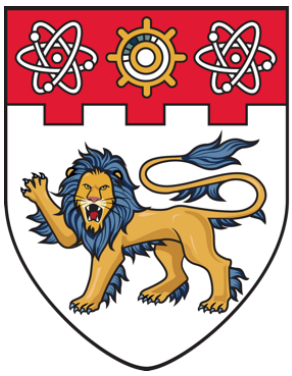


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Discovery and Characterization of Ligase Activity Determinants in Plant Asparaginyl Endopeptidases

CHAN Ning-Yu

SCHOOL OF BIOLOGICAL SCIENCES

2021

Discovery and Characterization of Ligase Activity Determinants in Plant Asparaginyl Endopeptidases

CHAN Ning-Yu

SCHOOL OF BIOLOGICAL SCIENCES

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

2021

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 10th 2021

. . . . . .

Date

. . . . . .

Ning-Yu Chan

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.

[Input Date Here]

Aug 10th 2021

Date

Professor James P. Tam

Authorship Attribution Statement

This thesis contains material from one paper published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 1 Section 3 is published as Tam, JP., Chan, N-Y, Liew, HT, Tan, SJ, Chen, Y, *Peptide asparaginyl ligases—renegade peptide bond makers.* Sci. China Chem, 2020. **63**: p. 296-307.

The contributions of the co-authors are as follows:

- Professor James P. Tam provided the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by Professor James P. Tam.
- Mr. Heng-Tai Liew wrote Section 2.4 Modified Subtilisin. The section was revised by Professor James P. Tam.
- Mr. Shaun Jun Hao Tan and Ms. Yu Chen wrote Section 2.3 Transpeptidation by Sortase A. The section was revised by Professor James P. Tam.
- All figures, tables, and sections except 2.3-2.4 were prepared by and revised by Professor James P. Tam.

Aug 10th 2021

. . . . . .

Date

Ning-Yu Chan

Acknowledgment

First and foremost, I would like to express my sincere gratitude and deepest appreciation to my supervisor Professor James P. Tam, for his strong support and meticulous guidance. The immense knowledge, invaluable scientific ideas, and passion for the science of my mentor Professor James P. Tam have been extremely inspiring, motivating, and life-changing during my Ph.D. studies. It is my greatest pleasure to be Professor James P. Tam's student, and this experience will continuously spur me to become better.

My appreciation also goes to the Thesis Advisory Committee, Professor Chuan-Fa Liu and Professor Oliver Mueller-Cajar, for their helpful advice, practical suggestions, and warm encouragement. I would also like to thank Professor Julien Lescar and Professor Siu Kwan Sze for their technical supports and constructive advice during this project.

I very much appreciate the help from the wonderful members and collaborators of Professor James P. Tam's group. They are Professor Xupeng Li, Professor Zhipei Sang, Dr. Xiaobao Bi, Dr. Bamaprasad Dutta, Dr. Wenjun He, Dr. Xinya Hemu, Dr. Mei Huang, Dr. Jiayi Huang, Dr. Antony Kam, Dr. Warren Liang, Dr. Shining Loo, Dr. Chris Looi, Dr. Yanling Ng, Dr. Giang Nguyen, Dr. Abbas El Sahili, Dr. Aida Serra, Dr. Yuping Shen, Dr. Weiliang Tan, Dr. Stephanie Victoria Tay, Dr. Janet To, Dr. Kaho Wong, Dr. Xiaohong Zhang, Mr. Side Hu, Ms. Yu Chen, Mr. Heng-Tai Liew, Mr. Dingpeng Zhang, Mr. Shuan Tan, Ms. Yeehwa Wong, Ms, Niying Chua, Mr. Ching Koon Lim, Mr. Dickson Sng, Mr. Fan Tang, Ms. Zhen Wang, Ms. Yiyin Xia, Ms. Weiqin Lim, Mr. Kar Jun Loh, Ms. Iris Tham, Mr. Kwan Ann Tan, Ms. Lee Choo Yee, Ms. Yalei Xu, Ms. Petrina Kong, and Ms. Hui Ying Yee. My special thanks go

to Dr. Xinya Hemu, Dr. Kaho Wong, and Mr. Heng-Tai Liew, who are particularly helpful and supportive during my Ph.D.

I am extremely grateful to my friend, Chia Mei Lin, a talented artist, for her warmest supports and love. Thanks also to all my supportive and helpful friends, Aidana, Alice, Christina, Patrick, Jeff, Jiawen, Kadiam, Oishi, Shikhar, Soheil, and Zarina of NTU SBS, as well as Nagyung, Yue Shu, Shuli, Howard, Zihan, Carol, Jiayu, and Kaji of NTU.

I would like to express my deepest gratitude to my wonderful family. I thank my Godfather Albert Wang and Wang's family for their kind support and wise suggestions. I very much appreciate my fiancé William Lloyd Vaughn and the Vaughns for the joys and love they brought into my life. Thanks also to my mothers, Shu-Ying Chen and Chiao-Yu Chiu, as well as Chen's family and Chiu's family for their warm supports and love. Lastly, I would like to give my special thanks and deepest gratitude to my father, Chih-Hui Chan, who always has a profound belief in my abilities and makes me who I am.

This work was supported by Academic Research Grant Tier 3 (MOE2016-T3-1-003) from the Singapore Ministry of Education.

Table of Contents

List of Figures

List of Tables

Abbreviations

Standard abbreviations are used for the amino acids and protecting groups [IUPAC-IUB Commission for Biochemical Nomenclature.

ACN	Acetonitrile
AEP	Asparaginyl endopeptidase
AP	Activation peptide
BLAST	Basic local alignment search tool
β-ME	β-mercaptoethanol
BrEA	Bromoethylamine
$CaCl_2$	Calcium chloride
Cy	Cycloviolacin
DARPins	Designed ankyrin repeat protein
DMSO	Dimethylsulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiolthreitol
ECL	Enhanced chemiluminescence
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
ET	Evolutionary trace
Extein	*Ex*ternal pro*tein*
FMOC	9-fluroenylmethyloxycarbonyl
FPLC	Fast protein liquid chromatography
FRET	Fluorescence resonance energy transfer
GFP	Green fluorescence protein
GmAMA1	α-amanitin precursor

HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPLC	High-performance liquid chromatography
IMAC	Immobilized metal affinity chromatography
Intein	*In*tervening pro*tein*
IPTG	Isopropyl β-D-1-thiogalactopyranoside
I-TASSER	Iterative Threading ASSEmbly Refinement
kB1	Kalata B1
kDa	Kilo Dalton
LB	Luria-Bertani broth
LC	Liquid chromatography
LC-MS	Liquid chromatography mass spectrometry
LSAM	Legumain stabilization and activity modulation
MALDI	Matrix-assisted laser desorption ionization
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MW	Molecular weight
NaCl	Sodium chloride
NCBI	National Center for Biotechnology Information
NCL	Native chemical ligation
OD	Optical density
OLP	Oligopeptidase
OneKP	1000 Plant Transcriptome Database
ORF	Open reading frame
PAL	Peptide asparaginyl ligase
PatEdm	PatE precursor

PCD	Programmed cell death
PCR	Polymerase chain reaction
PCY1	Peptide cyclase 1
PDB	Protein Data Bank
PDI	Protein disulfide isomerase
PMSF	Phenylmethane sulfonyl floride
POI	Protein of interest
POPB	Prolyl oligopeptidase
PVDF	Polyvinylidene fluoride
RiPPs	Ribosomally synthesized and post-translationally modified peptides
RP	Reverse phase
rpm	Runs per minute
RT	Retention time
rvET	Real-value evolutionaray trace
SAXS	X-ray scattering
SDS-PAGE	Sodium Dodecyl Sulphate–Polyacrylamide Gel Electrophoresis
SEC	Size exclusion chromatography
Sec pathway	Secretary pathway
SFTI-1	Sunflower seed Trypsin Inhibitor
SmCB1	*Schistosoma mansoni* cathepsin B-like endopeptidase
SOC	Super optimal broth with catabolite repression
SP	Signal peptidase
SPPS	Solid phase peptide synthesis
TCEP	Tris[2-carboxyethyl]phosphine

TFA	Trifluoroacetic acid
TIM	Triose phosphate isomerase
TOF	Time-of-fly
TSA	Transcriptome shotgun assembly
TTCF	Tetanus toxin antigen
Ub	Ubiquitin
UET	Universal evolutionary trace
UPGMA	Unweighted pair group method with arithmetic mean
visualCMAT	visual Correlated Mutation Analysis Tool
VPE	Vacuolar processing enzyme

Abstract

Ligases are naturally occurring protein catalysts involved in biological processes that make life possible. Certain ligases are ATP-independent, require no cofactors, and function in physiological conditions. These stand-alone ligases are versatile biotechnological and biochemical tools for a wide array of applications, including site-specific modifications of proteins, antibodies, and live cells, as well as the cyclization of peptides and proteins. Asparaginyl endopeptidases (AEPs), also known as legumains or vacuolar processing enzymes (VPEs), are cysteine proteases that break Asn/Asp(Asx)-peptide bonds. AEPs have several functions pertaining to their ability to cleave peptides and proteins, including seed maturation in plants, antigen display in mammals, and hemoglobin digestion in ticks.

AEPs are also known to function as peptide asparaginyl ligases (PALs), catalyzing Asx-bond formation, the reversed catalytic function of AEPs. The prototype of PALs is butelase-1 from *Clitoria ternatea*. Butelase-1 exhibits high catalytic efficiency, broad substrate scope, and simple tripeptide recognition signal. The discovery of more PALs with diverse substrate preferences expands the repertoire of enzyme-mediated ligation methods, allowing one-pot ligation, bio-orthogonal ligation, and tandem ligation. However, PALs share highly similar sequences and structures with the proteases AEPs, making it highly challenging to identify PALs in a sea of AEPs.

Our laboratory has reported that there exist sequence motifs differentiating PALs from the ubiquitous AEPs, termed ligase activity determinants (LADs). The LAD1 and LAD2 are located at the substrate-

binding pockets S2 and S1', respectively. We showed that a Gly in the middle of LAD1 and a Gly-Pro dipeptide of LAD2 are conserved in the protease AEPs. In contrast, a bulky residue replacing Gly at LAD1 and a small hydrophobic dipeptide in LAD2 in the sequence are the indicators for PALs.

In this thesis, my aims are (**1**) to classify putative PALs and AEPs based on the LAD amino acid sequences by bioinformatics, (**2**) to validate the putative PALs and AEPs identified from my dataset using recombinant expression and biochemical assays, and (**3**) to engineer novel PALs from AEPs by mutagenesis of LAD sites. The proposed study simplifies and expedites the process of searching novel PALs, and could yield new insights into the ligase activity of PALs.

We established a dataset of 1570 sequences of putative PALs and AEPs by mining the online databases. Through evolutionary trace (ET) analysis and sequence conservation analysis, the evolutionary importance of LADs was demonstrated. Through recombinant expression and characterization of the putative novel PALs, which exhibited butelase-1-like ligase activity, the versatility of the LADs was shown. The expression and mutations of more examples from the dataset also validated that LADs are useful sequence motifs indicating the ligase activity of PALs and AEPs.

Combining bioinformatic analyses, site-directed mutagenesis, and functional studies, it was shown that LADs play critical roles in determining the enzymatic directionality of ligase or protease activities of PALs and AEPs, respectively. In addition to the sequence and structure, various conditions confer the ligase activity of the PALs and AEPs, including reaction pH, reaction solvent, and substrate used. My results expand the catalog of available

PALs and could pave the way toward the understanding of the molecular basis underpinning the ligase activity of PALs.

Chapter 1 Introduction

1.1 Occurrences of Ligases from Nature

Ligases are enzymes that ligate the peptides or proteins together by amine bond formation. In contrast, proteases break peptide bonds. Proteases are ubiquitous and ligases are rare [1]. ATP-dependent peptide bond formation during the strictly controlled gene-encoded translational process takes place at the ribosome. The process involves a large ribosomal machinery, ATP, GTP, mRNA, tRNA, and enzymes [2]. ATP-independent and stand-alone ligases are frequently found in the ribosomally synthesized and post-translationally modified peptides (RiPPs) pathways [3, 4]. They have been discovered in plant [5-10], fungi [11], bacteria [12], cyanobacteria [13], and yeast [14, 15]. ATP-independent and stand-alone ligases have spawned considerable interest in the field of biochemistry and biotechnology as aqueous-compatible superglue, and one of the prototypic ligases, butelase-1 [6], has been utilized for a myriad of applications (**Figure 1**).

