 | 30.8 | 38.5 | +1 | 7.69 |
| gB7 | 30.8 | 38.5 | +2 | 8.20 |
| gB10 | 31.0 | 35.7 | +2 | 8.20 |

### 5.3.4. Sequence comparison of CKAIs

In 6C-CRP family, the unique feature of a consecutive -CC- motif was shared in two subfamilies, which are chitin-binding 6C-HLPs and CKAIs [20, 26]. CKAIs belong to the knottin family, which are ribosomally synthesized peptides with a knotted disulfide connectivity. The distinguishing characteristics of CKAIs comprise of three parts, which include the cystine knot arrangement of three interlocked disulfide bonds, proline-rich in sequence and the inhibitory anti-α-amylase effect.

Pairwise alignment between coffeetide cC1 and CKAIs revealed a low sequence identity (29.0-50.0%) and similarity (34.5-50.0%) when coffeetides have no proline-rich characteristic displayed in their sequences. Furthermore, the structure and disulfide connectivity of coffeetide cC1 revealed that coffeetides did not adopt the pure cystine-knot disulfide connectivity, which are the most important clues to classify CRPs under CKAIs family. In addition, CKAIs are much shorter in length (29-30 aa) and contain only six cysteine residues compared to the eight-cysteine-containing coffeetides. Taken together, the results suggested that coffeetides are not CKAIs.

### 5.3.5. Disulfide connectivity of coffeetides

To further prove that coffeetides are non-chitin-binding 8C-HLPs, disulfide mapping was performed on coffeetide cC1. Disulfide connectivity is an essential characteristic of the classification of CRPs. Thionins were known as plant toxins due to their toxicity towards bacteria [29], fungi [330], plant [331], and animal cells [133]. It can be classified based on cysteine numbers that *α/β*-thionins with eight Cys residues are referred to as 8C-thionins while those with six Cys are designated as 6C-thionins [12]. In 8C-thionins, the disulfide connectivity is CysI-CysVIII, CysII-CysVII, CysIII-CysVI and CysIV-CysV, an end-to-end disulfide bond linking the N- and C-termini, conferring a circular structural topology. Plant defensins are the best-known plant antimicrobial peptides distributed in more than 100 plant species. Based on the number of cysteine residues, plant defensins can be divided into 6C-, 8C- and 10-C defensins. In 8C-defensins, the linkage of disulfide bonds is CysI-CysVIII, CysII-CysV, CysIII-CysVI and CysIV-CysVII, a common characteristic containing an outer disulfide pair as an end-to-end inner disulfide bridge and the inner three pairs of disulfide bonds forming a cystine knot.

For chitin-binding 8C-HLPs, they contain a cystine knot motif with an additional disulfide bond at the C-termini. The disulfide connectivity is CysI-CysIV, CysII-CysV, CysIII-CysVI, and CysVII-CysVIII. Differently, non-chitin-binding 8C-HLPs like ginsentides, contain a

CysI–IV, CysIII–VII, CysII–VI and CysV–VIII disulfide connectivity. It can be observed that coffeetides adopted the same disulfide linkage as ginsentides. They contain a cystine-knot that is similar to that of chitin-binding HLPs, whereas their fourth penetrating disulfide bond CysV–VIII is unique to non-chitin-binding HLPs such as ginsentides [59]. Together, the findings further prove that coffeetides belong to the non-chitin-binding 8C-HLPs subfamily due to the sequence and disulfide linkages characteristics.

### 5.3.6. Highly constrained structure of coffeetides

NMR structural analysis revealed that coffeetide cC1 contains four disulfide bonds, three of which are embedded in the structural core (Cys II-VI, III-VII, and V-VIII). Therefore, the side chains of the other residues are exposed to the solvent and result in an amphipathic distribution of hydrophobic and hydrophilic side chains. Similar to ginsentide TP1, coffeetide cC1 possess a pseudocyclic structure in which both N- and C-terminus were fixed by the disulfide bond Cys I-IV and Cys V-VIII, respectively. The extra N-terminal tail was stabilized by the formation of pyroglutamate. These features combined with the tightly folded structure and intramolecular hydrogen bonds confer cC1 high stability against thermal, acidic, proteolytic and human serum-mediated degradation. To evaluate the structural similarities between coffeetide cC1 and ginsentides or other chitin-binding 8C-HLPs, pairwise structural alignment algorithm TM-align was employed and displayed as the TM-score. Coffeetide cC1 had a TM-score value of 0.302 and 0.283 as compared to chitin-binding 8C-HLPs vaccatide vH2 and morintide mO1, respectively. When compared to ginsentides, a TM-align score of 0.369 was observed. The results suggested that in terms of structure, coffeetide shared low structural similarity with both chitin-binding and non-chitin-binding 8C-HLPs.

### 5.3.7. Bioprocessing of coffeetides occurs through the secretory pathway

Our findings showed that coffeetide precursors comprise a three-domain architecture, which includes a signal peptide, a pro-peptide, and the mature peptide. It suggested that coffeetides undergo a secretory pathway that is similar to most of the plant CRPs [20, 26, 27, 332]. Secretory peptides were exported from cytoplasm with a signal peptide, which enables the peptide to go through the ER membrane. After being translocated, the signal peptide will be cleaved by signal peptidase (SPase). After the cleavage, the pro-peptide will be removed by endopeptidase and the mature peptide will be released for further post-translational modification. The N-terminal cleavage site of the mature domain is likely between G and Q of the pro-sequence, VHDGQEGE, because the longest isolated coffeetide cC1 contains QEGE (Q became Z after cleavage) at its N-terminus and VHDG is highly conserved in the precursor sequences.

The precursor arrangement comparison showed that coffeetides contain a different precursor architecture compared to 8C-defensins, 8C-thionins, and chitin-binding 8C-HLPs. In contrast, coffeetides adopts the same precursor arrangement as ginsentides, suggesting that coffeetides belong to the non-chitin-binding 8C-HLPs as ginsentides. However, the features of containing a shorter pro-peptide and longer mature peptide distinguish coffeetides from ginsentides. Taken together, coffeetides possess an absolutely conserved precursor arrangement which is different from other 8C-CRPs, providing insight into coffeetide biosynthesis and expand the existing library of non-chitin-binding HLPs.

### 5.3.8. Significance of converting coffee waste into the bioactive compound

The coffee industry has generated enormous amounts of coffee waste during the coffee processing stage. With the increasing production and need for coffee drinks, it is urgent to balance the production of coffee by-products with proper industrial applications. Previous studies have applied chemical and biotechnological processes to recover the fine chemicals or generate value-added products from the by-products. In the food industry, coffee pulps have been used to produce fruity aroma by solid-state fermentation while coffee husks were used to produce citric acid. In the health area, the bioactive compounds in the coffee waste have shown a potent inhibitory effect against enzymes and an effect on improving blood glucose metabolism, which makes coffee waste a source of drug development [333]. However, no such studies have been focused on the chemical spaces between 2 to 6 kDa of coffee by-products. With the advantages of large footprint, less off-target interactions, and highly compact structure, CRPs have gained increasing popularity in serving as potential therapeutics. For example, aglycin and vglycin extracted from pea and soybean, are CRPs that have a therapeutic effect on diabetes [334]. Additionally, roseltide rT7, which is a CRP isolated from *Hibiscus sabdariffa,* showed a mitochondria-targeting effect and thus delayed age-related diseases [335]. In this study, although the pharmacological effects of coffeetides have not been elucidated, it provides a clue of converting coffee waste into potential therapeutics.

### 5.3.9. Preliminary functional exploration of coffeetides

In our study, a few preliminary binding assays and cell-based assays were performed to investigate the potential functions of coffeetide cC1. ITC assay has shown that cC1 can bind to $Mg^{2+}$ and $Fe^{3+}$ with high binding affinity, which may be due to the presence of a highly negatively charged N-terminus. The 3D structure of cC1 showed that Glu2, Glu4, and Glu10 have formed a pocket at the N-terminus and hence became a possible binding site for ions. However, further studies still have to be done in order to discover the possible binding site.

By employing LDH and MTT assays, it can be observed that coffeetide cC1 is non-cytotoxic to HeLa cells, HUVEC-CS cells, H9c2 cells, and SH-SY5Y cells at the concentrations tested. However, cC1 was shown to enhance the metabolism of the human neuroblastoma SH-SY5Y cells. In addition, cC1 was shown to increase cell migration. These preliminary results suggest that cC1 may act as a potential non-cytotoxic biotherapeutics used for wound-healing and neurodegenerative diseases.

## 5.4. Conclusion

Here, we report the discovery and characterization of seven novel non-chitin-binding 8C-HLPs, coffeetides, from *C. canephora* and *C. liberica*, based on transcriptomic and proteomic analyses. Data mining revealed another 27 coffeetides from four *Coffea* species and 89 putative non-chitin-binding HLPs distributed in 12 plant families. Sequence analysis showed that coffeetides contain 35-40 residues and eight cysteines with a highly conserved –CC- motif at the third and fourth positions. Disulfide mapping and NMR analysis revealed connectivity of CysI–IV, CysIII–VII, CysII–VI and CysV–VIII and a highly compact pseudocyclic structure of coffeetides, which confer them resistance against thermal, acidic and enzymatic degradation. Biosynthesis analysis showed that coffeetide precursor sequences adopt a three-domain precursor arrangement like ginsentides but with a shorter pro-peptide, which is different from other 8C-CRPs. Phylogenetic analysis indicated that coffeetide form a distinct cluster of non-chitin-binding 8C-HLPs. Preliminary biological assays showed that cC1 is non-cytotoxic, iron- and magnesium-binding CRP that can improve cell migration and neuron cell metabolism. In addition, coffeetide cC1 can be chemically synthesized and folded within 3 h with the highest folding yield of 84.48%. In conclusion, our results suggest that coffeetides represent a new subfamily of non-chitin-binding 8C-HLPs, which expands the existing library of CRPs and provides new insight into their diversity and distribution in Plant kingdom. Furthermore, these coffeetides identified from *Coffea* husks revealed that coffee by-products could be a source for generating metabolically stable and orally bioavailable peptidyl drugs, which may benefit the sustainable development of nature.

# Summary, Conclusion and Future Outlook

Cysteine-rich peptides (CRPs) are naturally occurring peptides that play essential roles in plant host-defense and physiology. Their cross-linking disulfide bonds confer them high stability against thermal, acidic and enzymatic degradation, which provide them the structural basis for engineering potential orally active therapeutics. Based on different cysteine framework, CRPs are classified into different families and shown to display various bioactivities. For example, plant defensins are CRPs that possess anti-microbial, ion channel inhibition, and enzyme inhibitory effects. In contrast, knottins are CRPs that display hormone-like properties, enzyme inhibitory effect, insecticidal, antimicrobial, and anti-HIV functions. In my thesis, a rapid and general method using CRPs as unique chemical fingerprints to authenticate herbal products was developed. To further our understanding, CRPs from two specific medicinal plants have been isolated and extensively studied on their distribution, molecular diversity, and structures.

Chapter 3 describes a rapid and general method, designated as CRP fingerprinting, for quality control of medicinal plants. Using MALDI-TOF MS technique, the method employs CRPs as unique chemical markers for the authentication of species. Screening 100 medicinal plant extracts showed that CRP fingerprints are unique chemical markers regardless of the morphology, chemical composition, and origins of the plants. Compared to small molecules, CRPs are well-annotated and more easily detected by MALDI-TOF MS without being cluttered.

The roots of *Astragalus membranaceus* (RA) and the roots of *Hedysarum polybotrys* (RH) are two species widely used as traditional herbal medicine. However, with similar Chinese names and morphologies, RA is often substituted by RH, although their pharmacological activities are different. Herein, CRP fingerprinting was applied in differentiating these two species to validate the method. In this chapter, 40 RH and 51 RA samples were extracted with water and screened by MALDI-TOF MS. The screening result revealed the presence of two types of CRPs in RH samples, named hedytides hP1 and hP2. Similary, two types of CRPs, namely α-astratide aM1 and β-astratide bM1, was observed in RA samples. Sequence analysis revealed that both aM1 and hP1 share the same cysteine motif and high sequence similarity with PA1b-like peptides. In contrast, both bM1 and hP2 contain the same cysteine motif and share high sequence similarity with plant defensins. Although both species contain CRPs that belong to the same CRP family, the sequence variability makes these CRPs unique for distinguishing one species from another.

Furthermore, the conventional quality control method using UPLC requires two-day sample preparation time and 46 min for sample analysis, while the sample preparation time and

analytical time of CRP fingerprinting was 500-fold less, with 2h and 5 s, respectively. Unsupervised multivariate analyses such as PCA and HCA showed that RA and RH could be separated into two clusters based on their CRP fingerprints. In terms of classification performance, five models KNN, PLS-DA, SVM-DA, CART, and SIMCA revealed that the classification ability of CRP fingerprinting is as accurate as of UPLC. Therefore, CRP fingerprinting coupled with multivariate analyses can be used as a rapid and general approach for the authentication of medicinal plants.

Chapter 4 described the isolation and characterization of two families of CRPs, α- and *β*-astratides, from the roots of *Astragalus membranaceus*. The most abundant *α*-astratide with relative monoisotopic molecular weight $[M+H]^+$ of 3811.2 Da is named aM1, which contain 37 amino acid with six cysteine residues. Proteomic and transcriptomic analysis showed that *α*-astratide aM1 shares high sequence similarity with pea albumin 1-b (PA1b) and leginsulin, which are isolated from *Pisum sativum* (pea) and *Glycine max* (Soybean), respectively. PA1b is a disulfide-rich peptide that belongs to the cystine-knot inhibitor family and shows potent insecticidal activity [281]. In mammals, both PA1b and leginsulin were shown to be able to interfere with glucose homeostasis or have a relationship with insulin activity [284, 336]. Transcriptomic analysis revealed that *α*-astratide aM1 and other PA1b-like peptides are proline-rich, containing highly conserved loop 2 and unique five-domain architecture, which is different from other 6C-CRPs [1]. The cytotoxicity assay showed that aM1 is insecticidal against insect Sf9 cells with an $LD_{50}$ of 5.86 μM, whereas it is non-toxic to mammalian CHO-K1 cells. Structure prediction showed that aM1 contains the same binding site containing Phe10, Arg21, Ile23, and Leu27 as PA1b, which are the keys residues for the insecticidal activity [337]. The preliminary biological assay showed that aM1 reduces insulin secretion in mouse pancreatic β-cells at concentrations>5 μM. The action of aM1 may link to beta-cell function restoration and thus increase insulin sensitivity [313]. However, the details of aM1 putative mechanism on regulating blood glucose still remain unknown.

In contrast, the most abundant *β*-astratide with relative monoisotopic molecular weight $[M+H]^+$ of 4724.1 Da is named bM1, containing 45 amino acids and eight cysteine residues. Sequence analysis revealed that bM1 shows high sequence similarity with plant defensins. Similar to other plant defensins, *β*-astratide bM1 is antifungal and shown to inhibit the growth of four phytopathogenic fungi with $IC_{50}$ ranging from 2.7 to 130.3 μg/mL. Morphological observation of the fungi before and after bM1 treatment revealed that the fungi displayed shorter and highly branched hyphae. Plant defensins are classified based on their cysteine motifs, and the common motif found in 8C-plant defensins is $CX_{10}CX_5CX_3CX_{[9-10]}CX_{[6-8]}CXCX_3C$. However,

transcriptomic data mining revealed that bM1 contains a unique CXCXC motif at the C-terminus, which is a feature only found in plant defensins from the Fabaceae family.

Like other CRPs, both aM1 and bM1 are highly stable against heat, acid and enzymatic degradation. Importantly, aM1 is the first reported PA1b-like peptides isolated from medicinal plants while bM1 represents a new subfamily of plant defensins with a unique C-terminal motif. Both peptides expanded the existing library of plant CRPs and showed their potential to be orally active therapeutics.

Chapter 5 describes the isolation and characterization of a novel family of non-chitin-binding 8C-hevein-like peptides (HLPs) from *Coffea canephora* and *Coffea liberica,* designated as coffeetides cC1, cC1b, cC1c, cL1, cL1b, cL1c, and cL2. Transcriptomic data mining and phylogenetic analysis revealed that coffeetides and coffeetide-like peptides are widely distributed in 24 plant species from 12 different plant families. They are 8C-CRPs that share the same cysteine motif of CXnCXnCCXnCXnCXnCXnC as 8C-HLPs but lack of a chitin-binding domain. Disulfide mapping and NMR determination showed that coffeetide cC1 contains four disulfide bonds forming a pseudocyclic structure which confers high stability against heat, acid, and proteolytic degradation. Thus far, this disulfide connectivity is only found in ginsentides, which are a family of newly discovered non-chitin-binding 8C-HLPs [59]. In addition, coffeetides adopt a three-domain precursor arrangement comprising of a signal peptide, a pro-peptide domain, and a mature peptide, which is similar to ginsentides but different from all the other 8C-CRPs [12]. However, coffeetides are different from ginsentides in terms of lower abundance in Gly residues and their shorter pro-peptide domain.

A few preliminary biological assays have been performed on coffeetide cC1. MTT and LDH assays showed that cC1 is non-cytotoxic to HeLa cells, HUVEC-CS cells, H9c2 cardiomyocyte cells, and neuroblastoma SH-SY5Y cells but could enhance the metabolism of SH-SY5Y cells in a dose-dependent manner. In addition, ITC assay showed that cC1 could bind to $Fe^{3+}$ and $Mg^{2+}$ while cell migration assay revealed that cC1 could increase cell migration by approximately 25%. These preliminary results suggest that coffeetides may act as potential peptidyl therapeutics that could affect neurodegenerative diseases and wound-healing.

In conclusion, this thesis reports the identification of novel CRPs from *C. canephora*, *C. liberica,* and *A. membranaceus* as well as the roles of CRPs in the authentication of medicinal plants. The discovery of coffeetides and astratides shows the molecular diversity across different CRP families in nature. This molecular diversity of CRP scaffolds may play an essential role in the convergent evolution of plants to deliver peptides metabolic stability, ability to solve defensin-related problems. In contrast, their hypervariable sequences and

structures within a CRP family may be a result of divergent evolution, which enables a family of CRPs to exhibit multiple biological functions. The discovery of CRPs also increases our understanding of the diversity of CRP scaffolds in nature, which allows the development of peptide therapeutics. The cystine-stabilized feature provides a possible scaffold for grafting bioactive peptide to intracellular targets. Specifically, their backbone segments between cysteine residues are suitable for modification to incorporate bioactive peptides that are usually not stable against harsh conditions. In addition, our study showed that CRPs are unique fingerprints that are widely present in medicinal plants. Their presence revealed their bioactive roles of being responsible for some of the pharmacological effects in medicinal plants. Taken together, this thesis expanded the existing library of CRPs and explored their roles as chemical markers for authenticating medicinal plants. Furthermore, this work provides new avenues in the discovery and development of novel and stable peptidyl drugs.

# Publications and Presentations

1. **Huang J.,** Wong K. H., Tay S. V., et al. (2019). Astratides: Insulin-Modulating, Insecticidal, and Antifungal Cysteine-Rich Peptides from Astragalus membranaceus. Journal of Natural Products, 82 (2), 194-204.

2. Shen. Y., Xu L., **Huang J.,** et al. (2019). Potentides: Novel Cysteine-Rich Peptides with Unusual Disulfide Connectivity from Potentilla anserine. ChemBioChem

3. Dutta B., **Huang J. (equal contribution)**, To J. & Tam J.P. (2019). LIR Motif-Containing Hyperdisulfide β-Ginkgotide is Cytoprotective, Adaptogenic, and Scaffold-Ready. Molecules, 24(13): 2417.

4. **Huang J.**, Wong K. H., Tay S. V., How A., Tam J.P. (2019). Cysteine-rich peptide fingerprinting as a general method for herbal analysis to differentiate Radix Astragali and Radix Hedysarum. Frontiers in Plant Science, 10: 973.

5. **Huang J.,** Wong K. H., Tam J. P. (2017). Profiling and Authentication of Herbal Products. The FASEB Journal, 2017, 31(1_supplement): lb121-lb121.