1.1.1 Rarely Reported Ligase Acivity of Plant Asparaginyl Endopeptidase

Asparaginyl endopeptidases (AEPs), also frequently referred to as vacuolar processing enzymes (VPEs) or legumains [16], are proteases that cleave at carboxyl-terminal asparagine/aspartic acid (Asx) of peptides and proteins. AEPs were referred to as ‘vacuolar processing enzymes’ for their roles of proteolytic processing of the proteins in the vacuoles [17], the term VPE was first introduced in 1991 to describe the enzymes purified from dry castor beans (*Ricinus communis*) capable of processing the proprotein precursors [18]. The use of the term ‘legumain’ was first recommended by the Nomenclature Committee of International Union of Biochemistry and

Molecular Biology (NC-IUBMB) in the first edition of Enzyme Nomenclature in 1992 [19], the term was then first used by Kembhavi *et al*. to describe the cysteine endopeptidase found in legume seeds [20].

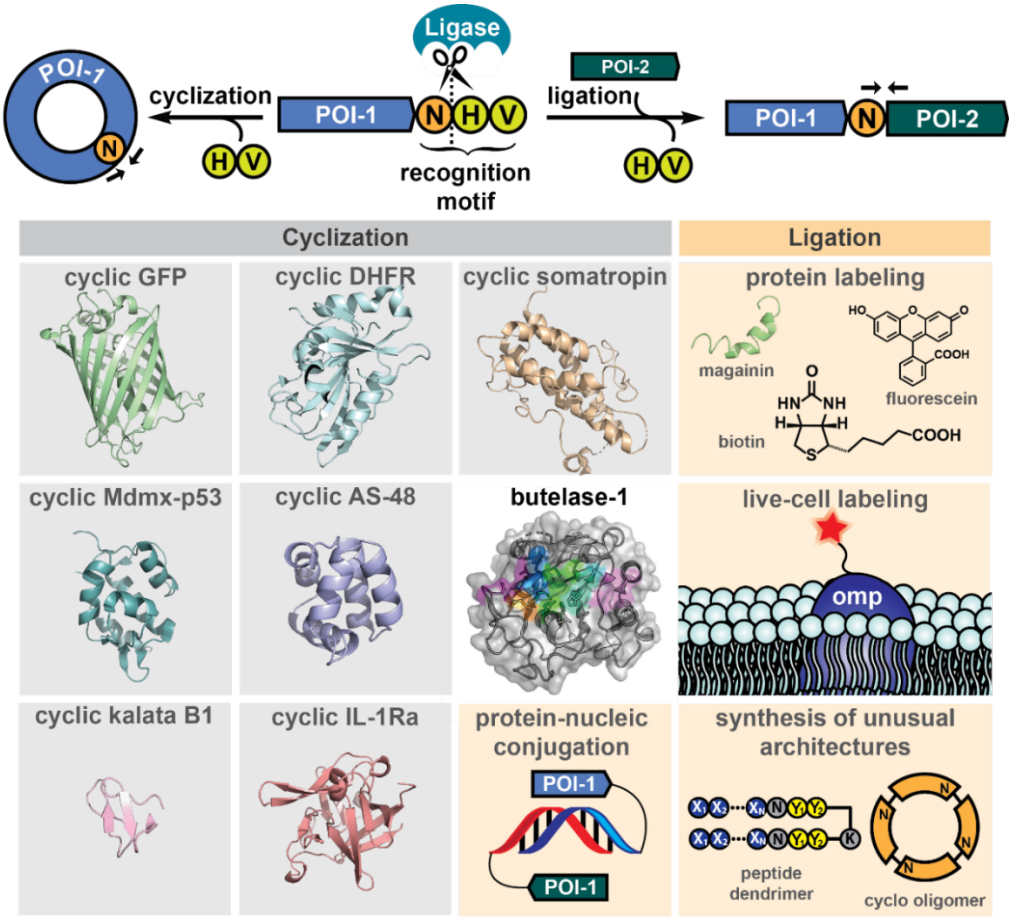


Figure 1. Applications using the prototypic ligase butelase-1. Butelase-1-mediated cyclization has been applied to various proteins, such as kalata B1, green fluorescence protein (GFP), and bacteriocin AS-48. The butelase-1-mediated intermolecular ligation has been applied to protein, live-cell labeling, production of protein-nucleic conjugation, and synthesis of peptides and proteins with unusual architectures, such as peptide dendrimer. Figure adapted from [21].

AEPs are classified as members of the C13 family, clan CD, of cysteine proteases (EC 3.4.22.34) featuring a His-Gly-spacer-Ala-Cys motif, where His and Cys are the catalytic residues [1, 22, 23]. In the 1980s, the proteolytic activity of AEPs was first observed in the developing plant seeds [17, 24-26]. Purified AEPs from plant seeds were then identified as proteases [27] recognizing the asparagine residues [18, 28, 29]. Later, plant AEPs were further characterized and found to be important for the processing of seed storage proteins, such as albumins and globulins, in various plants [30-32], and the mutation of AEPs resulted in the accumulation of precursors of seed storage protein [33, 34].

Compared to the protease activity, ligase activity of AEPs is rarely reported, although early in the 1990s their bifunctionality was proposed [35]. In 2014, an AEP, designated as butelase-1, with predominant ligase activity, simple tripeptide recognition motif Asn-Xaa-Yaa (Xaa and Yaa are any natural amino acid) at the P1-P1'-P2' position (nomenclature according to [36]), and fast kinetics was discovered, extracted, and purified from the pod of *Clitoria ternatea* [6]. The small groups of AEPs with predominant ligase activity were specifically termed peptide asparaginyl ligases (PALs). Since the discovery of butelase-1, several PALs were discovered, they include VyPAL1-2, HeAEP3, PxAEP3b, OaAEP1b, and OaAEP3-5 (**Table 1**) [7, 9, 10]. Those versatile PALs have been utilized for a myriad of applications, including macrocyclization, live cell-labeling, and synthesis of peptides with unusual architectures [21].

Table 1. List of reported peptide asparaginyl ligases [21].

Species	**Name**	**Activity Confirmed by**	**Accession no.**	**Ref.**
Clitoria ternatea	Butelase-1	*In vitro* assays	KF918345.1	[6]
Viola yedoensis	VyPAL1	*In vitro* assays	PRJNA494974	[10]
Viola yedoensis	VyPAL2	*In vitro* assays	PRJNA494974	[10]
Oldenlandia affinis	OaAEP1b	*In vitro* assays	KR259377	[7]
Oldenlandia affinis	OaAEP3	*In vitro* assays	KR259379	[9]
Oldenlandia affinis	OaAEP4	*In vitro* assays	LQ854853.1	[9]
Oldenlandia affinis	OaAEP5	*In vitro* assays	LQ854855.1	[9]
Petunia x hybrida	PxAEP3b	*In planta* assays	MG720076.1	[8]
Hybanthus enneaspermus	HeAEP3	*In planta* assays	MG720074.1	[8]

1.1.2 Cyclization of Prosegetalin A by Serine Protease-Like Ligase PCY1

Orbitides, also known as Caryophyllaceae-like cyclopeptides, are cyclic peptides lacking disulfide bonds from plants, they were first isolated from the flaxseed oil in 1959 [37]. The sizes of orbitides usually range from 5 to 12 amino acids [3]. Currently, there are 191 peptide sequences classified as members of orbitides on the online database Cybase [38]. Bioactivities of orbitides include but are not restricted to cytotoxicity, antiplatelet activity, and antimalarial activity [39].

Segetalin A is a cyclic hexapeptide that possesses estrogen-like activity found in *Saponaria vaccaria* [40]. Segetalin A was found to be derived from the ribosomally-synthesized precursor Presegetalin A1 of 32 amino acids [41]. Later, Peptide Cyclase 1 (PCY1), a serine protease-like ligase discovered in *Saponaria vaccaria*, and the oligopeptidase OLP1 are found to be responsible for the post-translational processing of Segetalin A [5]. The precursor Presegetalin A1 (MSPILAHDVVKPQGVPVWAFQAKDVENASAPV) is cleaved by OLP1 at the N-terminus, following the cleavage and cyclization of the C-terminal remaining peptide (GVPVWAFQAKDVENASAPV), resulting in the mature cyclic Segatalin A of six residues (GVPVWA), and the linear follower peptide (FQAKDVENASAPV) (**Figure 2**).

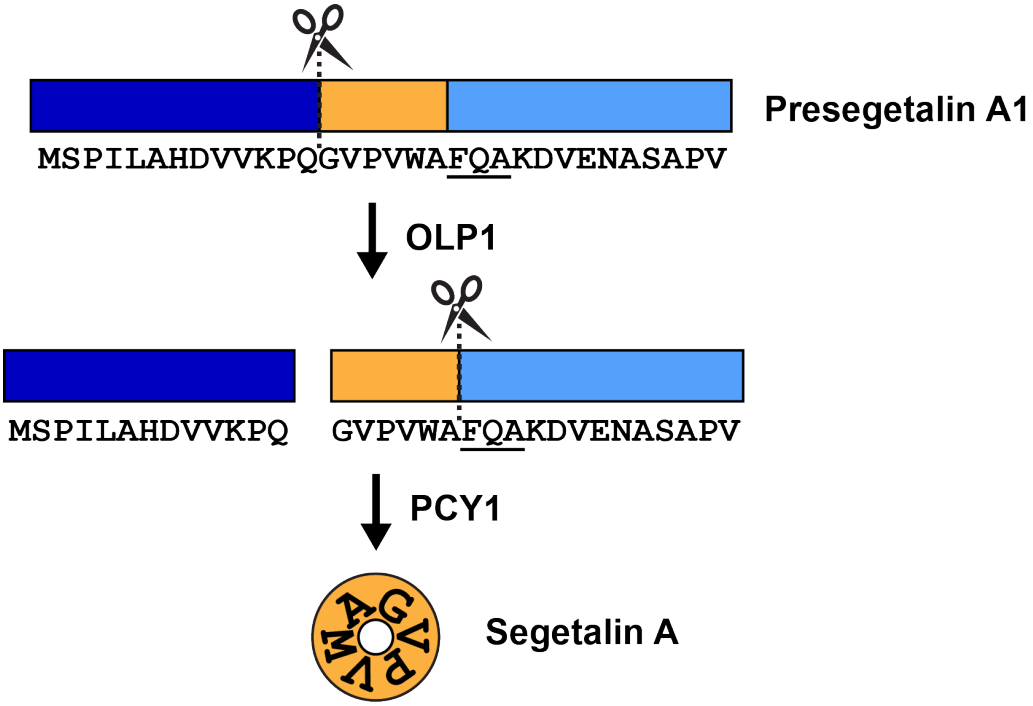


Figure 2. Post-translational modification of segetalin A by OLP1 and PCY1. The precursor of segetalin A, presegetalin A1, was first cleaved by OLP1 at the Gln and Gly, then cleaved and cyclized by PCY1 at the C-terminus to form the mature cyclic segetalin A. Figure adapted from [5].

PCY1 belongs to the S9 family of clan SC [1], and can cyclize various Presegetalin from *S. vaccaria* [5]. Kinetics analysis using different peptide substrates revealed that a C-terminal tripeptide motif FQA at P1'-P3' position [36] is preferred by PCY1 [42]. Notably, conformational change of PCY1 is induced by the binding of presegetalin A1, and the addition of the follower peptides led to the concentration-dependent decrease in turnover rate [43].

1.1.3 Bioprocessing of Petellamides by Serine Protease-Like Ligase PatG

Patellamides, which belong to the cyanobactin superfamily, have been found in ascidians of the Diemnidae family [44]. The origin of patellamides, the symbiont *Prochloron spp.* or the host ascidians, was solved by Schmidt *et al.* by identification of genes responsible for the biosynthesis, the *pat* gene cluster, and recombinant expression [45].

The *pat* gene cluster contains seven genes, including *patA*, *patB*, *patC*, *patD*, *patE*, *patF,* and *patG*. Among them, *patA*, *patD*, *patE*, *patF*, and *patG* were found central to the biosynthesis of patellamides. PatA and PatD were both predicted to contain two domains. The N-terminal region of PatA was found to be similar to subtilisin-like proteases. The C-terminus of PatD was proposed to be involved in the cyclization of Cys and Thr residues of PatE. The gene *patE* encodes the 71-residue peptide precursor containing patellamides A and C. PatG was predicted to be a protein containing multiple domains. The N-terminal domain of PatG was found to be homologous to NAD(P)H oxidoreductases, and the C-terminal domain of PatG, like the PatA N-terminus, contains a subtilisin-like serine protease region [45].