6. **Huang J.,** Wong K. H., Tam J. P. (2019). Coffeetides: Iron-chelating Cysteine-rich Peptides from Coffee Waste, *2019 American Peptide Symposium*, Monterey, USA

7. **Huang J.**, Wong K. H., Tan W. L., et al. Identification and characterization of a carboxypeptidase inhibitor from *Lycium barbarum* (In preparation)

8. **Huang J.**, Wong K. H., Stephanie V. T., Serra A., Sze S. K., and Tam J.P. Coffeetides: a new class of non-chitin-binding hevein-like peptides from coffee waste (In preparation)

9. Stephanie T., Wong, K.H., **Huang, J.**, & Tam, J. P. Discovery and characterization of cysteine-rich peptides repeats *in planta* (in preparation)

## References


1. Huang, J., et al., *Astratides: Insulin-Modulating, Insecticidal, and Antifungal Cysteine-Rich Peptides from Astragalus membranaceus.* J. Nat. Prod., 2019. **82**(2): p. 194-204.
2. Li, F., et al., *Are we seeing a resurgence in the use of natural products for new drug discovery?* Expert Opin. Drug Discov., 2019. **14**(5): p. 417-420.
3. Atanasov, A.G., et al., *Discovery and resupply of pharmacologically active plant-derived natural products: A review.* Biotechnol Adv, 2015. **33**(8): p. 1582-1614.
4. Craik, D.J., et al., *The future of peptide-based drugs.* Chem. Biol. Drug Des., 2013. **81**(1): p. 136-147.
5. Newman, D.J. and G.M. Cragg, *Natural products as sources of new drugs over the 30 years from 1981 to 2010.* J. Nat. Prod., 2012. **75**(3): p. 311-335.
6. Fosgerau, K. and T. Hoffmann, *Peptide therapeutics: current status and future directions.* Drug Discov. Today, 2015. **20**(1): p. 122-128.
7. Katz, C., et al., *Studying protein–protein interactions using peptide arrays.* Chem. Soc. Rev., 2011. **40**(5): p. 2131-2145.
8. Padhi, A., et al., *Antimicrobial peptides and proteins in mycobacterial therapy: current status and future prospects.* Tuberculosis, 2014. **94**(4): p. 363-373.
9. Giordano, C., et al., *Neuroactive peptides as putative mediators of antiepileptic ketogenic diets.* Front. Neurol., 2014. **5**: p. 63.
10. Market Peptide Therapeutics , *Global Industry Analysis, Size, Share, Growth, Trends and Forecast 2014–2020. April 2017*.
11. Sels, J., et al., *Plant pathogenesis-related (PR) proteins: a focus on PR peptides.* Plant Physiol. Bioch., 2008. **46**(11): p. 941-950.
12. Tam, J.P., et al., *Antimicrobial Peptides from Plants.* Pharmaceuticals, 2015. **8**(4): p. 711-757.
13. Selitrennikoff, C.P., *Antifungal proteins.* Appl. Environ. Microbiol., 2001. **67**(7): p. 2883-2894.
14. Van Loon, L.C., M. Rep, and C.M. Pieterse, *Significance of inducible defense-related proteins in infected plants.* Annu. Rev. Phytopathol., 2006. **44**: p. 135-162.
15. Shai, Y., *Mode of action of membrane active antimicrobial peptides.* J. Pept. Sci., 2002. **66**(4): p. 236-248.
16. Rao, A.G., *Antimicrobial peptides.* Mol. Plant-Microbe Interact, 1995. **8**(6): p. 13.
17. Montalbán-López, M., et al., *Discovering the bacterial circular proteins: bacteriocins, cyanobactins, and pilins.* J. Biol. Chem., 2012. **287**(32): p. 27007-27013.

18. Silverstein, K.A., et al., *Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants.* Plant J., 2007. **51**(2): p. 262-280.
19. Hammami, R., et al., *PhytAMP: a database dedicated to antimicrobial plant peptides.* Nucleic Acids Res., 2008. **37**(suppl_1): p. D963-D968.
20. Kini, S.G., et al., *Studies on the Chitin Binding Property of Novel Cysteine-Rich Peptides from Alternanthera sessilis.* Biochemistry, 2015. **54**(43): p. 6639-6649.
21. Kini, S.G., et al., *Morintides: cargo-free chitin-binding peptides from Moringa oleifera.* BMC Plant Biol., 2017. **17**(1): p. 68-80.
22. Porto, W.F., et al., *In silico identification of novel hevein-like peptide precursors.* Peptides, 2012. **38**(1): p. 127-136.
23. Wong, K.H., et al., *beta-Ginkgotides: Hyperdisulfide-constrained peptides from Ginkgo biloba.* Sci. Rep., 2017. **7**(1): p. 6140-6152.
24. Shen, Y., et al., *Potentides: Novel Cysteine-Rich Peptides with Unusual Disulfide Connectivity from Potentilla anserina.* ChemBioChem, 2019.
25. Kumari, G., et al., *Cysteine-Rich Peptide Family with Unusual Disulfide Connectivity from Jasminum sambac.* J. Nat. Prod., 2015. **78**(11): p. 2791-2799.
26. Nguyen, P.Q., et al., *Allotides: Proline-Rich Cystine Knot alpha-Amylase Inhibitors from Allamanda cathartica.* J. Nat. Prod., 2015. **78**(4): p. 695-704.
27. Wong, K.H., et al., *Ginkgotides: Proline-Rich Hevein-Like Peptides from Gymnosperm Ginkgo biloba.* Front. Plant Sci., 2016. **7**: p. 1639-1653.
28. Kaas, Q., J.-C. Westermann, and D.J. Craik, *Conopeptide characterization and classifications: an analysis using ConoServer.* Toxicon, 2010. **55**(8): p. 1491-1509.
29. De Caleya, R.F., et al., *Susceptibility of phytopathogenic bacteria to wheat purothionins in vitro.* Appl. Environ. Microbiol., 1972. **23**(5): p. 998-1000.
30. Balls, A., W. Hale, and T. Harris, *A crystalline protein obtained from a lipoprotein of wheat flour.* Cereal Chem., 1942. **19**(19): p. 279-288.
31. Stec, B., *Plant thionins–the structural perspective.* Cell Mol. Life Sci., 2006. **63**(12): p. 1370-1385.
32. Epple, P., K. Apel, and H. Bohlmann, *An Arabidopsis thaliana thionin gene is inducible via a signal transduction pathway different from that for pathogenesis-related proteins.* Plant Physiol., 1995. **109**(3): p. 813-820.
33. Stec, B., U. Rao, and M.M. Teeter, *Refinement of purothionins reveals solute particles important for lattice formation and toxicity. Part 2: Structure of β-purothionin at 1.7 Å resolution.* Acta Crystallogr., Sect. D: Biol. Crystallogr., 1995. **51**(6): p. 914-924.
34. Hughes, P., et al., *The cytotoxic plant protein, β-purothionin, forms ion channels in lipid membranes.* J. Biol. Chem., 2000. **275**(2): p. 823-827.
35. Pelegrini, P.B. and O.L. Franco, *Plant γ-thionins: novel insights on the mechanism of action of a multi-functional class of defense proteins.* Int. J. Biochem. Cell Biol., 2005. **37**(11): p. 2239-2253.

36. Thomma, B.P., B.P. Cammue, and K. Thevissen, *Plant defensins.* Planta, 2002. **216**(2): p. 193-202.
37. Sharma, P. and A. Lönneborg, *Isolation and characterization of a cDNA encoding a plant defensin-like protein from roots of Norway spruce.* Plant Mol. Biol., 1996. **31**(3): p. 707-712.
38. Terras, F.R., et al., *Small cysteine-rich antifungal proteins from radish: their role in host defense.* The Plant Cell, 1995. **7**(5): p. 573-588.
39. Chiang, C. and L. Hadwiger, *The Fusarium solani-induced expression of a pea gene family encoding high cysteine content proteins.* Mol. Plant-Microbe Interact, 1991. **4**(4): p. 324-331.
40. Park, H.C., et al., *Characterization of a stamen-specific cDNA encoding a novel plant defensin in Chinese cabbage.* Plant Mol. Biol., 2002. **50**(1): p. 57-68.
41. Shafee, T.M., et al., *The Defensins Consist of Two Independent, Convergent Protein Superfamilies.* Mol. Biol. Evol., 2016. **33**(9): p. 2345-2356.
42. Bontems, F., et al., *Three-dimensional structure of natural charybdotoxin in aqueous solution by 1H-NMR Charybdotoxin possesses a structural motif found in other scorpion toxins.* Eur. J. Biochem., 1991. **196**(1): p. 19-28.
43. Gao, A.-G., et al., *Fungal pathogen protection in potato by expression of a plant defensin peptide.* Nat. Biotechnol., 2000. **18**(12): p. 1307.
44. de Oliveira Carvalho, A. and V.M. Gomes, *Plant defensins—prospects for the biological functions and biotechnological properties.* Peptides, 2009. **30**(5): p. 1007-1020.
45. Lay, F.T., F. Brugliera, and M.A. Anderson, *Isolation and properties of floral defensins from ornamental tobacco and petunia.* Plant Physiol., 2003. **131**(3): p. 1283-1293.
46. Chen, K.-C., et al., *A novel defensin encoded by a mungbean cDNA exhibits insecticidal activity against bruchid.* J. Agric. Food Chem., 2002. **50**(25): p. 7258-7263.
47. Méndez, E., et al., *Primary structure of ω-hordothionin, a member of a novel family of thionins from barley endosperm, and its inhibition of protein synthesis in eukaryotic and prokaryotic cell-free systems.* Eur. J. Biochem., 1996. **239**(1): p. 67-73.
48. Terras, F., et al., *Analysis of two novel classes of plant antifungal proteins from radish (Raphanus sativus L.) seeds.* J. Biol. Chem., 1992. **267**(22): p. 15301-15309.
49. Bloch, C. and M. Richardson, *A new family of small (5 kDa) protein inhibitors of insect α-amylases from seeds or sorghum (Sorghum bicolor (L) Moench) have sequence homologies with wheat γ-purothionins.* FEBS Lett., 1991. **279**(1): p. 101-104.
50. Thevissen, K., et al., *Fungal membrane responses induced by plant defensins and thionins.* J. Biol. Chem., 1996. **271**(25): p. 15018-15025.

51. Van Der Weerden, N.L., F.T. Lay, and M.A. Anderson, *The plant defensin, NaD1, enters the cytoplasm of Fusarium oxysporum hyphae.* J. Biol. Chem., 2008. **283**(21): p. 14445-14452.
52. Archer, B., *The proteins of Hevea brasiliensis latex. 4. Isolation and characterization of crystalline hevein.* Biochemical Journal, 1960. **75**(2): p. 236.
53. Boller, T. and J. Metraux, *Extracellular localization of chitinase in cucumber.* Physiol. Mol. Plant P., 1988. **33**(1): p. 11-16.
54. Rice, R.H. and M.E. Etzler, *Subunit structure of wheat germ agglutinin.* Biochem. Bioph. Res. Co., 1974. **59**(1): p. 414-419.
55. Younes, I. and M. Rinaudo, *Chitin and chitosan preparation from marine sources. Structure, properties and applications.* Mar. Drugs, 2015. **13**(3): p. 1133-1174.
56. Koo, J.C., et al., *Two hevein homologs isolated from the seed of Pharbitis nil L. exhibit potent antifungal activity.* BBA - Protein Structure and Molecular Enzymology, 1998. **1382**(1): p. 80-90.
57. Wong, K.H., et al., *Vaccatides: Antifungal Glutamine-Rich Hevein-Like Peptides from Vaccaria hispanica.* Front. Plant Sci., 2017. **8**: p. 1100-1113.
58. Jimenez-Barbero, J., et al., *Hevein domains: an attractive model to study carbohydrate–protein interactions at atomic resolution.* Adv. Carbohyd. Chem. Bi., 2006. **60**: p. 303-354.
59. Tam, J.P., et al., *Ginsentides: Cysteine and Glycine-rich Peptides from the Ginseng Family with Unusual Disulfide Connectivity.* Sci. Rep., 2018. **8**(1): p. 16201-16215.
60. Asensio, J.L., et al., *NMR investigations of protein-carbohydrate interactions: Studies on the relevance of Trp/Tyr variations in lectin binding sites as deduced from titration microcalorimetry and NMR studies on hevein domains. Determination of the NMR structure of the complex between pseudohevein and N, N', N "-triacetylchitotriose.* Proteins, 2000. **40**(2): p. 218-236.
61. Andersen, N.H., et al., *Hevein: NMR assignment and assessment of solution-state folding for the agglutinin-toxin motif.* Biochemistry, 1993. **32**(6): p. 1407-1422.
62. Lee, O.S., et al., *Pn-AMPs, the hevein-like proteins from Pharbitis nil confers disease resistance against phytopathogenic fungi in tomato, Lycopersicum esculentum.* Phytochemistry, 2003. **62**(7): p. 1073-1079.
63. Fujimura, M., et al., *Purification, characterization, and sequencing of a novel type of antimicrobial peptides, Fa-AMP1 and Fa-AMP2, from seeds of buckwheat (Fagopyrum esculentum Moench.).* Biosci. Biotech. Bioch., 2003. **67**(8): p. 1636-1642.
64. Huang, R.-H., et al., *Two novel antifungal peptides distinct with a five-disulfide motif from the bark of Eucommia ulmoides Oliv.* FEBS Lett., 2002. **521**(1-3): p. 87-90.

65. Huang, R.-H., et al., *Solution structure of Eucommia antifungal peptide: a novel structural model distinct with a five-disulfide motif.* Biochemistry, 2004. **43**(20): p. 6005-6012.
66. Odintsova, T.I., et al., *A novel antifungal hevein-type peptide from Triticum kiharae seeds with a unique 10-cysteine motif.* FEBS J., 2009. **276**(15): p. 4266-4275.
67. Van den Bergh, K.P., et al., *Five disulfide bridges stabilize a hevein-type antimicrobial peptide from the bark of spindle tree (Euonymus europaeus L.).* FEBS Lett., 2002. **530**(1-3): p. 181-185.
68. Van Parijs, J., et al., *Hevein: an antifungal protein from rubber-tree (Hevea brasiliensis) latex.* Planta, 1991. **183**(2): p. 258-264.
69. Aboitiz, N., et al., *NMR and modeling studies of protein–carbohydrate interactions: synthesis, three-dimensional structure, and recognition properties of a minimum hevein domain with binding affinity for chitooligosaccharides.* ChemBioChem, 2004. **5**(9): p. 1245-1255.
70. Chávez, M.I., et al., *Effect of a serine-to-aspartate replacement on the recognition of chitin oligosaccharides by truncated hevein. A 3D view by using NMR.* Carbohydr. Res., 2010. **345**(10): p. 1461-1468.
71. Malanovic, N. and K. Lohner, *Gram-positive bacterial cell envelopes: The impact on the activity of antimicrobial peptides.* BBA-Biomembranes, 2016. **1858**(5): p. 936-946.
72. McDonald, N.Q. and W.A. Hendrickson, *A structural superfamily of growth factors containing a cystine knot motif.* Cell, 1993. **73**(3): p. 421-424.
73. Rees, D. and W. Lipscomb, *Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2.5 Å resolution.* J. Mol. Biol., 1982. **160**(3): p. 475-498.
74. Craik, D.J., et al., *Plant cyclotides: a unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif.* J. Mol. Biol., 1999. **294**(5): p. 1327-1336.
75. Gruber, C.W., et al., *Distribution and evolution of circular miniproteins in flowering plants.* The Plant Cell, 2008. **20**(9): p. 2471-2483.
76. Poth, A.G., et al., *Cyclotides associate with leaf vasculature and are the products of a novel precursor in petunia (Solanaceae).* J. Biol. Chem., 2012. **287**(32): p. 27033-27046.
77. Nguyen, G.K., 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., 2011. **286**(27): p. 24275-24287.
78. Mylne, J.S., et al., *Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase–mediated biosynthesis.* The Plant Cell, 2012. **24**(7): p. 2765-2778.
79. Nguyen, P.Q.T., et al., *Discovery and characterization of pseudocyclic cystine-knot α-amylase inhibitors with high resistance to heat and proteolytic degradation.* FEBS J., 2014. **281**(19): p. 4351-4366.

80. Nguyen, P.Q., et al., *Antiviral Cystine Knot alpha-Amylase Inhibitors from Alstonia scholaris.* J. Biol. Chem., 2015. **290**(52): p. 31138-31150.
81. Pallaghy, P.K., et al., *A common structural motif incorporating a cystine knot and a triple-stranded β-sheet in toxic and inhibitory polypeptides.* Protein Sci., 1994. **3**(10): p. 1833-1839.
82. Le-Nguyen, D., et al., *Characterization and 2D NMR study of the stable [9–21, 15–27] 2 disulfide intermediate in the folding of the 3 disulfide trypsin inhibitor EETI II.* Protein Sci., 1993. **2**(2): p. 165-174.
83. Heitz, A., D. Le-Nguyen, and L. Chiche, *Min-21 and Min-23, the smallest peptides that fold like a cystine-stabilized β-sheet motif: design, solution structure, and thermal stability.* Biochemistry, 1999. **38**(32): p. 10615-10625.
84. Kolmar, H., *Biological diversity and therapeutic potential of natural and engineered cystine knot miniproteins.* Curr. Opin. Pharmacol., 2009. **9**(5): p. 608-614.
85. Colgrave, M.L. and D.J. Craik, *Thermal, chemical, and enzymatic stability of the cyclotide kalata B1: the importance of the cyclic cystine knot.* Biochemistry, 2004. **43**(20): p. 5965-5975.
86. Ireland, D.C., et al., *Discovery and characterization of a linear cyclotide from Viola odorata: implications for the processing of circular proteins.* J. Mol. Biol., 2006. **357**(5): p. 1522-1535.
87. Chagolla-Lopez, A., et al., *A novel alpha-amylase inhibitor from amaranth (Amaranthus hypocondriacus) seeds.* J. Biol. Chem., 1994. **269**(38): p. 23675-23680.
88. Gao, G.-H., et al., *Solution structure of PAFP-S: a new knottin-type antifungal peptide from the seeds of Phytolacca americana.* Biochemistry, 2001. **40**(37): p. 10973-10978.
89. Hwang, J.-S., et al., *Isolation and characterization of Psacotheasin, a novel Knottin-type antimicrobial peptide, from Psacothea hilaris.* J. Microbiol. Biotechn. , 2010. **20**(4): p. 708-711.
90. Bode, W., et al., *The refined 2.0 Å X-ray crystal structure of the complex formed between bovine β-trypsin and CMTI-I, a trypsin inhibitor from squash seeds (Cucurbita maxima) Topological similarity of the squash seed inhibitors with the carboxypeptidase A inhibitor from potatoes.* FEBS Lett., 1989. **242**(2): p. 285-292.
91. Hernandez, J.-F., et al., *Squash trypsin inhibitors from Momordica cochinchinensis exhibit an atypical macrocyclic structure.* Biochemistry, 2000. **39**(19): p. 5722-5730.
92. Heitz, A., et al., *Solution Structure of the Squash Trypsin Inhibitor MCoTI-II. A New Family for Cyclic Knottins.* Biochemistry, 2001. **40**(27): p. 7973-7983.
93. Ryan, C.A., G.M. Hass, and R.W. Kuhn, *Purification and properties of a carboxypeptidase inhibitor from potatoes.* J. Biol. Chem., 1974. **249**(17): p. 5495-54999.

94. Hass, G.M. and M.A. Hermodson, *Amino acid sequence of a carboxypeptidase inhibitor from tomato fruit.* Biochemistry, 1981. **20**(8): p. 2256-2260.
95. Arolas, J.L., et al., *Secondary binding site of the potato carboxypeptidase inhibitor. Contribution to its structure, folding, and biological properties.* Biochemistry, 2004. **43**(24): p. 7973-7982.
96. Craik, D.J., *Plant cyclotides: circular, knotted peptide toxins.* Toxicon, 2001. **39**(12): p. 1809-1813.
97. Daly, N.L., K.J. Rosengren, and D.J. Craik, *Discovery, structure and biological activities of cyclotides.* Adv. Drug Deliv. Rev., 2009. **61**(11): p. 918-930.
98. Gran, L., *On the Effect of a Polypeptide Isolated from "Kalata-Kalata" (Oldenlandia affinis DC) on the Oestrogen Dominated Uterus.* Acta Pharmacol. Toxicol., 1973. **33**(5-6): p. 400-408.
99. Rosengren, K.J., et al., *Twists, Knots, and Rings in Proteins: STRUCTURAL DEFINITION OF THE CYCLOTIDE FRAMEWORK.* J. Biol. Chem., 2003. **278**(10): p. 8606-8616.
100. Gran, L., K. Sletten, and L. Skjeldal, *Cyclic Peptides from Oldenlandia affinis DC. Molecular and Biological Properties.* Chem. Biodivers., 2008. **5**(10): p. 2014-2022.
101. Wong, C.T., et al., *Optimal oxidative folding of the novel antimicrobial cyclotide from Hedyotis biflora requires high alcohol concentrations.* Biochemistry, 2011. **50**(33): p. 7275-7283.
102. Shenkarev, Z.O., et al., *Conformation and mode of membrane interaction in cyclotides.* FEBS J., 2006. **273**(12): p. 2658-2672.
103. Wang, C.K., et al., *Despite a conserved cystine knot motif, different cyclotides have different membrane binding modes.* Biophys. J., 2009. **97**(5): p. 1471-1481.
104. Duvick, J.P., et al., *Purification and characterization of a novel antimicrobial peptide from maize (Zea mays L.) kernels.* J. Biol. Chem., 1992. **267**(26): p. 18814-18820.
105. Nolde, S.B., et al., *Disulfide-stabilized helical hairpin structure and activity of a novel antifungal peptide EcAMP1 from seeds of barnyard grass (Echinochloa crus-galli).* J. Biol. Chem., 2011. **286**(28): p. 25145-25153.
106. Altschul, S.F., et al., *Basic local alignment search tool.* J. Mol. Biol., 1990. **215**(3): p. 403-410.
107. Mount, D.W., *Using the basic local alignment search tool (BLAST).* Cold Spring Harb. Protoc., **2007**(7): p. 1-17.
108. Mount, D.W., *Strategies for sequence similarity database searches.* Cold Spring Harb. Protoc., **2007**(7): p. 1-15.
109. Tuteja, R., *Type I signal peptidase: An overview.* Arch. Biochem. Biophys., 2005. **441**(2): p. 107-111.
110. von Heijne, G., *Signal sequences: The limits of variation.* J. Mol. Biol., 1985. **184**(1): p. 99-105.