Later, to scrutinize the functions of PatA and PatG during the biosynthesis of patellamides, PatA and PatG were both recombinantly

expressed, purified, and subjected to biochemical assays. Incubation of PatA and the artificial PatE precursor (PatEdm) resulted in N-terminal cleavage of PatEdm. On the contrary, PatG was found to cyclize the peptide substrates with a C-terminal AYDG motif without the presence of PatA and cofactors. It was thus established that the production of mature patellamides involves the N-terminal cleavage by PatA and cyclization by PatG [13].

Structural and biochemical analyses of PatG later showed that PatG belongs to the family S8 of clan SB, with a catalytic Asp-His-Ser triad [1, 46]. During the bioprocessing of patellamide A and C, PatD first catalyzes cyclodehydration reaction, following cleavage of the precursor by PatA. PatG then cleaves before the C-terminal tetrapeptide motif AYDG at the P1'-P4' sites [36] and mediates the cyclization of the remaining peptide, resulting in the mature patellamide A and C (**Figure 3**) [47].

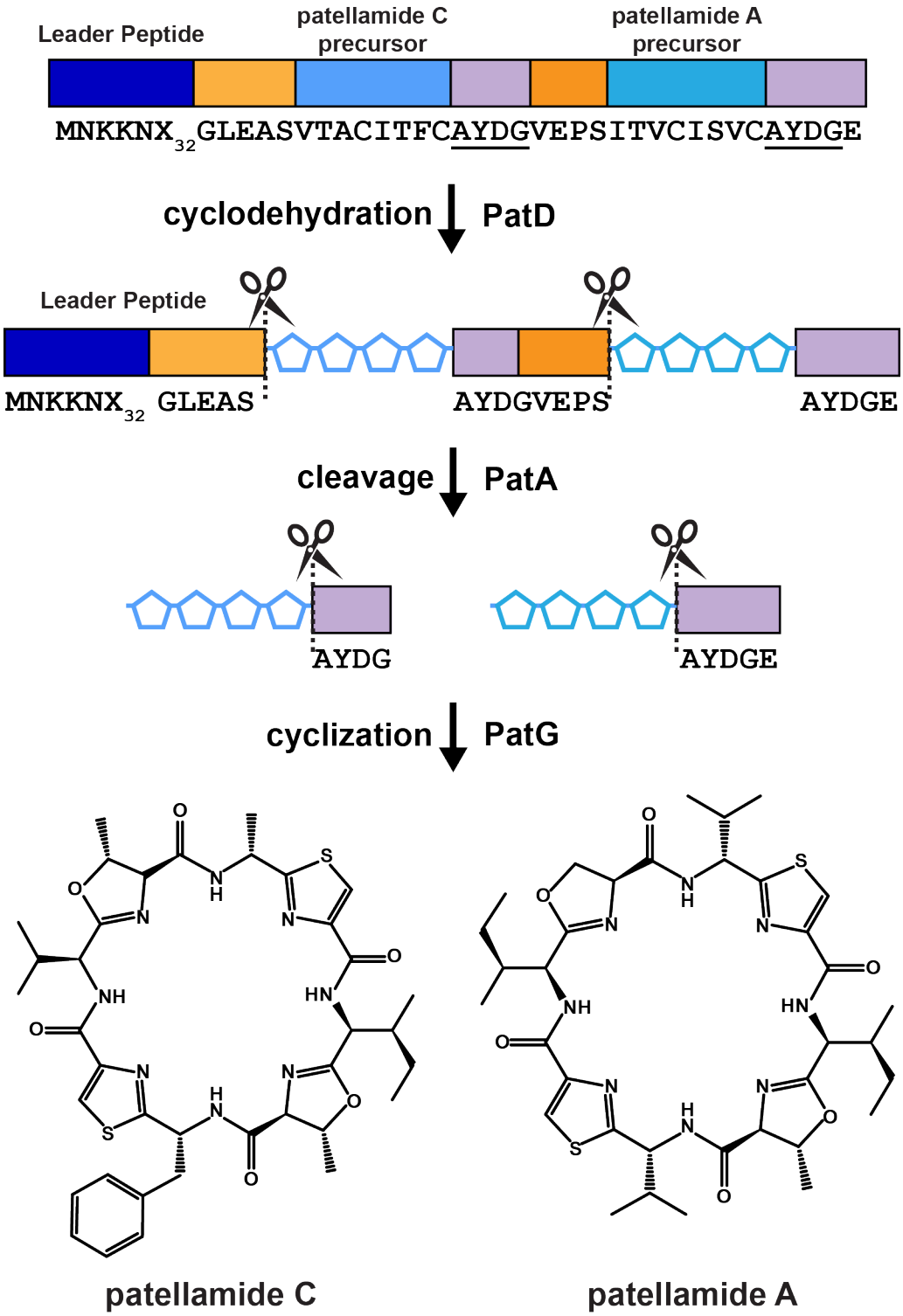


Figure 3. Post-translational modification of patellamide A and patellamide C by PatD, PatA, and PatG. During the maturation of patellamide A and patellamide C, cyclodehydration reaction of the Cys, Ser, and Thr residues of the precursor was catalyzed by PatD. Next, PatA catalyzes the cleavage of the precursor. Lastly, PatG cleaves and cyclizes the cleaved fragments of precursor to form patellamide A and patellamide C. Figure adapted from [47].

1.1.4 Maturation of Amatoxin by Mushroom Serine Protease-Like Ligase POPB

Amatoxins are frequently found in mushrooms and are RiPPs. They are notorious for mushroom poisonings because of their ability to inhibit RNA polymerase II. α-amanitin is a ribosomally encoded bicyclic octapeptide containing a Trp-Cys cross-bridge [48]. α-amanitin interacts with the RNA polymerase II bridge helix and inhibits the translocation of RNA polymerase II [49].

The genes encoding prolyl oligopeptidase (POP) were found to only present in those amatoxin-producing species, including *Galerina badipes*, *Galerina venenata*, *Galerina hybrid*, and *Galerina marginata*. It was thus proposed that the POPs in those species, POPA and POPB, might be involved in the post-translational processing of α-amanitin precursor [48]. Luo *et al.* later obtained the two POPs by yeast expression and assayed them using the α-amanitin precursor (GmAMA1) of 35 amino acids [11]. POPA was unable to process the GmAMA1, while POPB converted the GmAMA1 to the mature and cyclic α-amanitin (IWGIGCNP). The transient accumulation of the intermediate of 25 amino acids indicates that POPB first catalyzes the cleavage of the octapeptide at the N-terminus, following the cyclization of the mature octapeptide and the release of C-terminal peptide of 17 amino acids (**Figure 4**) [11].

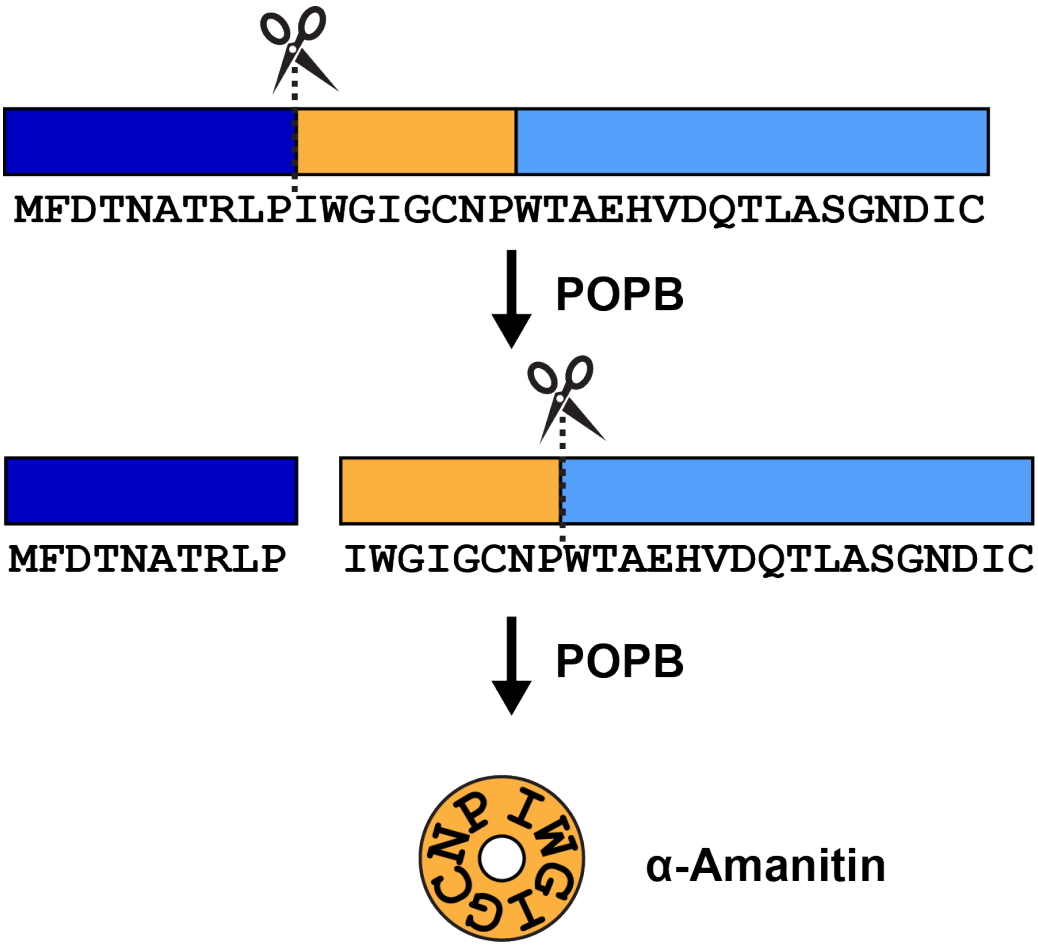


Figure 4. Post-translational modification of α-amanitin by POPB. POPB catalyzes the cleavage between the Pro and Ile as well as Pro and Trp at the N-terminus and the C-terminus, respectively, of the mature α-amanitin. POPB then cyclizes the octapeptide IWGIGCNP, resulting in cylic α-amanitin. Figure adapted from [11].

The substrate scope of POPB was screened by Sgambelluri *et al.* using the octapeptides IWGIGCNP with substitutions at P2 to P7 position with Ala, Ser, Leu, and Asn. POPB was able to cyclize all of the peptide substrates with amino acid substitution at different rates. Replacement of P7-Trp with polar residue, Ser and Asn, resulted in decreased cyclic product yields, suggesting that POPB prefers a nonpolar residue at the P7 position. Gly-to-Ala mutations at P4 and P6 positions of the octapeptide substrates did not affect the yield, while mutations of the P4-Gly and P6-Gly to Ser, Leu, and Asn led to a lower yield of cyclic products. The minimum length of the peptide substrates required for POPB to catalyze cyclization was suggested to be eight residues, as POPB only cleaves the peptide substrates with six residues (IWGIGP) and seven residues (IWGIGCP) [50].

1.1.5 Cell Wall Sorting of Surface Proteins by Sortase A

Sortases covalently link the surface proteins to the peptidoglycan cell wall of Gram-positive bacteria, allowing virulence factors to be displayed [51]. These enzymes are thus named 'sortases' for their roles in 'sorting' proteins to the cell wall. Before the discovery and isolation of sortases, Schneewind *et al.* found a conserved motif LPXTG (where X stands for all amino acids) in homologous sequences of many surface proteins found in Gram-positive bacteria [52]. Later, it was found that the cleavages of the protein precursors are between the Thr and Gly residue of the LPXTG pentapeptide motif [53], and the carboxyl of Thr is linked to the pentaglycine cross-bridge of the cell wall [54]. By mutational studies, Sortase A was confirmed to be the enzyme responsible for the cell wall sorting of surface proteins [12]. Sortase A catalyzes transpeptidation reaction by cleaving between the Thr residue and

Gly residue of the LPXTG motif of the protein, forming acyl-enzyme intermediate with Cys of Sortase A linked to the carbonyl of Thr, the intermediate is then attacked by the pentaglycine on the cell wall of Gram-positive bacteria (**Figure 5**) [55].

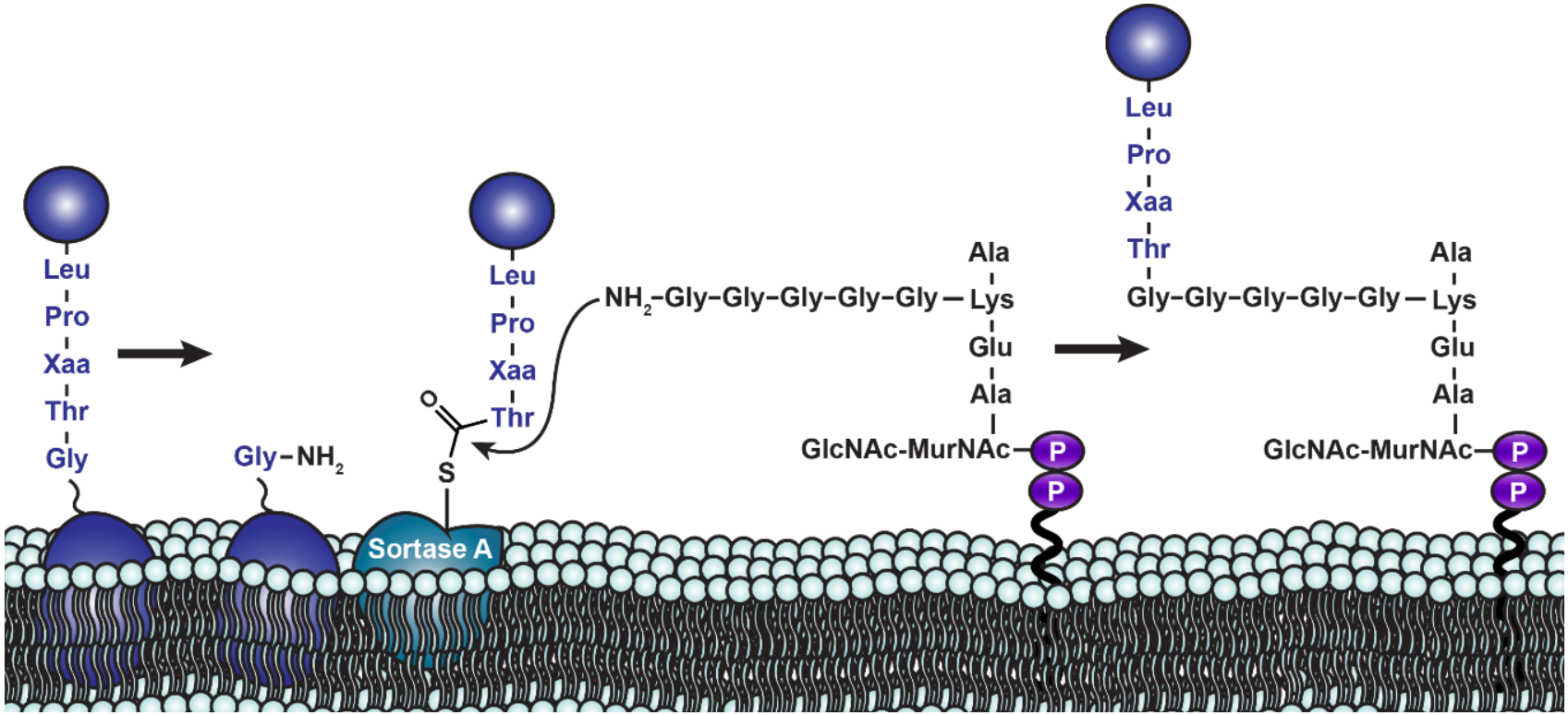


Figure 5. Anchoring of surface protein at the peptidoglycan cell wall of Gram-positive bacteria by sortase A. Sortase A recognizes the pentapeptide motif LPXTG of the substrate (colored in blue), cleaves between the Gly and Thr residue, and covalently links the carboxyl of Thr to the lipid II (colored in purple) of the cell wall. Figure adapted from [56].

Compared to Sortase B, which is only found in a few bacteria, Sortase A is widely found in Gram-positive bacteria and catalyzes transpeptidation reaction for anchoring various proteins to the cell wall [51]. Sortase A has been utilized for a myriad of applications as it is easily expressed in *Escherichia coli*, and it can link the peptides and proteins easily with the presence of the LPXTG motif and the pentaglycine as incoming nucleophile [56, 57].

1.1.6 Intein-Mediated Protein Splicing

Protein splicing is one of the post-translational modifications, during which multiple proteins can be coded from one gene. Inteins (*in*tervening pro*teins*) are a class of auto-processing proteins, they excise themselves out from the polypeptide precursors, following the peptide bond formation between the neighboring exteins (*ex*ternal pro*teins*), which are separated by inteins before the excision (**Figure 6**) [58, 59].

Inteins were first reported by Hirata *et al.* and Kane *et al*. in the yeast *Saccharomyces cerevisiae* [14, 15] and are widely found in all domains of life. Inteins can be distinguished into two groups, the contiguous inteins and the split inteins. Split inteins are less common compared to the contiguous inteins, which only involve one transcript. Split inteins are transcribed and translated from two different genes, and the removal of inteins and ligation of exteins happens after the split inteins assemble non-covalently (**Figure 7**) [58].

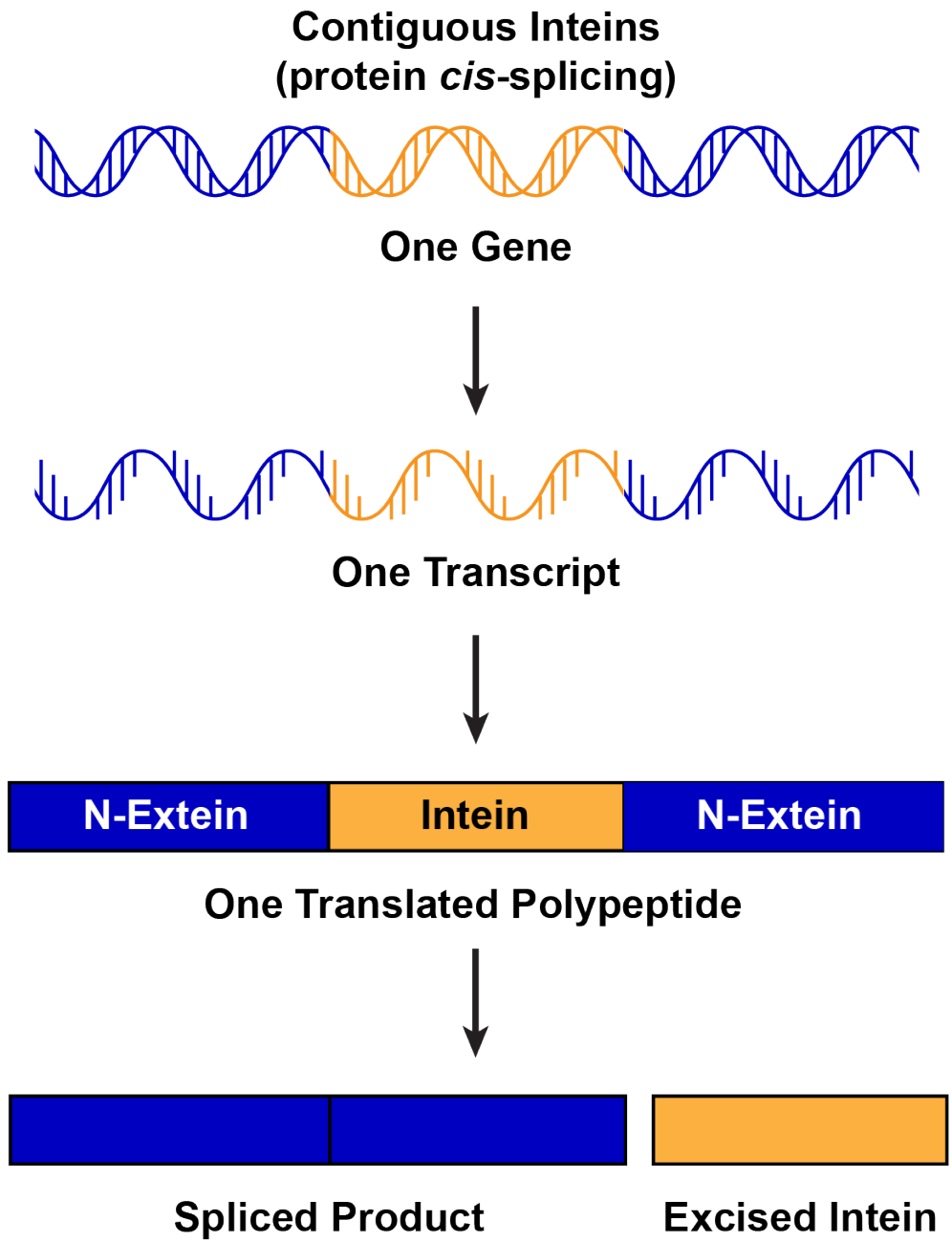


Figure 6. Protein splicing by inteins. Contiguous intein catalyzes spontaneous protein *cis*-splicing, which involves only one gene, one transcript, and one translated polypeptide. Intein excises itself from the polypeptide, resulting in the spliced product. Figure adapted from [58].

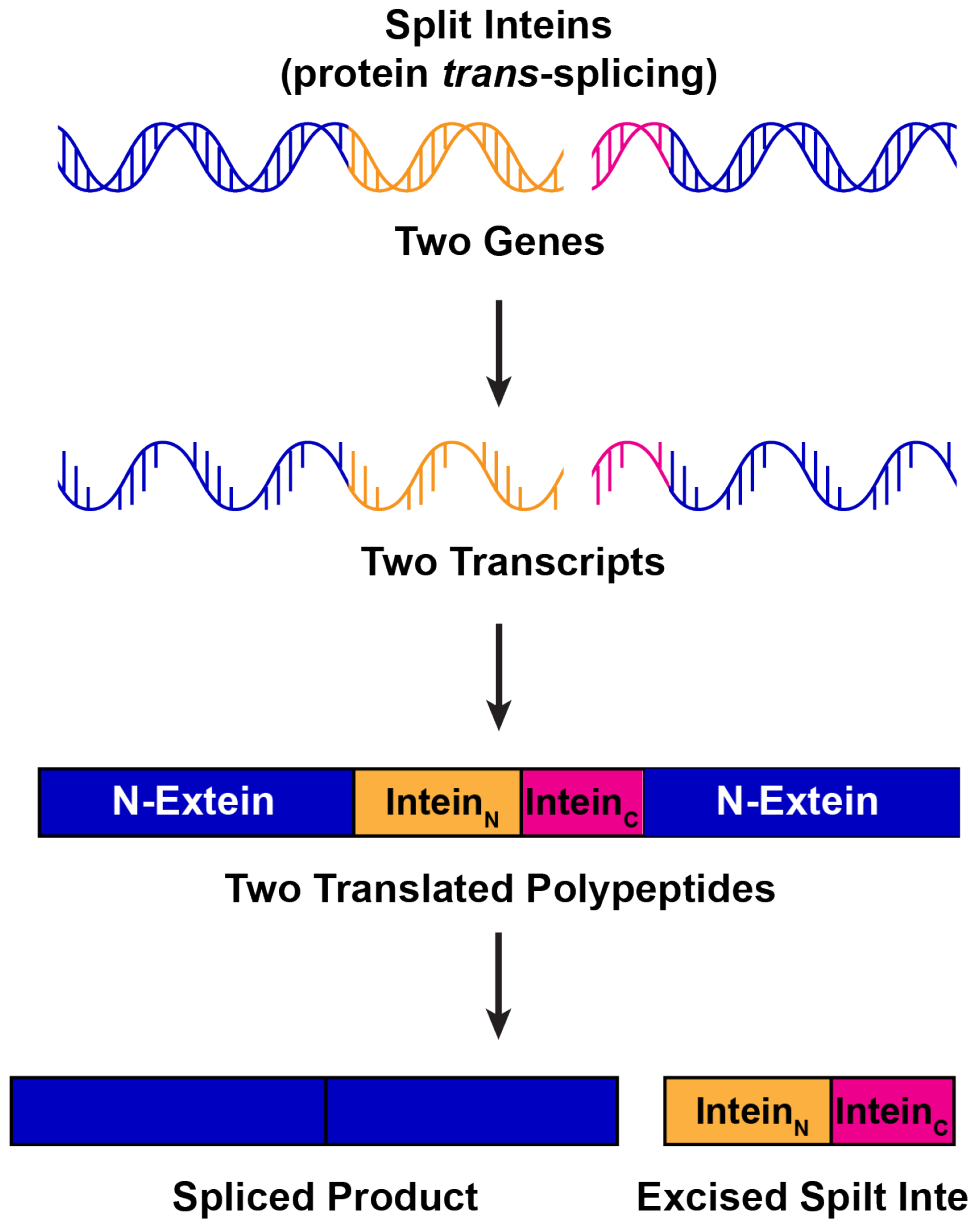


Figure 7. Protein splicing by split inteins. Split inteins catalyze spontaneous protein *trans*-splicing, which involves two genes, two transcripts, and two translated polypeptides. Figure adapted from [58].