111. Gruber, C.W., et al., *A Novel Plant Protein-disulfide Isomerase Involved in the Oxidative Folding of Cystine Knot Defense Proteins.* J. Biol. Chem., 2007. **282**(28): p. 20435-20446.
112. Ponz, F., et al., *Synthesis and processing of thionin precursors in developing endosperm from barley (Hordeum vulgare L.).* EMBO J., 1983. **2**(7): p. 1035-1040.
113. Lay, F. and M. Anderson, *Defensins-components of the innate immune system in plants.* Curr. Protein Pept. Sci., 2005. **6**(1): p. 85-101.
114. Merrifield, R., *Solid-phase peptide synthesis.* Adv. Enzymol. Relat. Areas Mol. Biol., 2006: p. 221-296.
115. Reinwarth, M., et al., *Chemical synthesis, backbone cyclization and oxidative folding of cystine-knot peptides—promising scaffolds for applications in drug design.* Molecules, 2012. **17**(11): p. 12533-12552.
116. Kadokura, H., *Oxidative protein folding: many different ways to introduce disulfide bonds.* Antioxid. Redox. Redox. Sign., 2005. **8**(5-6): p. 731-733.
117. Anfinsen, C.B. and E. Haber, *Studies on the reduction and re-formation of protein disulfide bonds.* J. Biol. Chem., 1961. **236**(5): p. 1361-1363.
118. Tam, J.P. and C.T. Wong, *Chemical synthesis of circular proteins.* J. Biol. Chem., 2012. **287**(32): p. 27020-27025.
119. Kiefhaber, T., et al., *Protein aggregation in vitro and in vivo: a quantitative model of the kinetic competition between folding and aggregation.* Nat. Biotechnol., 1991. **9**(9): p. 825-829.
120. Rudolph, R., *Renaturation of recombinant, disulfide-bonded proteins from inclusion bodies.* Biotechnol. Adv., 1990: p. 149-172.
121. Ahmed, A.K., S. Schaffer, and D. Wetlaufer, *Nonenzymic reactivation of reduced bovine pancreatic ribonuclease by air oxidation and by glutathione oxidoreduction buffers.* J. Biol. Chem., 1975. **250**(21): p. 8477-8482.
122. Rudolph, R. and H. Lilie, *In vitro folding of inclusion body proteins.* FASEB J., 1996. **10**(1): p. 49-56.
123. Chatrenet, B. and J. Chang, *The disulfide folding pathway of hirudin elucidated by stop/go folding experiments.* J. Biol. Chem., 1993. **268**(28): p. 20988-20996.
124. Arolas, J.L., et al., *Folding of small disulfide-rich proteins: clarifying the puzzle.* Trends Biochem. Sci., 2006. **31**(5): p. 292-301.
125. Akerele, O., *Nature's medicinal bounty: don't throw it away.* 1993. World Health Forum 14, p. 390-395
126. Gurib-Fakim, A., *Medicinal plants: traditions of yesterday and drugs of tomorrow.* Mol. Asp. Med., 2006. **27**(1): p. 1-93.
127. Wong, G., *Biotech scientists bank on big pharma's biologics push.* Nat. Biotechnol., 2009. **27**(3): p. 293.
128. Nguyen, G.K.T., et al., *Butelase-mediated cyclization and ligation of peptides and proteins.* Nat. Protoc., 2016. **11**: p. 1977.

129. Schafmeister, C.E., J. Po, and G.L. Verdine, *An All-Hydrocarbon Cross-Linking System for Enhancing the Helicity and Metabolic Stability of Peptides.* J. Am. Chem. Soc., 2000. **122**(24): p. 5891-5892.
130. Ward, P., et al., *Potent and highly selective neurokinin antagonists.* J. Med. Chem., 1990. **33**(7): p. 1848-1851.
131. Wong, C.T.T., et al., *Orally Active Peptidic Bradykinin B1 Receptor Antagonists Engineered from a Cyclotide Scaffold for Inflammatory Pain Treatment.* Angew. Chem. Int. Ed. Engl., 2012. **51**(23): p. 5620-5624.
132. CARRASCO, L., et al., *Thionins: Plant Peptides that Modify Membrane Permeability in Cultured Mammalian Cells.* Eur. J. Biochem., 1981. **116**(1): p. 185-189.
133. Kramer, K.J., et al., *Toxicity of purothionin and its homologues to the tobacco hornworm, Manduca sexta (L.)(Lepidoptera: Sphingidae).* Toxicol. Appl. Pharm., 1979. **48**(1): p. 179-183.
134. Goldman, M.H., R.B. Goldberg, and C. Mariani, *Female sterile tobacco plants are produced by stigma-specific cell ablation.* EMBO J., 1994. **13**(13): p. 2976-2984.
135. Vita, C., et al., *Scorpion toxins as natural scaffolds for protein engineering.* Proc. Natl. Acad. Sci. U.S.A, 1995. **92**(14): p. 6404-6408.
136. Li, C., et al., *Turning a Scorpion Toxin into an Antitumor Miniprotein.* J. Am. Chem. Soc., 2008. **130**(41): p. 13546-13548.
137. Qiu, Y., et al., *An Orally Active Bradykinin B1 Receptor Antagonist Engineered as a Bifunctional Chimera of Sunflower Trypsin Inhibitor.* J. Med. Chem., 2017. **60**(1): p. 504-510.
138. Ekor, M., *The growing use of herbal medicines: issues relating to adverse reactions and challenges in monitoring safety.* Front. Pharmacol., 2014. **4**(177): p. 1-10.
139. Choudhary, N. and B.S. Sekhon, *An overview of advances in the standardization of herbal drugs.* J. Pharm. Educ. Res., 2011. **2**(2): p. 55.
140. IARC, *Some traditional herbal medicines, some mycotoxins, naphthalene and styrene*. IARC Monogr Eval Carcinog Risks Hum 82: p. 1-556
141. Achike, F.I. and C.-Y. Kwan, *Characterization of a novel tetrandrine-induced contraction in rat tail artery.* Acta Pharmacologica Sinica, 2002. **23**(8): p. 698-704.
142. Debelle, F.D., J.-L. Vanherweghem, and J.L. Nortier, *Aristolochic acid nephropathy: a worldwide problem.* Kidney Int., 2008. **74**(2): p. 158-169.
143. Zhao, S., et al., *Internal transcribed spacer 2 barcode: a good tool for identifying Acanthopanacis cortex.* Front. Plant Sci., 2015. **6**(840): p. 1-7.
144. World Health Organization (WHO), *Quality control methods for herbal materials*. WHO: Geneva, Switzerland, 2011.
145. Folashade, O., H. Omoregie, and P. Ochogu, *Standardization of herbal medicines-A review.* Int. J. Biodivers. Conserv., 2012. **4**(3): p. 101-112.
146. World Health Organization (WHO)., *General guidelines for methodologies on research and evaluation of traditional medicine*. WHO: Geneva, Switzerland, 2000, Geneva.

147. Jiang, Y., et al., *Recent analytical approaches in quality control of traditional Chinese medicines—a review.* Anal. Chim. Acta, 2010. **657**(1): p. 9-18.
148. Xie, P., et al., *Chromatographic fingerprint analysis—a rational approach for quality assessment of traditional Chinese herbal medicine.* J. Chromatogr. A, 2006. **1112**(1-2): p. 171-180.
149. Wong, K.H., et al., *Differentiating Puerariae Lobatae Radix and Puerariae Thomsonii Radix using HPTLC coupled with multivariate classification analyses.* J. Pharmaceut. Biomed., 2014. **95**: p. 11-19.
150. Kasthuri, K., et al., *Development of GC-MS for a polyherbal formulation-MEGNI".* Int. J. Pharm. Sci., 2010. **2**(2): p. 81-83.
151. Jiang, Y., et al., *Optimization of pressurized liquid extraction of five major flavanoids from Lysimachiaclethroide.* J. Pharmaceut. Biomed., 2007. **43**(1): p. 341-345.
152. Chen, J., Y. Song, and P. Li, *Capillary high-performance liquid chromatography with mass spectrometry for simultaneous determination of major flavonoids, iridoid glucosides and saponins in Flos Lonicerae.* J. Chromatogr. A, 2007. **1157**(1-2): p. 217-226.
153. Ma, S., et al., *Off-line comprehensive two-dimensional high-performance liquid chromatography system with size exclusion column and reverse phase column for separation of complex traditional Chinese medicine Qingkailing injection.* J. Chromatogr. A, 2006. **1127**(1-2): p. 207-213.
154. Ganzera, M., *Quality control of herbal medicines by capillary electrophoresis: Potential, requirements and applications.* Electrophoresis, 2008. **29**(17): p. 3489-3503.
155. Hsieh, S.-C., et al., *Determination of aristolochic acid in Chinese herbal medicine by capillary electrophoresis with laser-induced fluorescence detection.* J. Chromatogr. A, 2006. **1105**(1-2): p. 127-134.
156. Lai, Z., P. Xu, and P. Wu, *Multi-steps infrared spectroscopic characterization of the effect of flowering on medicinal value of Cistanche tubulosa.* J. Mol. Struct., 2009. **917**(2-3): p. 84-92.
157. Lu, F., et al., *A new method for testing synthetic drugs adulterated in herbal medicines based on infrared spectroscopy.* Anal. Chim. Acta, 2007. **589**(2): p. 200-207.
158. Lau, C.-C., et al., *Rapid analysis of Radix puerariae by near-infrared spectroscopy.* J. Chromatogr. A, 2009. **1216**(11): p. 2130-2135.
159. Lu, J., et al., *Application of two-dimensional near-infrared correlation spectroscopy to the discrimination of Chinese herbal medicine of different geographic regions.* Spectrochim. ACTA A, 2008. **69**(2): p. 580-586.
160. Bailey, N.J., et al., *Multi-component metabolic classification of commercial feverfew preparations via high-field 1H-NMR spectroscopy and chemometrics.* Planta Medica, 2002. **68**(08): p. 734-738.
161. Pauli, G.F., *qNMR—a versatile concept for the validation of natural product reference compounds.* Phytochem. Anal., 2001. **12**(1): p. 28-42.

162. Yip, P.Y., et al., *DNA methods for identification of Chinese medicinal materials.* Chin. Med., 2007. **2**(1): p. 9.
163. European Medicines Agency, *Note for guidance on quality, of herbal medicinal products*. European Medicines Agency London, London, 2001. p. 6.
164. Chinese Pharmacopoeia Comission, *Pharmacopoeia of the People's Republic of China; China Medical Science Press: Beijing, China.* 2015: p. 232.
165. Food and Drug Administration, *Guidance for industry botanical drug products.* Rockville, MD: Center for Drug Evaluation and Research, 2004. p. 10.
166. Zhou, J., L. Qi, and P. Li, *Quality control of Chinese herbal medicines with chromatographic fingerprint.* Chin. J. Chromatogr., 2008. **26**(2): p. 153-159.
167. Yan, S.-k., et al., *An approach to develop two-dimensional fingerprint for the quality control of Qingkailing injection by high-performance liquid chromatography with diode array detection.* J. Chromatogr. A, 2005. **1090**(1): p. 90-97.
168. Liau, B.-C., et al., *LC-APCI-MS method for detection and analysis of tryptanthrin, indigo, and indirubin in daqingye and banlangen.* J. Pharm. Biomed. Anal., 2007. **43**(1): p. 346-351.
169. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res., 2005. **33**(7): p. 2302-2309.
170. Li, F., et al., *Strategy and Chromatographic Technology of Quality Control for Traditional Chinese Medicines.* Chin. J. Chromatogr., 2006. **24**(6): p. 537-545.
171. Biancolillo, A. and F. Marini, *Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis.* Front. Chem., 2018. **6**(576).
172. Chatfield, C., *Exploratory data analysis.* Eur. J. Oper. Res., 1986. **23**(1): p. 5-13.
173. Martens, H., T. Naes, and T. Naes, *Multivariate calibration*. 1992: John Wiley & Sons.
174. Bautista, R.D., et al., *Simultaneous spectrophotometric determination of drugs in pharmaceutical preparations using multiple linear regression and partial least-squares regression, calibration and prediction methods.* Talanta, 1996. **43**(12): p. 2107-2115.
175. Mazurek, S. and R. Szostak, *Quantitative determination of captopril and prednisolone in tablets by FT-Raman spectroscopy.* J. Pharmaceut. Biomed., 2006. **40**(5): p. 1225-1230.
176. da Silva Fernandes, R., et al., *Non-destructive detection of adulterated tablets of glibenclamide using NIR and solid-phase fluorescence spectroscopy and chemometric methods.* J. Pharm. Biomed. Anal. , 2012. **66**: p. 85-90.

177. de Peinder, P., et al., *Detection of Lipitor® counterfeits: A comparison of NIR and Raman spectroscopy in combination with chemometrics.* J. Pharmaceut. Biomed., 2008. **47**(4): p. 688-694.
178. Murthy, P.S. and M. Madhava Naidu, *Sustainable management of coffee industry by-products and value addition—A review.* Resour. Conserv. Recy., 2012. **66**: p. 45-58.
179. Bisht, S. and S.S. Sisodia, *Coffea arabica: A wonder gift to medical science.* J. Nat. Pharm., 2010. **1**(1): p. 58.
180. Los Santos-Briones, D. and S. Hernández-Sotomayor, *Coffee biotechnology.* Brazilian J. Plant Physiol., 2006. **18**(1): p. 217-227.
181. Barone, J. and H. Roberts, *Caffeine consumption.* Food Chem. Toxicol., 1996. **34**(1): p. 119-129.
182. Namba, T. and T. Matsuse, *A historical study of coffee in Japanese and Asian countries: focusing the medicinal uses in Asian traditional medicines.* J. Jpn. Hist. Pharm., 2001. **37**(1): p. 65-75.
183. Dogasaki, C., et al., *Identification of chemical structure of antibacterial components against Legionella pneumophila in a coffee beverage.* J. Jpn. Hist. Pharm., 2002. **122**(7): p. 487-494.
184. Ochiai, R., et al., *Green coffee bean extract improves human vasoreactivity.* Hypertens. Res., 2004. **27**(10): p. 731-737.
185. Yukawa, G., et al., *Effects of coffee consumption on oxidative susceptibility of low-density lipoproteins and serum lipid levels in humans.* Biochemistry, 2004. **69**(1): p. 70-74.
186. Naismith, D., et al., *The effect in volunteers of coffee and decaffeinated coffee on blood glucose, insulin, plasma lipids and some factors involved in blood clotting.* Ann. Nutr. Metab., 1970. **12**(3): p. 144-151.
187. Salazar-Martinez, E., et al., *Coffee consumption and risk for type 2 diabetes mellitus.* Ann. Intern. Med., 2004. **140**(1): p. 1-8.
188. Dórea, J.G. and T.H.M. da Costa, *Is coffee a functional food?* Brit. J. Nutr., 2005. **93**(06): p. 773-782.
189. Parliment, T. and H.D. Stahl, *What makes that coffee smell so good?* Chemtech, 1995. **25**(8): p. 38-47.
190. Spiller, M.A., *The chemical components of coffee.* Caffeine, 1998. **1998**: p. 97-161.
191. James, J.E., *Critical review of dietary caffeine and blood pressure: a relationship that should be taken more seriously.* Psychosom. Med., 2004. **66**(1): p. 63-71.
192. Carrillo, J.A. and J. Benitez, *Clinically significant pharmacokinetic interactions between dietary caffeine and medications.* Clin. Pharmacokinet., 2000. **39**(2): p. 127-153.
193. Bonita, J.S., et al., *Coffee and cardiovascular disease: in vitro, cellular, animal, and human studies.* Pharmacol. Res., 2007. **55**(3): p. 187-198.
194. Rustan, A.C., et al., *Effect of coffee lipids (cafestol and kahweol) on regulation of cholesterol metabolism in HepG2 cells.* Arter. Thromb. Vasc. Biol., 1997. **17**(10): p. 2140-2149.

195. Clifford, M.N., *Chlorogenic acids and other cinnamates–nature, occurrence and dietary burden[1].* J. Sci. Food Agric., 1999. **79**: p. 362-372.
196. Ohnishi, M., et al., *Inhibitory effects of chlorogenic acids on linoleic acid peroxidation and haemolysis.* phytochemistry, 1994. **36**(3): p. 579-583.
197. Fu, J., et al., *Review of the botanical characteristics, phytochemistry, and pharmacology of Astragalus membranaceus (Huangqi).* Phytother. Res., 2014. **28**(9): p. 1275-1283.
198. Kim, C., et al., *Induction of growth hormone by the roots ofAstragalus membranaceus in pituitary cell culture.* Arch. Pharm. Res., 2003. **26**(1): p. 34-39.
199. Kai, Z., et al., *Biological active ingredients of traditional Chinese herb Astragalus membranaceus on treatment of diabetes: a systematic review.* Mini-rev. Med. Chem., 2015. **15**(4): p. 315-329.
200. Ma, X.Q., et al., *Chemical analysis of Radix Astragali (Huangqi) in China: a comparison with its adulterants and seasonal variations.* J. Agr. Food. Chem., 2002. **50**(17): p. 4861-4866.
201. Yu, Q.T., et al., *Two new saponins from the aerial part of Astragalus membranaceus var. mongholicus.* Chin. Chem. Lett., 2007. **18**(5): p. 554-556.
202. Li-Man, S., et al., *Chemical constituents from Astragalus ernestii.* Chin. J. Nat. Medicines, 2011. **9**(1): p. 38-41.
203. Jin, M., et al., *Structural features and biological activities of the polysaccharides from Astragalus membranaceus.* Int. J. Biol. Marcromol., 2014. **64**: p. 257-266.
204. Yin, F., et al., *Dietary supplementation with Astragalus polysaccharide enhances ileal digestibilities and serum concentrations of amino acids in early weaned piglets.* Amino Acids, 2009. **37**(2): p. 263-270.
205. Choi, H.-S., et al., *Quality characteristic of hwangki (Astragalus membranaceus) chungkukjang during fermentation.* Korean J. Food Preserv., 2007. **14**(4): p. 356-363.
206. Nalbantsoy, A., et al., *Evaluation of the immunomodulatory properties in mice and in vitro anti-inflammatory activity of cycloartane type saponins from Astragalus species.* J. Ethnopharmacol., 2012. **139**(2): p. 574-581.
207. Hongfeng, L., *The effect of APS to FPG and blood lipids on insulin resistance of rats with type 2 biabetes.* Journal of Mudanjiang Medical College, 2007. **5**.
208. Yu, J., et al., *Inhibitory effects of astragaloside IV on diabetic peripheral neuropathy in rats.* Can. J. Physol. Pharm., 2006. **84**(6): p. 579-587.
209. Li, R., et al., *Chemical constituents of Astragalus membranaceus Bge. var. mongholicus (Bge) Hsiao.* Journal of Shenyang Pharmaceutical University, 2007. **1**.
210. Jiangwei, M., Q. Zengyong, and X. Xia, *Aqueous extract of Astragalus mongholicus ameliorates high cholesterol diet induced oxidative injury in experimental rats models.* J. Med. Plant Res., 2011. **5**(5): p. 855-858.

211. Gasteiger, E., et al., *ExPASy: the proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res., 2003. **31**(13): p. 3784-3788.
212. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat. Methods, 2011. **8**(10): p. 785-786.
213. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res., 2004. **14**(6): p. 1188-1190.
214. Roy, A., A. Kucukural, and Y. Zhang, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nat. Protoc., 2010. **5**(4): p. 725-738.
215. DeLano, W.L., *Pymol: An open-source molecular graphics tool.* CCP4 Newsletter On Protein Crystallography, 2002. **40**: p. 82-92.
216. Jeener, J., et al., *Investigation of exchange processes by two-dimensional NMR spectroscopy.* J. Chem. Phys., 1979. **71**(11): p. 4546-4553.
217. Kumar, A., R. Ernst, and K. Wüthrich, *A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules.* Biochem. Bioph. Res. Co., 1980. **95**(1): p. 1-6.
218. Bax, A. and D.G. Davis, *MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy.* J. Magn. Reson., 1985. **65**(2): p. 355-360.
219. Delaglio, F., et al., *NMRPipe: a multidimensional spectral processing system based on UNIX pipes.* J. Biomol. NMR, 1995. **6**(3): p. 277-293.
220. Wüthrich, K., M. Billeter, and W. Braun, *Pseudo-structures for the 20 common amino acids for use in studies of protein conformations by measurements of intramolecular proton-proton distance constraints with nuclear magnetic resonance.* J. Mol. Biol., 1983. **169**(4): p. 949-961.
221. Lee, W., M. Tonelli, and J.L. Markley, *NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy.* Bioinformatics, 2014. **31**(8): p. 1325-1327.
222. Brünger, A.T., et al., *Crystallography & NMR system: A new software suite for macromolecular structure determination.* Acta Crystallographica Section D, 1998. **54**(5): p. 905-921.
223. Laskowski, R.A., et al., *AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR.* J. Biomol. NMR, 1996. **8**(4): p. 477-486.
224. Ye, X. and T. Ng, *A new antifungal peptide from rice beans.* Chem. Biol. Drug. Des., 2002. **60**(2): p. 81-87.
225. Wiegand, I., K. Hilpert, and R.E. Hancock, *Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances.* Nat. Protoc., 2008. **3**(2): p. 163-175.
226. Gebäck, T., et al., *TScratch: a novel and simple software tool for automated analysis of monolayer wound healing assays: Short Technical Reports.* Biotechniques, 2009. **46**(4): p. 265-274.