The post-translational protein splicing is spontaneous. Generally, intein-mediated protein splicing starts with the conserved Cys or Ser of inteins nucleophilically attacking the C-terminal carbonyl carbon of the N-extein. The N-to-O/S acyl shift results in a linear thioester or ester intermediate. Subsequently, a branched intermediate is formed through transesterification, nucleophilic attack of the conserved Cys, Ser, or Thr of C-extein. The cyclization of C-terminal invariant Asn of intein leads to the removal of intein succinimide, and the O/S-to-N acyl shift results in peptide bond formation and the spliced product [58].

1.1.7 Spontaneous Isopeptide Bond Formation by Engineered Ligases from *Streptococcus pyogenes*

Intramolecular Lys-Asn isopeptide bond, amide bond involving the ε-amino group of lysine, was first discovered in the crystal structure of *Streptococcus pyogenes* major pilin protein Spy0128 [60]. Zakeri and Howarth dissected the Spy0128 into two segments, the 'isopeptage,' which contains the reactive Asn (TDKDMTITFTNKKDAE) and the pilin-C, which contains the reactive Lys. They found that the intermolecular amide bond formation happens spontaneously between the sidechain of Lys and Asn of the two fragments [61].

Zakeri *et al.* split and modified the collagen adhesion domain (CnaB2) from *Streptococcus pyogenes* and developed the SpyTag, the C-terminal 13-mer peptide tag containing the reactive Asp, and the SpyCatcher, the N-terminal protein fragment of 138 amino acids containing the reactive Lys and catalytic Glu (**Figure 8**). The carbonyl carbon of Asp is nucleophilically attacked by the unprotonated amine of Lys, the intramolecular isopeptide bond

formation is catalyzed by the Glu in close proximity. Similar to the isopeptag and pilin-C, SpyTag and SpyCatcher also associate covalently and spontaneously [62].

To reduce the size of the tag left in the ligated product, Zakeri *et al.* next further dissected the SpyCatcher into KTag (ATHIKFSKRD), the β-stand of CnaB2 containing the reactive Lys, and SpyLigase, which contains the catalytic Glu. Mixing of the three fragments, SpyTag, KTag, and SpyLigase, resulted in the irreversible and covalent isopeptide bond formation (**Figure 9**) [63, 64]. Later, applying the same principle, Buldun *et al.* developed SnoopLigase, SnoopTagJr, and DogTag from *Streptococcus pneumoniae* for spontaneous transamidation. Covalent association of the SnoopTagJr and DogTag catalyzed by SnoopLigase was proposed to be more efficient than that by Spy Ligase [65, 66].

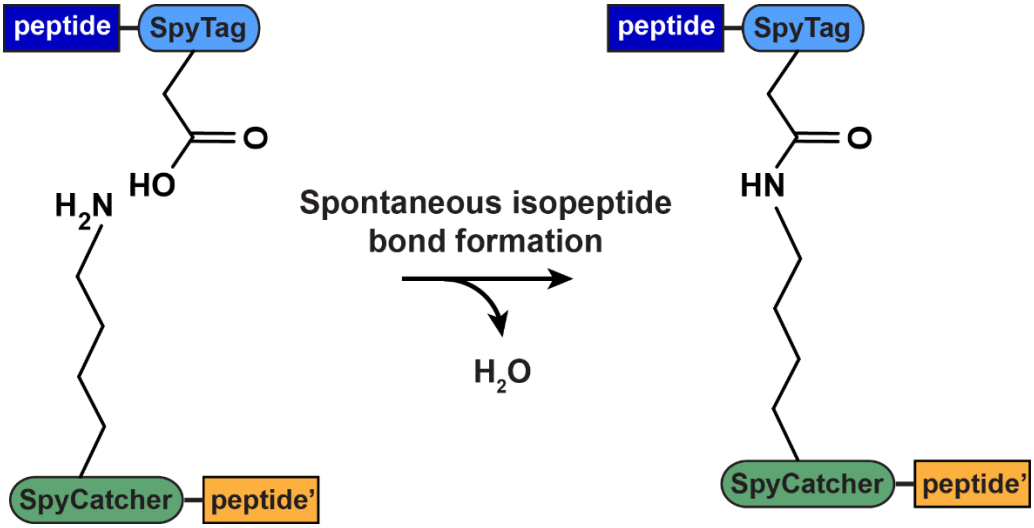


Figure 8. Spontaneous Lys-Asn isopeptide bond formation by SpyTag and SpyCatcher. The CnaB2 of FbaB from *Streptococcus pyogenes* was dissected and modified into two fragments by Zakeri *et al.*, resulting in Spytag containing the reactive Asp and the SpyCatcher containing the reactive Lys and catalytic Glu. Incubating the two fragments together leads to spontaneous isopeptide formation. Figure adapted from [67].

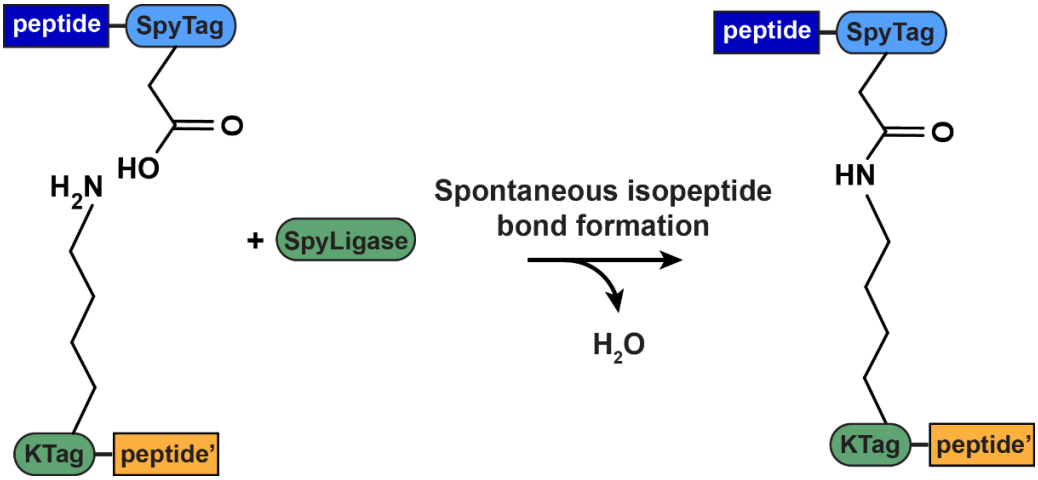


Figure 9. Spontaneous isopeptide bond formation by SpyTag, KTag, and SpyLigase. By further dissecting the SpyCatcher into KTag (ATHIKFSKRD), which contains the reactive Lys, and SpyLigase, which contains the catalytic Glu, the size of protein left on the product can be reduced. Mixing the three fragments leads to irreversible isopeptide bond formation. Figure adapted from [67].

1.1.8 Peptide Bond Formation by Engineered Subtilisin BPN' Variants

Subtilisin BPN' is a serine protease of S8 family, clan SB, from *Bacillus amyloliquefaciens*. Subtilisin BPN' catalyzes hydrolysis reaction through attacking the peptide bond by its catalytic Ser then hydrolyze the acyl-enzyme intermediate subjected to subsequent hydrolysis (**Figure 10**) [68]. With two key mutations, subtilisin BPN' was successfully converted to a peptide ligase with broad substrate specificity, subtiligase, which has been applied for engineering, synthesis, and conjugation of proteins [68]. Thiolsubtilisin, a subtilisin variant with largely reduced amidase activity, was first generated by the chemical mutagenesis approach in the 1960s by replacing the catalytic Ser of subtilisin BPN' to Cys (S221C) [69, 70]. Another variant of subtilisin, selenosubtilisin, was later generated by replacing the catalytic Ser with selenocysteine. Although selenosubtilisin has a 20-fold faster aminolysis kinetic compared to thiolsubtilisin, it is more susceptible to oxidative inactivation [71].

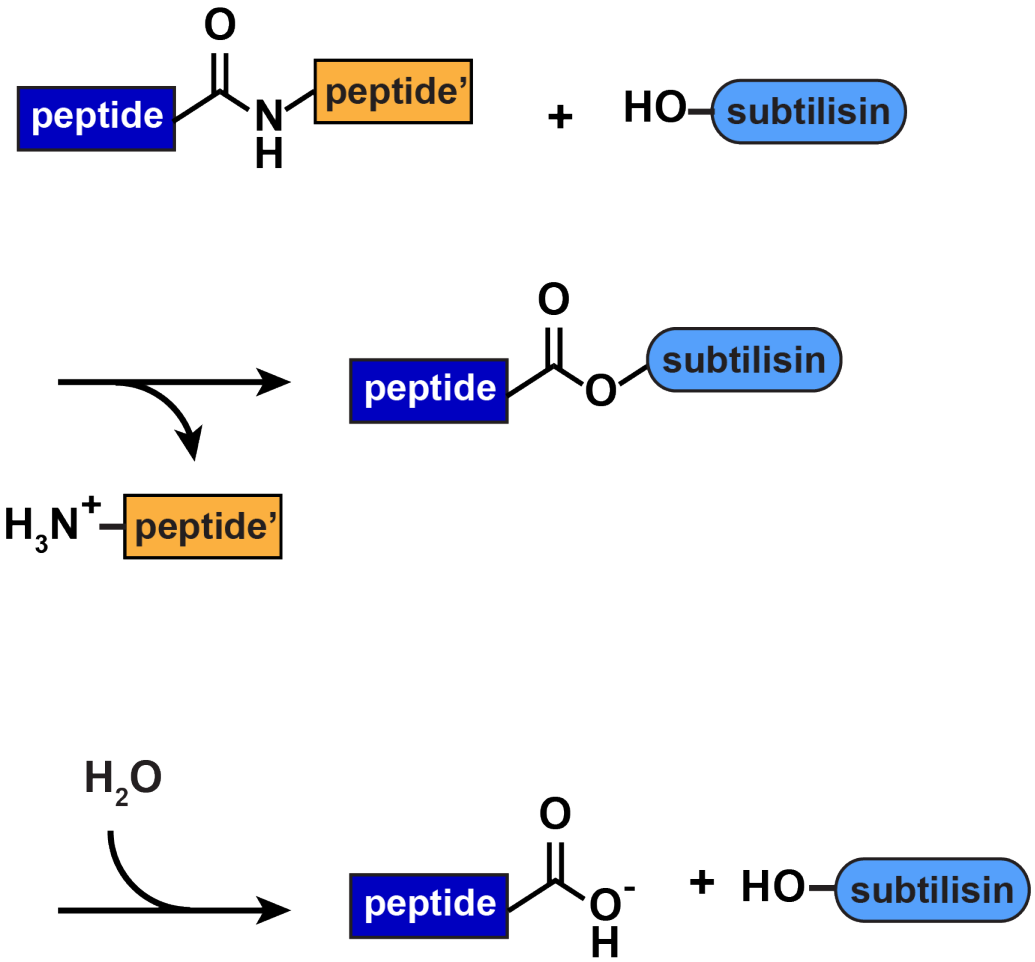


Figure 10. Subtilisin-catalyzed hydrolysis. Subtilisin BPN' is a serine protease and it catalyzes peptide bond hydrolysis. Figure adapted from [68].

To improve the efficiency of thiolsubtilisin, a Pro to Ala mutation (P225A) at helix α6 was introduced to thiolsubtilisin. With the two mutations (S221C/P225A), the subtilisin variant with enhanced ligase activity and 100-fold lower amidase activity, subtiligase, was generated (**Figure 11**) [72]. Later, more mutations were introduced to create more stable and more efficient subtilisin BPN' variants. Five mutations that were previously found to improve stability of subtilisin BPN', M50F, N76D, N109S, K213R, and N218S, were introduced to the subtiligase to obtain variants with enhanced stability against heat, alkali, and organic solvents [73]. Toplak *et al.* then designed a stable calcium-independent subtilisin BPM' variant compatible with organic co-solvents and denaturing agents, peptiligase, by deleting nine amino acids composing the calcium-binding domain and adding 18 mutations for stability enhancement. Peptiligase catalyzes irreversible peptide bond formation between an N-terminal amine and a C-terminal carboxamidomethyl (Cam)-ester in aqueous conditions [74].

In addition to mutating the subtilisin BPN' variants, substrates can also improve the ligation efficiency. Liu and coworkers used peptide thioesters, instead of esters, as the substrate to facilitate acyl-enzyme thioester intermediate formation and largely enhanced the catalytic efficiencies [75].

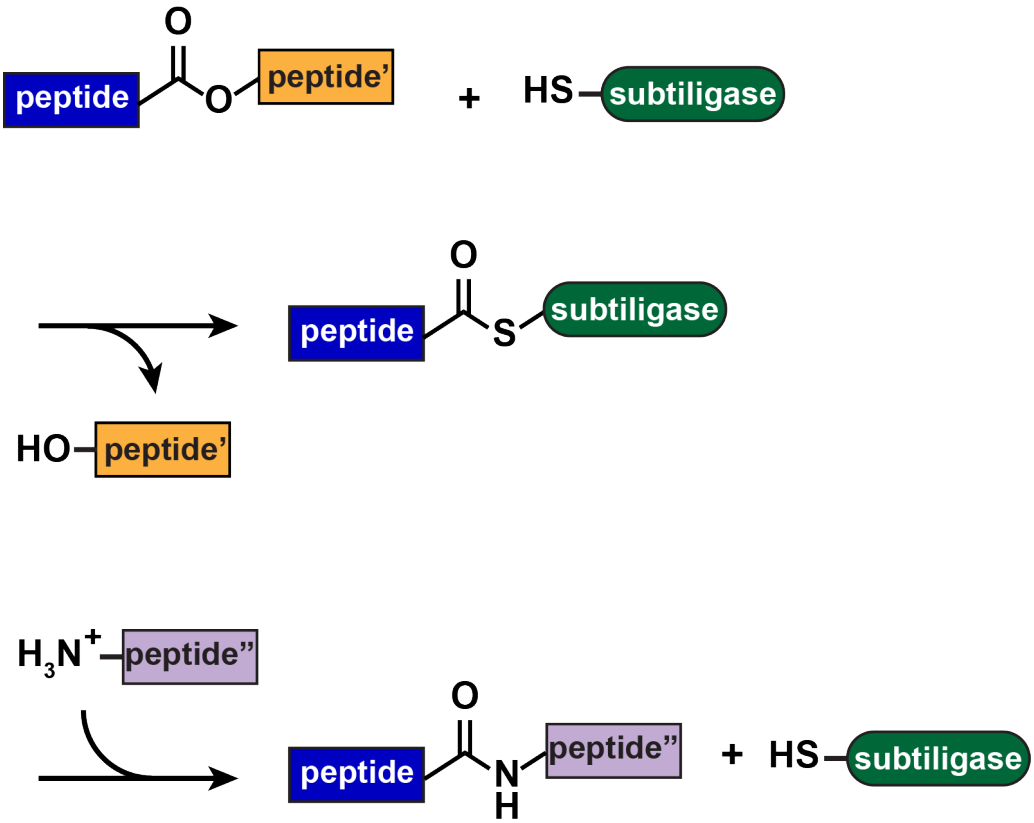


Figure 11. Subtiligase-catalyzed peptide bond formation. Subtiligase is a variant of subtilisin with two key mutations (S221C and P225A) that performs ligation instead of hydrolysis. Catalytic Cys of Subtiligase attacks the peptide esters or thioesters, leading to the formation of a thioacyl-enzyme intermediate. Subsequently, the intermediate is intercepted by the incoming amine, leading to the formation of peptide bond. Figure adapted from [68].

1.1.9 Peptide Bond Formation by Engineered Trypsin — Trypsiligase

Trypsin is a serine protease that mainly cleaves the Lys and Arg residues at the P1 position. Many efforts have been made to generate trypsin variants with diverse substrate scopes. Trypsin specifically recognizes residues with a basic side chain. Crystal structure of bovine trypsin in complex with pancreatic trypsin inhibitor showed that Asp of the base of substrate binding pocket S1 interacts with Lys15 of the inhibitor and may be responsible for the specificity toward basic substrates [76]. Graf *et al*. thus proposed that mutating the Asp of the base of substrate binding pocket S1 to Lys (D189K) can alter the substrate specificity of trypsin. Trypsin variant D189K showed low catalytic efficiency against P1-Tyr compared to the wild-type trypsin and altered substrate specificity, the trypsin variant D189K is able to cleave between Leu and Arg. Computer modeling of the trypsin variant D189K suggested that Leu interacts with neighboring oxygen, making the positively charged NH_3^+ group of Leu inaccessible to the incoming substrate; the resulted hydrophobic substrate-binding pocket thus allows the sidechain of Leu to bind [77].

Trypsin has little preference toward the P2' position of the substrate [78]. Computer modeling showed that Asn143 and Glu151 are potential metal-binding sites that bridge to P2'-His of the substrate. By introducing two mutations, N143H and E151H, to the trypsin, Willett *et al.* showed that trypsin variant N143H/E151H is able to efficiently recognize and cleave the His at the P2' position of the substrate, AGPYAHSS) when a nickel or zinc is present [79].

To modify the substrate specificity of the primed side of trypsin, Kurth *et al.* performed molecular modeling and proposed that Lys60 could determine the substrate preference for trypsin at the P1' and the P3' residues. By mutating the Lys60 to Glu, the substrate specificity toward P1' residue was altered. The trypsin variant K60E prefers a P1'-Arg-containing substrate, which was proposed to be a result of the salt bridge between Glu60 of trypsin variant K60E and Arg at the P1' position of the substrate by molecular modeling [80].

By incorporating the abovementioned four mutations D189K, K60E, N143H, and E151H, which alter the substrate preference at the P1, P1', P2' and P2' position, respectively, Leibscher *et al.* generated a variant of trypsin, Trypsiligase, which specifically cleaves between Tyr and Arg of the tripeptide motif Tyr-Arg-His (**Figure 12**) [81]. It was postulated by Liehscher *et al.* that with the presence of YRH-containing substrate and zinc, trypsiligase is converted into active conformation and catalyzes the cleavage between the Tyr and Arg of the Tyr-Arg-His motif. On the other hand, with the presence of reactive *O*-guanidinophenyl esters and zinc ions, ligation reaction is favored by trypsiligase [81].

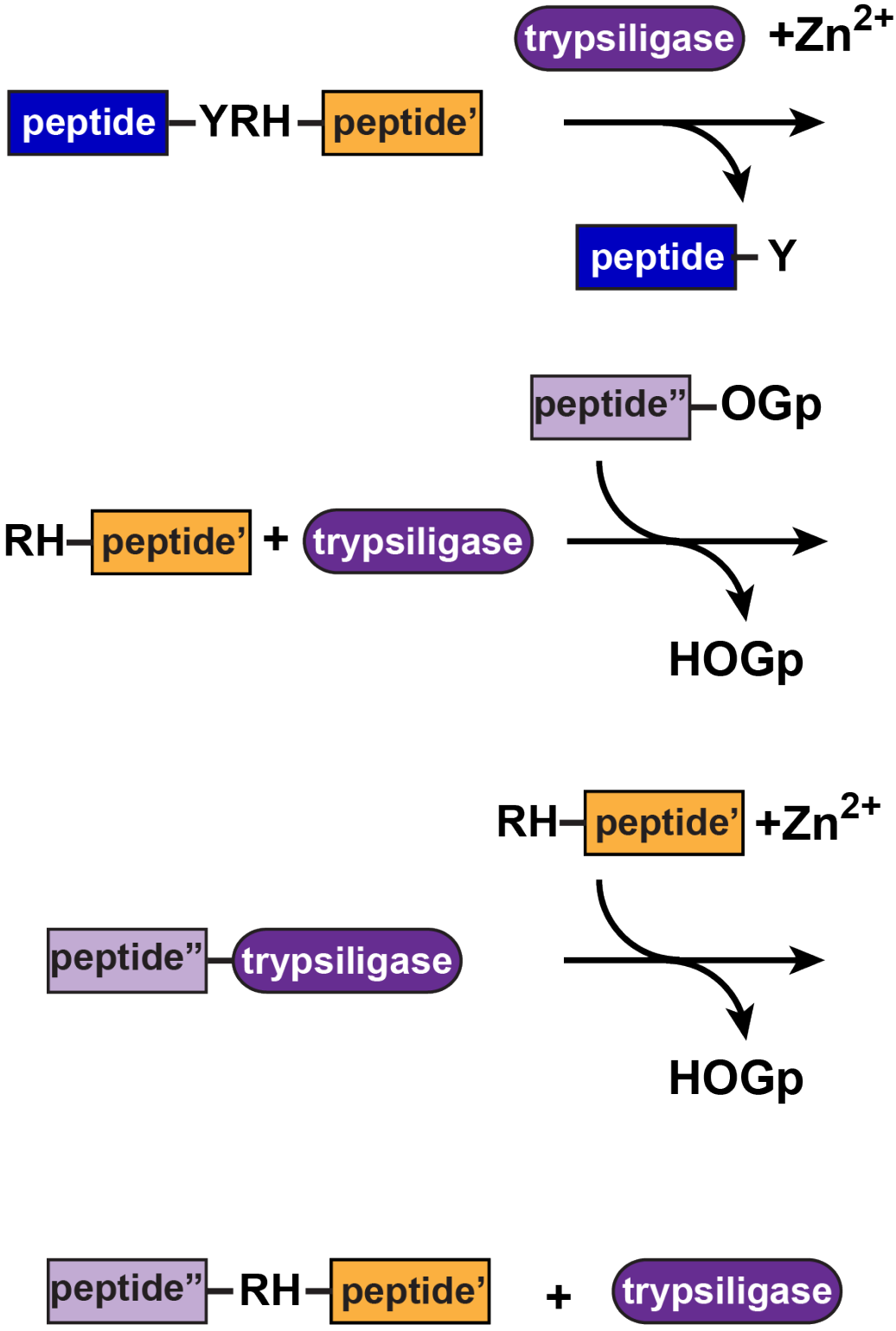


Figure 12. Peptide bond formation by Trypsiligase. Trypsiligase specifically cleaves between the Tyr and Arg of the YRH motif, and the reactive *O*-guanidinophenyl esters will then be ligated to the cleaved peptide carrying the Arg-His motif. Figure adapted from [81].

1.2 The Peptide Asparaginyl ligases (PALs) and Asparaginyl Endopeptidase (AEPs)

Commonly identified in plant seeds during seed maturation, AEPs are also widely found in the vegetative organs of plants [82-85], mammals [86, 87], and parasites [88]. Interestingly, multiple isoforms of AEPs are frequently observed in plants but not in mammals [1].

1.2.1 Discovery of Peptide Asparaginyl Ligases (PALs) in Plants

Several peptide asparaginyl ligases (PALs) have been discovered and isolated or purified from the plants, they are frequently found in the cyclic peptide-producing plant families, the Cucurbitacae, Fabaceae, Rubiaceae, Solanaceae, and Violaceae family [89]. Shortly after the discovery of butelase-1 of the Fabaceae family, the prototype of PAL, OaAEP1b was reported in the African flowering plant *Oldenlandia affinis* of the Rubiaceae family [7]. Kalata B1 is a cyclic peptide of 29 amino acids found in *O. affinis*, and it possesses various biological activities, such as anti-HIV activity, insecticidal activity, and uterotonic activity [90]. The mature Kalata b1 is excised from its precursor Oak1, and Harris *et al.* showed that OaAEP1b is able to cyclize the precursor of Kalata B1 with the native C-terminal motif after the P1-Asn. However, OaAEP1b is unable to process the N-terminus of the Kalata b1 precursor [7]. Later, three more PALs from *O. affinis*, OaAEP3-5, were identified and shown to be more efficient than OaAEP1b [9].

Due to the abundance of cyclic peptides reported in the Violaceae family (**Figure 13**) (**Appendix A**) [38], it is not surprising that several PALs were found in this family. HeAEP3 from *Hybanthus enneaspermus* was reported to function preferably as a ligase compared to its isoforms by *in planta*

assay, however, the activity was not tested *in vitro* [8]. In 2019, more PALs from the viola family were reported. VyPAL1-2 were discovered in *Viola yedoensis* and they catalyze efficient cyclization reaction with negligible hydrolysis product *in vitro* [10].