227. Lo, M.-C., et al., *Evaluation of fluorescence-based thermal shift assays for hit identification in drug discovery.* Anal. Biochem., 2004. **332**(1): p. 153-159.
228. Whitmore, L. and B.A. Wallace, *DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data.* Nucleic Acids Res., 2004. **32**(WEB SERVER ISS.): p. W668-W673.
229. Boguski, M.S., T.M. Lowe, and C.M. Tolstoshev, *dbEST—database for "expressed sequence tags".* Nat. Genet., 1993. **4**(4): p. 332.
230. Matasci, N., et al., *Data access for the 1,000 Plants (1KP) project.* Gigascience, 2014. **3**(1): p. 17.
231. Hall, T., *BioEdit: an important software for molecular biology.* GERF Bull. Biosci., 2011. **2**(1): p. 60-61.
232. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European molecular biology open software suite.* Trends Genet., 2000. **16**(6): p. 276-277.
233. Tamura, K., et al., *MEGA6: molecular evolutionary genetics analysis version 6.0.* Mol. Biol. Evol., 2013. **30**(12): p. 2725-2729.
234. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.* Nucleic Acids Res., 2016. **44**(W1): p. W242-W245.
235. Skov, T., et al., *Automated alignment of chromatographic data.* J. Chemom., 2006. **20**(11-12): p. 484-497.
236. Lever, J., M. Krzywinski, and N. Altman, *Points of significance: Principal component analysis*. 2017, Nature Publishing Group.
237. Rokach, L. and O. Maimon, *Clustering methods*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 321-352.
238. Chlebda, D.K., et al., *Hyperspectral imaging coupled with chemometric analysis for non-invasive differentiation of black pens.* Applied Physics A, 2016. **122**(11): p. 957.
239. Daszykowski, M., B. Walczak, and D. Massart, *Representative subset selection.* Anal. Chim. Acta, 2002. **468**(1): p. 91-103.
240. Rosipal, R. and N. Krämer. *Overview and recent advances in partial least squares*. in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*. 2005. Springer.
241. Wong, K.H., et al., *Differentiation of Pueraria lobata and Pueraria thomsonii using partial least square discriminant analysis (PLS-DA).* J. Pharmaceut. Biomed., 2013. **84**: p. 5-13.
242. Zadeh, L.A., *Fuzzy sets.* Inf. Control, 1965. **8**(3): p. 338-353.
243. Breiman, L., *Classification and regression trees*. 2017: Routledge.
244. Frank, I.E. and S. Lanteri, *Classification models: discriminant analysis, SIMCA, CART.* Chemometr. Intell. Lab., 1989. **5**(3): p. 247-256.
245. Luna, A.S., et al., *Rapid characterization of transgenic and non-transgenic soybean oils by chemometric methods using NIR spectroscopy.* Spectrochim. ACTA A, 2013. **100**: p. 115-119.

246. Vapnik, V., *The nature of statistical learning theory*. 2013: Springer science & business media.
247. Menesatti, P., et al., *Estimation of plant nutritional status by Vis–NIR spectrophotometric analysis on orange leaves [Citrus sinensis (L) Osbeck cv Tarocco].* Biosyst. Eng., 2010. **105**(4): p. 448-454.
248. Gad, H.A., et al., *Application of chemometrics in authentication of herbal medicines: a review.* Phytochem. Anal., 2013. **24**(1): p. 1-24.
249. Pradhan, V., et al., *A overview of species identification by DNA barcoding.* Int. J. Curr. Microbiol. App. Sci., 2015. **4**(4): p. 127-140.
250. Liang, Y., P. Xie, and F. Chau, *Chromatographic fingerprinting and related chemometric techniques for quality control of traditional Chinese medicines.* J. Sep. Sci., 2010. **33**(3): p. 410-421.
251. Wong, K.H., et al., *The quality control of two Pueraria species using Raman spectroscopy coupled with partial least squares analysis.* J. Raman Spectrosc., 2015. **46**(4): p. 361-368.
252. Ho, C.S., et al., *Electrospray ionisation mass spectrometry: principles and clinical applications.* Clin. Biochem. Rev., 2003. **24**(1): p. 3.
253. Jackson, P.E., P.F. Scholl, and J.D. Groopman, *Mass spectrometry for genotyping: an emerging tool for molecular medicine.* Mol. Med. Today, 2000. **6**(7): p. 271-276.
254. El-Aneed, A., A. Cohen, and J. Banoub, *Mass spectrometry, review of the basics: electrospray, MALDI, and commonly used mass analyzers.* Appl. Spectrosc. Rev., 2009. **44**(3): p. 210-230.
255. Cai, Z. and S. Liu, *Applications of MALDI-TOF spectroscopy*. Vol. 331. 2014: Springer.
256. Packeu, A., et al., *Fast and accurate identification of dermatophytes by matrix-assisted laser desorption ionization–time of flight mass spectrometry: validation in the clinical laboratory.* J. Clin. Microbiol., 2014. **52**(9): p. 3440-3443.
257. Fraige, K., E.R. Pereira-Filho, and E. Carrilho, *Fingerprinting of anthocyanins from grapes produced in Brazil using HPLC–DAD–MS and exploratory analysis by principal component analysis.* Food Chem., 2014. **145**: p. 395-403.
258. Chambery, A., et al., *Peptide fingerprint of high quality Campania white wines by MALDI-TOF mass spectrometry.* Food Chem., 2009. **113**(4): p. 1283-1289.
259. Cho, W.C.S. and K.N. Leung, *In vitro and in vivo immunomodulating and immunorestorative effects of Astragalus membranaceus.* J. Ethnopharmacol., 2007. **113**(1): p. 132-141.
260. Liu, Y., et al., *Comparative chemical analysis of Radix Astragali and Radix Hedysari by HPLC.* Nat. Prod. Res., 2012. **26**(20): p. 1935-1938.
261. Liu, J., et al., *Comparison of the immunoregulatory function of different constituents in radix astragali and radix hedysari.* J. Biomed. Biotechnol., 2010. **2010**: p. 479426-479437.

262. Song, Q.-H., et al., *Effects of Astragali root and Hedysari root on the murine B and T cell differentiation.* J. Ethnopharmacol. , 2000. **73**(1-2): p. 111-119.
263. Lee, I.-J., et al., *Investigation of Two Species of Huang-qi (Aatragalus membranaceus and Hedysarum polybotrys) by HPLC, ITS, Microscopic Morphology and Antioxidant Activities.* J. Food. Drug. Anal., 2012. **20**(3): p. 603-610.
264. Xiao, W., L. Han, and B. Shi, *Isolation and purification of flavonoid glucosides from Radix Astragali by high-speed counter-current chromatography.* J. Chromatogr. B, 2009. **877**(8-9): p. 697-702.
265. Ma, X., et al., *Species identification of Radix Astragali (Huangqi) by DNA sequence of its 5S-rRNA spacer domain.* Phytochemistry, 2000. **54**(4): p. 363-368.
266. Tan, W.L., et al., *Lybatides from Lycium barbarum Contain An Unusual Cystine-stapled Helical Peptide Scaffold.* Sci. Rep., 2017. **7**(1): p. 5194-5204.
267. Nguyen, G.K., et al., *Discovery of a linear cyclotide from the bracelet subfamily and its disulfide mapping by top-down mass spectrometry.* J. Biol. Chem., 2011. **286**(52): p. 44833-44844.
268. Louis, S., et al., *Broad screening of the legume family for variability in seed insecticidal activities and for the occurrence of the A1b-like knottin peptide entomotoxins.* Phytochemistry, 2007. **68**(4): p. 521-535.
269. hui, Z.R.G.w.s.b.y.d.w.y., *Pharmacopoeia of the People's Republic of China*. Vol. 1. 2000: Chemical Industry Press.
270. Liu, Y., et al., *Comparative chemical analysis of Radix Astragali and Radix Hedysari by HPLC.* Nat. Prod. Res., 2012. **26**(20): p. 1935-8.
271. Viapiana, A., et al., *An approach based on HPLC-fingerprint and chemometrics to quality consistency evaluation of Matricaria chamomilla L. commercial samples.* Front. Plant Sci., 2016. **7**: p. 1561.
272. Varmuza, K. and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*. 2016: CRC press.
273. Barnes, R., M.S. Dhanoa, and S.J. Lister, *Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra.* Applied spectroscopy, 1989. **43**(5): p. 772-777.
274. Liu, J., et al., *Comparison of the immunoregulatory function of different constituents in radix astragali and radix hedysari.* Journal of Biomedicine and Biotechnology, 2010. **2010**: p. 479426-479437.
275. Martins, L.R.R., E.R. Pereira-Filho, and Q.B. Cass, *Chromatographic profiles of Phyllanthus aqueous extracts samples: a proposition of classification using chemometric models.* Anal. Bioanal. Chem., 2011. **400**(2): p. 469-481.
276. Schmidt, B.M., et al., *Revisiting the ancient concept of botanical therapeutics.* Nat. Chem. Biol., 2007. **3**(7): p. 360-366.