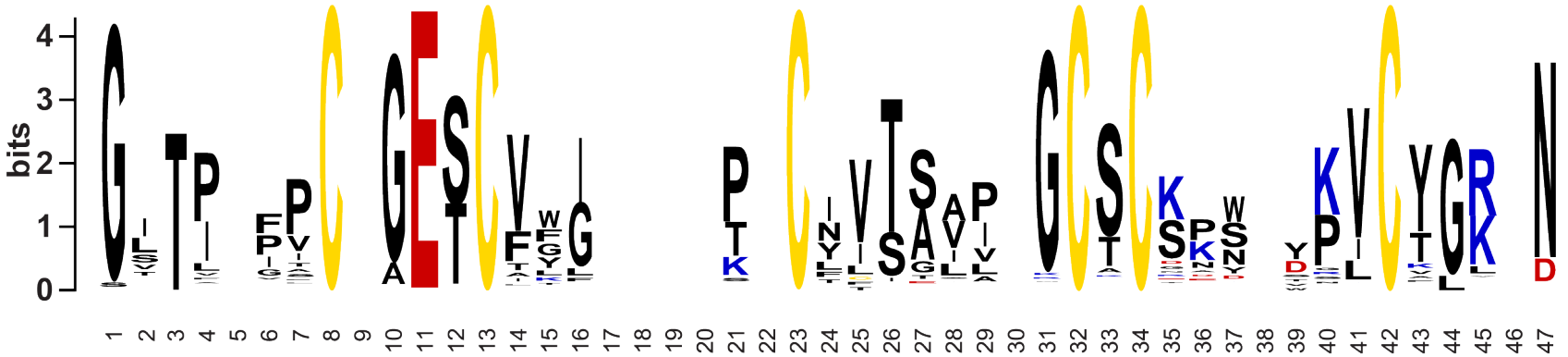


Figure 13. Cyclic peptides reported in the Violaceae family. Blank indicates predominantly deletion. See Appendix A for the sequences of cyclic peptides used to generate this WebLogo. The sequences are retrieved from CyBase [38], and aligned by JalView [93, 94] with ProbCons [95]. The figure was generated using WebLogo 2.8.2 (available online at: https://weblogo.berkeley.edu/logo.cgi) [96].

1.2.2 Functions of PALs and AEPs in Plants, Mammals, and Parasites

PALs and AEPs are responsible for various biological processing *in vivo*, include the post-translational processing of peptide and protein precursors, such as seed storage proteins and cyclic peptides, and enzymes, such as hydrolytic enzymes in the plants and digestive enzymes in parasites (**Table 2**) [91].

Functions of PALs and AEPs can be categorized into four groups, the hydrolysis of seed storage proteins, the programmed cell death (PCD), the processing of RiPPs, and the circular permutation of lectin [92].

Expression of some plant AEPs is strictly controlled in a certain part of the plants because they are responsible for the developmental cell death during embryogenesis. During seed development, the seed storage protein precursors are subjected to processing by AEPs, leading to the production of mature proteins such as albumins and globulins. AtAEPδ (or δVPE) was only transiently expressed in the inner integument during the early stage of *Arabidopsis thaliana* seed development, and AtAEPδ deficiency resulted in delayed seed coat formation [93]. It is also important to note that when the vegetative AEPs, such as AtAEPδ, are absent in the *Arabidopsis* mutants, AtAEPβ (or βVPE) can compensate for their activities [30], and the seed mutants lacking AtAEPα (or αVPE), AtAEPβ, and AtAEPδ, no mature seed storage protein was detected [32].

Table 2. List of peptides and proteins post-translationally processed by PALs and AEPs. Table adapted from [91].

Species	Organ	Name	Cleavage site	Accession no.
Abrus precatorius	Seed	Abrin-a	-CNPPN|ANQSP-	P28590
Acacia confusa	Seed	Trypsin inhibitor	-YCEGN|SDDES-	P24924
Actinidia chinensis	Fruit	Actinidin protease	-VKYNN|QNYPE-	AAK06862
Arabidopsis thaliana	Seed	2S albumin 1	-DDATN|PIGPK-	CAA80870
Arabidopsis thaliana	Seed	2S albumin 1	-DDMEN|PQGQQ	
Arabidopsis thaliana	Seed	2S albumin 2	-DDASN|PMGPR-	CAA80871
Arabidopsis thaliana	Seed	2S albumin 2	-DDIEN|PQGQQ	
Arabidopsis thaliana	Seed	2S albumin 3	-DDASN|PVGPR-	CAA80868
Arabidopsis thaliana	Seed	2S albumin 4	-DDASN|PIGPI-	CAA80869
Arabidopsis thaliana	Seed	2S albumin 4	-DDIEN|PQRRQ	
Arabidopsis thaliana	Seed	2S albumin 5	-DDVSN|PQQGK-	NP_200285
Arabidopsis thaliana	Seed	2S albumin 5	-EDDEN|PMGPQ	
Arabidopsis thaliana	Seed	12S gluobulin 1	-GRHGN|GLEET-	NP_199225
Arabidopsis thaliana	Seed	12S gluobulin 2	-HEIAN|GLEET-	NP_171884
Arabidopsis thaliana	Seed	12S gluobulin 3	-SPGGN|GLEET-	NP_194581
Avena sativa	Seed	12S globulin	-DQSFN|GLEEN-	1515394A
Bertholletia excelsa	Seed	2S albumin	-VEEEN|QEECR-	P04403
Brassica napus	Seed	2S albumin	-DDATN|SAGPF- -DDMEN|PQGPQ	XP_013688210

Canavalia ensiformis	Seed	Concanavalin A	-FPDAN|VIRNS-	CAA25787
			-TIDFN|AAYN|ADTIV-	
			-KLKSN|EIPDI-*	
Clitoria ternatea	Leaf	Cyclotide cter-M	-ICMKN|HIIAA-*	
Curcurbita maxima	Seed	2S albumin	-EVEEN|RQGRE-	
			-RGIEN|PWRRE	
Curcurbita maxima	Seed	11S globulin	-SESEN|GLEET-	AAA33110
Curcurbita maxima	Seed	Membrane protein 27-32	-FGNEN|RDKTK-	BAA06186
Curcurbita maxima	Seed	PV100	-GCGVN|QRHSP-	BAA34056
			-GRGED|VD|EVERR-	
			-EDDEN|QRDPD-	
			-RGGRD|DEDEN|QRDPD-	
			-RSRVN|QVAIR	
Glycine max	Seed	Glycinin A1aBx	-KSRRN|GIDET-	1309256A
Glycine max	Seed	Glycinin A2B1a	-KRSRN|GIDET-	1402179A
Glycine max	Seed	Glycinin A3B4	-CQTRN|GVEEN-	1303273A
Glycine max	Seed	Glycinin A5A4B3	-QEQSN|RRGSR-	BAD72975
			-CETRN|GVEEN-	
Glycine max	Seed	Protease P34	-IKMAN|KKMKK-	P22895
Helianthus annuus	Seed	2S albumin HaG5	-VTESN|IDIPF-	P15461
Helianthus annuus	Seed	11S globulin HaG3	-GGWSN|GVEET-	P19084
Nicotiana alata	Leaf	Proteinase inhibitor II	-EEKKN|DRICT-	AAF14181

Nicotiana tabacum	Leaf	Endochitinase B	-RSFGN|GLLVD-	P24091
Nicotiana tabacum	Leaf	CBP20	-NCGDN|MNVLL-	AAB29959
Oldenlandia affinis	Leaf	Kalata-B1	-VCTRN|GLPSL-*	P56254
Oldenlandia affinis	Leaf	Kalata-B3/B6	-ICTRD|GLPKR-* -ICTRD|GLPSL-*	P58455
Oldenlandia affinis	Leaf	Kalata-B2	-ICTRD|SLPMR-* -ICTRD|SLPMS-*	P58454
Oryza sativa	Seed	Lectin	-GRNAN|GELCP-	XP_015633956
Phaseolus vulgaris	Seed	a-Amylase inhibitor	-HRQAN|SAVGL-	P02873
Pisum sativum	Seed	Legumin A	-RQGDN|GLEET-	P02857
Pisum sativum	Seed	Legumin B	-EERKN|GLEET-	P05692
Pisum sativum	Seed	Vicilin	-QGKEN|DKEEE-	CBK38920
Pisum sativum	Seed	Lectin	-LEEEN|VTSYT-	P02867
Ricinus communis	Seed	2S albumin	-QTRTN|PSQQG- -RRSDN|QERSL	S11500
Ricinus communis	Seed	Ricin	-VPNFN|ADVCM-	P02879
Ricinus communis	Seed	Agglutinin	-VPNFN|ADVCM-	XP_002534220
Solanum lycopersicum	Leaf	Proteinase inhibitor I	-EFDSN|LMCEG-	P05118
Vigna mungo	Leaf	Protease SH-EP	-GSKVN|HHHKM-	CAA33753

Cleavage sites with a star (*) indicate peptide ligation.

The vegetative AEPs are responsible for various functions, including programmed cell death [91, 94], which is modulated by caspase in animals. In tobacco (*Nicotiana benthamiana*), it was found that silencing the *VPE* gene reduces the virus-induced hypersensitive cell death [95]. During pathogen-induced and AEP-dependent programmed cell death, vacuolar collapse, and the following cell death were not observed in virus-infected and AEP-deficient plants [95].

AEPs play important roles in the biosynthesis of cyclic and circular-permutated peptides, such as Concanavalin A [96, 97], Sunflower Trypsin Inhibitor 1 (SFTI-1) [98], Kalata B1 [99], and McoTI-I [99]. As these post-translationally modified peptides frequently possess antimicrobial and insecticidal activities to maintain the plant immune system, AEPs are essential for plant defense. For instance, Kalata B1 possesses antimicrobial activity, and *Helicoverpa punctigera* larvae fed on Kalata B1 were shown to have impaired growth and development [100].

The post-translational modifications of peptides by AEPs were first reported in jack bean (*Canavalia ensiformis*) for the processing of the lectin that binds to mannose and glucose, Concanavalin A [101]. The maturation of Concanavalin A was previously reported to be post-translational and requires both peptide cleavage and transpeptidation [102, 103]. In 1993, AEPs from the mature seed of jack bean (*Canavalia ensiformis*) were purified and proposed to be responsible for the maturation of Concanavalin A [104]. Later, Min and Jones successfully obtained the mature Concanavalin A by jack bean AEP-catalyzed *in vitro* splicing, during which the precursor is cleaved, and the cleaved sequences are then shuffled and ligated together (**Figure 14**) [96].

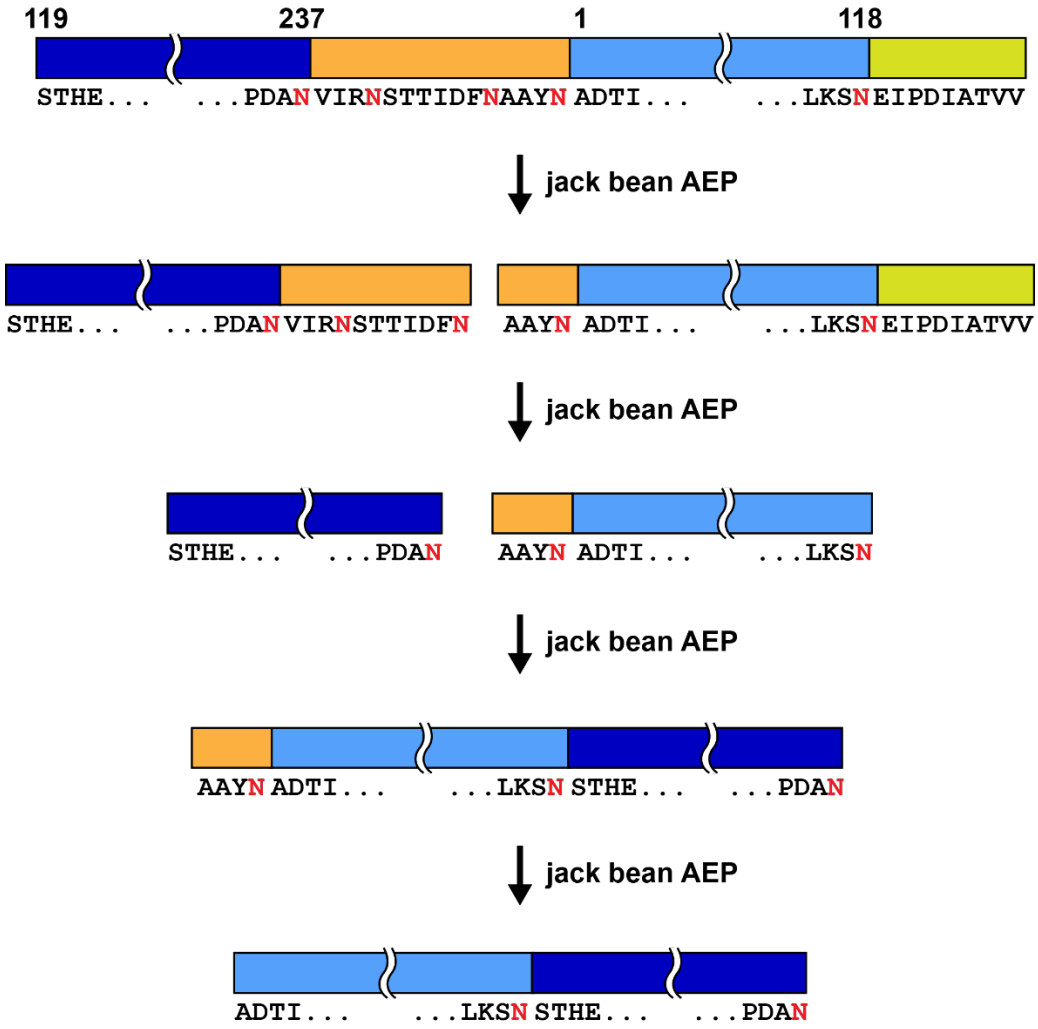


Figure 14. Step-wise maturation of ConA by jack bean AEP-mediated protein splicing. ConA was first cleaved at multiple sites, following the shuffling and ligation of two fragments. Figure adapted from [96].