277. Qi, S., et al., *Lychee (Litchi chinensis Sonn.) seed water extract as potential antioxidant and anti-obese natural additive in meat products.* Food Control, 2015. **50**: p. 195-201.
278. Higgins, T., et al., *Gene structure, protein structure, and regulation of the synthesis of a sulfur-rich protein in pea seeds.* J. Biol. Chem., 1986. **261**(24): p. 11124-11130.
279. Da Silva, P., et al., *A folded and functional synthetic PA1b: an interlocked entomotoxic miniprotein.* Biopolymers, 2009. **92**(5): p. 436-444.
280. Da Silva, P., et al., *Molecular requirements for the insecticidal activity of the plant peptide pea albumin 1 subunit b (PA1b).* J. Biol. Chem. , 2010. **285**(43): p. 32689-32694.
281. Gressent, F., et al., *Pea Albumin 1 subunit b (PA1b), a promising bioinsecticide of plant origin.* Toxins, 2011. **3**(12): p. 1502-1517.
282. Eyraud, V., et al., *The interaction of the bioinsecticide PA1b (Pea Albumin 1 subunit b) with the insect V-ATPase triggers apoptosis.* Sci. Rep., 2017. **7**(1): p. 4902-4911.
283. Yamazaki, T., et al., *A possible physiological function and the tertiary structure of a 4-kDa peptide in legumes.* FEBS J., 2003. **270**(6): p. 1269-1276.
284. Dun, X.P., et al., *The effect of pea albumin 1F on glucose metabolism in mice.* Peptides, 2008. **29**(6): p. 891-897.
285. Dun, X.P., et al., *Activity of the plant peptide aglycin in mammalian systems.* FEBS J., 2007. **274**(3): p. 751-759.
286. Lu, J., et al., *The soybean peptide aglycin regulates glucose homeostasis in type 2 diabetic mice via IR/IRS1 pathway.* J. Nutr. Biochem., 2012. **23**(11): p. 1449-1457.
287. Shafee, T.M., et al., *Convergent evolution of defensin sequence, structure and function.* Cell. Mol. Life Sci., 2017. **74**(4): p. 663-682.
288. Broekaert, W.F., et al., *Plant defensins: novel antimicrobial peptides as components of the host defense system.* Plant Physiol., 1995. **108**(4): p. 1353-1358.
289. Rios, J. and P. Waterman, *A review of the pharmacology and toxicology of Astragalus.* Phytother. Res., 1997. **11**(6): p. 411-418.
290. Wang, Z., J. Wang, and P. Chan, *Treating type 2 diabetes mellitus with traditional chinese and Indian medicinal herbs.* J. Evidence-Based Complementary Altern. Med., 2013. **2013**: p. 343594-343611.
291. Agyemang, K., et al., *Recent Advances in Astragalus membranaceus Anti-Diabetic Research: Pharmacological Effects of Its Phytochemical Constituents.* J. Evidence-Based Complementary Altern. Med., 2013. **2013**: p. 654643-654652.
292. Louis, S., et al., *Molecular and biological screening for insect-toxic seed albumins from four legume species.* Plant Sci., 2004. **167**(4): p. 705-714.
293. Watanabe, Y., et al., *A peptide that stimulates phosphorylation of the plant insulin-binding protein: isolation, primary structure and cDNA cloning.* Eur. J .Biochem., 1994. **224**(1): p. 167-172.

294. Jianzhong, T., *Analysis of leginsulin gene in soybean cultivar (Glycine max) and wild species (Glycine Soja).* Chinese Journal of Applied and Environmental Biology, 1999. **5**(3): p. 259-263.
295. Taylor, W.G., et al., *Sequence determination by MALDI-TOF mass spectrometry of an insecticidal lentil peptide of the PA1b type.* Phytochem. Lett., 2015. **12**: p. 105-112.
296. Meindre, F., et al., *The nuclear magnetic resonance solution structure of the synthetic AhPDF1. 1b plant defensin evidences the structural feature within the γ-motif.* Biochemistry, 2014. **53**(49): p. 7745-7754.
297. Zhang, Y. and K. Lewis, *Fabatins: new antimicrobial plant peptides.* FEMS Microbiol. Lett., 1997. **149**(1): p. 59-64.
298. Lay, F.T., et al., *The three-dimensional solution structure of NaD1, a new floral defensin from Nicotiana alata and its application to a homology model of the crop defense protein alfAFP.* J. Mol. Biol., 2003. **325**(1): p. 175-188.
299. Chen, J.-J., et al., *Cloning and functional expression of a mungbean defensin VrD1 in Pichia pastoris.* J. Agric. Food Chem., 2004. **52**(8): p. 2256-2261.
300. Finkina, E.I., et al., *A novel defensin from the lentil Lens culinaris seeds.* Biochem. Biophys. Res. Commun., 2008. **371**(4): p. 860-865.
301. Kvansakul, M., et al., *Binding of phosphatidic acid by NsD7 mediates the formation of helical defensin–lipid oligomeric assemblies and membrane permeabilization.* Proc. Natl. Acad. Sci. U. S. A, 2016. **113**(40): p. 11202-11207.
302. Wong, K.H., et al., *Ginkgotides: proline-rich hevein-like peptides from gymnosperm Ginkgo biloba.* Front. Plant Sci., 2016. **7**: p. 1639.
303. Hellinger, R., et al., *Inhibition of human prolyl oligopeptidase activity by the cyclotide psysol 2 isolated from Psychotria solitudinum.* J. Nat. Prod., 2015. **78**(5): p. 1073-1082.
304. Männistö, P.T., et al., *Prolyl oligopeptidase: a potential target for the treatment of cognitive disorders.* Drug News Perspect., 2007. **20**(5): p. 293-305.
305. Eyraud, V., et al., *Expression and biological activity of the cystine knot bioinsecticide PA1b (Pea Albumin 1 Subunit b).* PloS one, 2013. **8**(12): p. e81619.
306. Nguyen, G.K., et al., *Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis.* Nat. Chem. Biol., 2014. **10**(9): p. 732-738.
307. Adel, E.-G., et al., *Classification of the Leguminosae-Papilionoideae: A Numerical Re-assessment.* Nat. Sci. Biol., 2013. **5**(4): p. 499-507.
308. Doyle, J.J. and M.A. Luckow, *The rest of the iceberg. Legume diversity and evolution in a phylogenetic context.* Plant Physiol., 2003. **131**(3): p. 900-910.
309. Jennings, C., et al., *Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from Oldenlandia affinis.* Proc. Natl. Acad. Sci. U. S. A, 2001. **98**(19): p. 10614-10619.

310. Daly, N.L. and D.J. Craik, *Bioactive cystine knot proteins.* Curr. Opin. Chem. Biol., 2011. **15**(3): p. 362-368.
311. Briscoe, C.P., et al., *Pharmacological regulation of insulin secretion in MIN6 cells through the fatty acid receptor GPR40: identification of agonist and antagonist small molecules.* Br. J. Pharmacol., 2006. **148**(5): p. 619-628.
312. Leahy, J., *Beta-cell dysfunction with chronic hyperglycemia:" overworked beta-cell" hypothesis.* Diabetes Rev., 1996. **4**: p. 298-319.
313. Lamontagne, J., et al., *Pioglitazone acutely reduces energy metabolism and insulin secretion in rats.* Diabetes, 2013: p. DB_120428.
314. Nicolas, P., *Multifunctional host defense peptides: intracellular-targeting antimicrobial peptides.* FEBS J., 2009. **276**(22): p. 6483-6496.
315. Vriens, K., B. Cammue, and K. Thevissen, *Antifungal plant defensins: mechanisms of action and production.* Molecules, 2014. **19**(8): p. 12280-12303.
316. dos Passos Assis, B., et al., *Growth Response of Four Conilon Coffee Varieties (Coffea canephora Pierre ex A. Froehner) to Different Shading Levels.* J. Agr. Sci., 2019. **11**(7).
317. Manach, C., et al., *Polyphenols: food sources and bioavailability.* Am. J. Clin. Nutr., 2004. **79**(5): p. 727-747.
318. Olthof, M.R., P.C. Hollman, and M.B. Katan, *Chlorogenic acid and caffeic acid are absorbed in humans.* J. Nutr. , 2001. **131**(1): p. 66-71.
319. Geraets, L., et al., *Caffeine metabolites are inhibitors of the nuclear enzyme poly (ADP-ribose) polymerase-1 at physiological concentrations.* Biochem. Pharmacol., 2006. **72**(7): p. 902-910.
320. Mori, H., et al., *Inhibitory effect of chlorogenic acid on methylazoxymethanol acetate-induced carcinogenesis in large intestine and liver of hamsters.* Cancer Lett., 1986. **30**(1): p. 49-54.
321. Lin, L., et al., *Maca (Lepidium meyenii) as a source of macamides and polysaccharide in combating of oxidative stress and damage in human erythrocytes.* Int. J. Food Sci. Tech., 2017.
322. Heitz, A., et al., *Knottin cyclization: impact on structure and dynamics.* BMC Struct. Biol., 2008. **8**(1): p. 54.
323. Tam, J.P., et al., *An unusual structural motif of antimicrobial peptides containing end-to-end macrocycle and cystine-knot disulfides.* Proc. Natl. Acad. Sci. U.S.A, 1999. **96**(16): p. 8913-8918.
324. Olli, S. and P. Kirti, *Cloning, characterization and antifungal activity of defensin Tfgd1 from Trigonella foenum-graecum L.* BMB Rep., 2006. **39**(3): p. 278-283.
325. Lin, C., et al., *Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts.* Theor. Appl. Genet., 2005. **112**(1): p. 114-130.
326. Gongora, C., et al. *Differential gene expression response of Coffee arabica and C. liberica to coffee berry borer attack*. in *22nd International*

*Conference on Coffee Science, ASIC 2008, Campinas, SP, Brazil, 14-19 September, 2008*.
327. Vidal, R.O., et al., *A high-throughput data mining of single nucleotide polymorphisms in Coffea species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid Coffea arabica.* Plant Physiol., 2010. **154**(3): p. 1053-1066.
328. Vieira, L.G.E., et al., *Brazilian coffee genome project: an EST-based genomic resource.* Brazilian J. Plant Physiol., 2006. **18**(1): p. 95-108.
329. Wilcox, D.E., *Isothermal titration calorimetry of metal ions binding to proteins: An overview of recent studies.* Inorganica Chim. Acta, 2008. **361**(4): p. 857-867.
330. Ebrahim-Nesbat, F., et al., *Cultivar-related differences in the distribution of cell-wall-bound thionins in compatible and incompatible interactions between barley and powdery mildew.* Planta, 1989. **179**(2): p. 203-210.
331. Evans, J., et al., *Cellular responses to Pyrularia thionin are mediated by Ca2+ influx and phospholipase A2 activation and are inhibited by thionin tyrosine iodination.* Proc. Natl. Acad. Sci. U. S. A, 1989. **86**(15): p. 5849-5853.
332. Twardzik, D.R. and A. Peterkofsky, *Glutamic acid as a precursor to N-terminal pyroglutamic acid in mouse plasmacytoma protein.* Proc. Natl. Acad. Sci. U.S.A, 1972. **69**(1): p. 274-277.
333. Del Castillo, M., et al., *Coffee by-products*, in *Coffee*. 2019. p. 309-334.
334. Jiang, H., et al., *Oral administration of soybean peptide Vglycin normalizes fasting glucose and restores impaired pancreatic function in Type 2 diabetic Wistar rats.* J. Nutr. Biochem., 2014. **25**(9): p. 954-963.
335. Kam, A., et al., *Plant-derived mitochondria-targeting cysteine-rich peptide modulates cellular bioenergetics.* J. Biol. Chem., 2019. **294**(11): p. 4000-4011.
336. Hirano, H., *Structures and Functions of Leginsulin and Leginsulin-Binding Protein*, in *Protein Structure — Function Relationship*, Z.H. Zaidi and D.L. Smith, Editors. 1996, Springer US: Boston, MA. p. 91-96.
337. Eyraud, V., et al., *Expression and biological activity of the cystine knot bioinsecticide PA1b (Pea Albumin 1 Subunit b).* PloS one, 2013. **8**(12): p. e81619-e81627.