Mammalian AEP was first discovered and sequenced in humans, isolated and characterized from pig kidneys [105]. In mammals, AEPs are important for the immune system. AEPs localized to the lysosomes of B cells are involved in the processing of antigen and the subsequent display of antigen on the MHC II complex [106]. Watts and coworkers first noticed that inhibition of lysosomal cathepsins, the proteases involved in many cellular and physiological functions [107], does not abolish the digestion of tetanus toxin antigen (TTCF), indicating the existence of other proteases [108]. Hinted by the activity and substrate specificity of AEP purified from pig kidney [105], which is P1-Asn-specific, they found that it is hAEP, or human legumain, the enzyme responsible for the cleavage of TTCF, and inhibition of hAEP leads to inhibited TTCF processing *in vitro* and slower TTCF presentation to T cells *in vivo*.

In addition to the immune system, AEPs are also involved in many physiological functions, such as bone remodeling, and diseases, including cancer and Alzheimer's disease [109]. It was first reported in 2003 by *Liu et al.* that AEPs are overexpressed in multiple human solid tumors, and overexpression of AEPs may promote metastasis, tissue invasion, and progelatinase A activation [110]. Inhibition of AEPs thus becomes one of the strategies for anti-cancer therapy.

AEPs are commonly found in the parasites, such as ticks, the ectoparasites feeding on blood, and were first reported in the *Schistosoma enzyme* [111]. In hard tick *Ixodes ricinus*, the AEP was found to localize to the digestive vesicles in the gut, and the AEP was termed IrAE. Profiling of the peptidase activities of *Ixodes ricinus* gut tissue extract showed that IrAEs are

involved in the proteolysis of host hemoglobin [112]. The recombinantly expressed IrAE1 was shown to digest human hemoglobin efficiently at pH 4.5 by cleaving after the Asn residue [113]. Similarly, recombinantly expressed AEP from the gut of *Haemaphysalis longicornis*, termed H1Lgm2, was shown to digest the blood proteins *in vitro* [114]. In *Schistosoma mansoni*, Sajid *et al.* found that the cathepsin B-like endopeptidase (SmCB1), a hemoglobin processing enzyme in the tick gut, is *trans*-activated by the AEP, termed SmAE, at pH 5.5 [115]. In short, in the guts of ticks, the AEPs have two major functions reported hitherto, digestion of the host hemoglobin and activation of other proteases in the guts (**Figure 15**) [116].

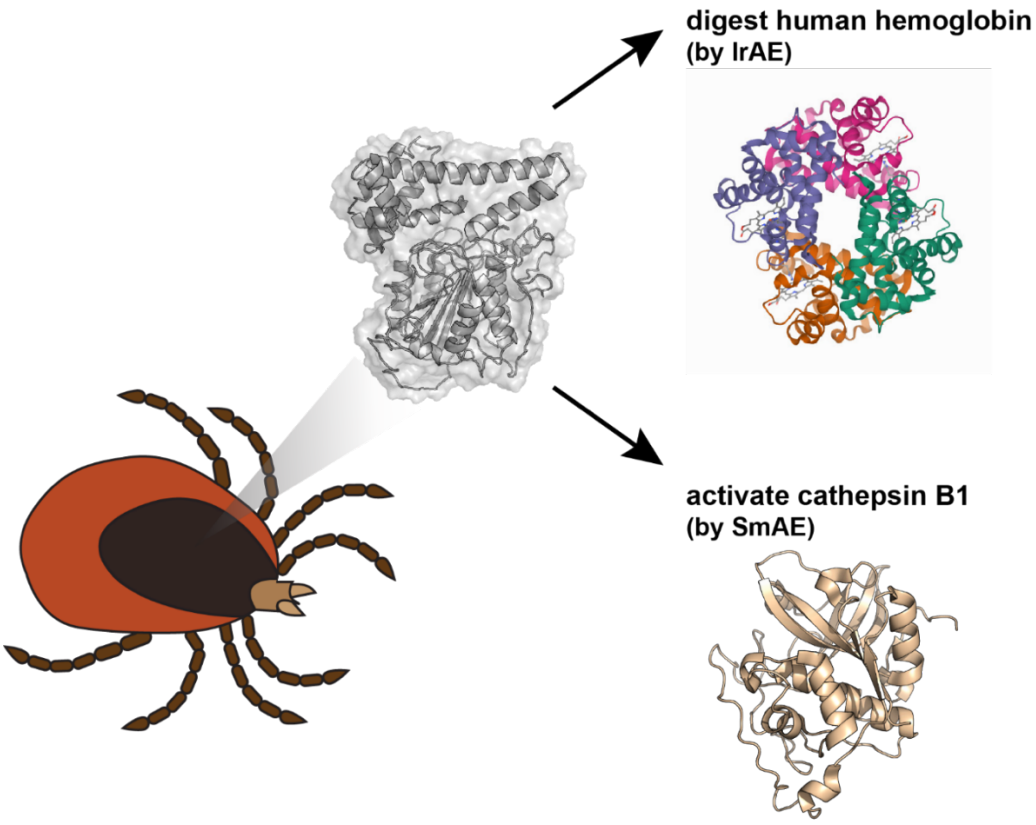
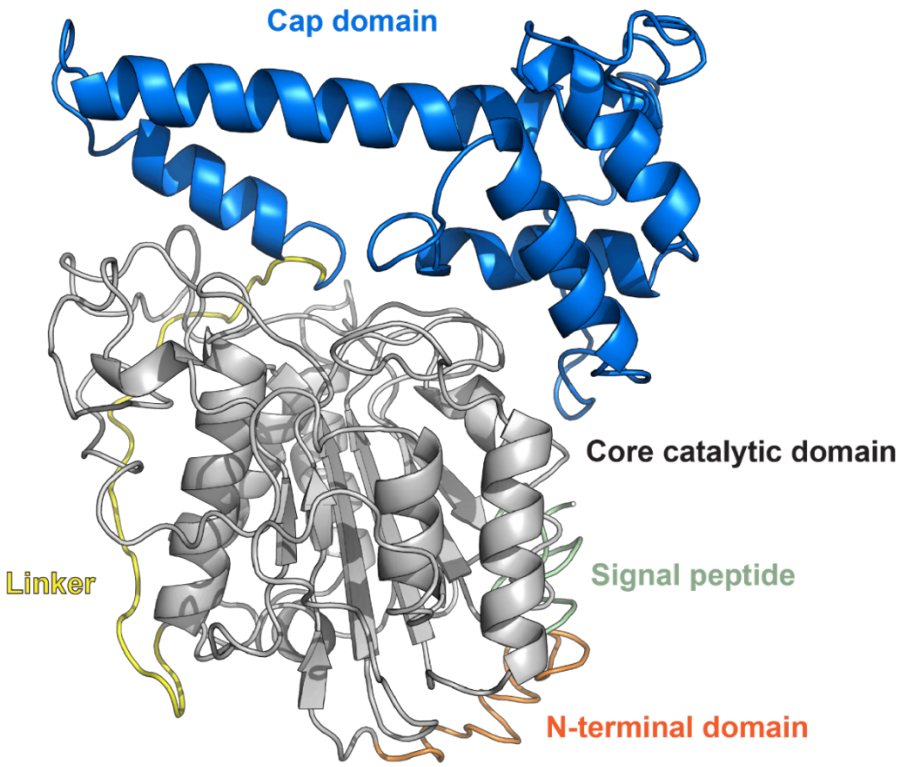


Figure 15. The function of AEPs in hard ticks. IrAE digests the host hemoglobin in *Ixodes ricinus* and SmAE activates the cathepsin B1 in *Schistosoma mansoni*. The structure of IrAE was modeled by I-TASSER [117-119]. The PDB accession codes for Hemoglobin and cathepsin are 2DHB [120] and 3S3Q [121], respectively.

1.2.3 Overall Architecture of PALs and AEPs

The overall architecture of PALs and AEPs is revealed by several crystal structures from different groups [10, 122-131]. AEPs are synthesized as an inactive zymogen containing a signal peptide, an N-terminal domain, core catalytic domain, a flexible linker, and the cap domain at the C-terminus, which can be further divided into the legumain stabilization and activity modulation (LSAM) domain and activation peptide (AP) (**Figure 16**). The core domain exhibits a caspase-like Triose phosphate isomerase (TIM)-barrel fold and contains six-stranded β-sheet flanked by five major α-helices, covered by two antiparallel β-sheets. The C-terminal cap domain is death domain-like [109] and contains six α-helices, it connects to the core domain by the flexible linker [124].

Through crystal structure, X-ray scattering (SAXS), and size exclusion chromatography (SEC) elution profile, it is revealed that the AEP zymogens form homodimer at neutral pH [124, 127]. An antiparallel helix bundle formed by four helices (α6 and α7) of the C-terminal cap domains of two AtAEPγ is proposed to mediate the dimerization, and the stabilization of the dimer is postulated to be facilitated by a plant-specific poly-proline loop (PPL) between the two cysteines, Cys252 and Cys266 according to AtAEPγ numbering, of the S3 substrate-binding pocket [127]. Apart from dimerization, the AEP zymogen is also stabilized by two plant-specific salt bridges.



MKNPLAILFLIATVVAVVSGIRDDFLRLPSQASKFFQADDNVEGTRWAVLVAGSKGY
VNYRHQADVCHAYQILKKGGLKDENIIVFMYDDIAYNESNPHPGVIINHPYGSDVYK
GVPKDYVGEDINPPNFYAVLLANKSALTGTGSGKVLDSGPNDHVFIYYTDHGGAGVL
GMPSKPYIAASDLNDVLKKKHASGTYKSIVFYVESCESGSMFDGLLPEDHNIYVMGA
SDTGESSWVTYCPLQHPSPPPEYDVCVGDLFSVAWLEDCDVHNLQTETFQQQYEVVK
NKTIVALIEDGTHVVQYGDVGLSKQTLFVYMGTDPANDNNTFTDKNSLGTPRKAVSQ
RDADLIHYWEKYRRAPEGSSRKAEAKKQLREVMAHRMHIDNSVKHIGKLLFGIEKGH
KMLNNVRPAGLPVVDDWDCFKTLIRTFETHCGSLSEYGMKHMRSFANLCNAGIRKEQ
MAEASAQACVSIPDNPWSSLHAGFSV

Figure 16. Domain architecture of the prototypic PAL butelase-1. The enzyme can be distinguished into five domains, they are the signal peptide (light green), the N-terminal domain (orange), the core catalytic domain (grey), the linker (yellow), and the C-terminal cap domain (blue). PDB accession code of butelase-1: 6HDI [132].

The substrate-binding sites of enzymes were originally defined and divided into seven subsites, S4-S3' substrate-binding pocket, by Schechter and Berger in 1967, and each subsite can be composed of one or more residues but only accommodates one amino acid of the substrate [36]. The substrate is cleaved between the P1 position and the P1' position, which bind to the S1 and S1' pocket of the enzyme, respectively. The position at the N-terminal side of the P1 position is termed the P2 position, and the position at the C-terminal side of the P1' position is termed the P2' position. The primed side of the substrate, such as the P1' and the P2' position, is the leaving group, which is released after the cleavage (**Figure 17**).

The substrate-binding pockets of PALs and AEPs were defined based on the crystal structure in complex with inhibitor. In 2018, the crystal structure of one of the AEP isoforms from *Arabidopsis thaliana* AtAEPγ in complex with the peptide inhibitor Ac-YVAD chloromethyl ketone (CMK) was solved [128]. Together with the docking of SFTI, Zauner *et al.* identified the substrate-binding pockets and an oxyanion hole, formed by amide nitrogen of Cys219, amide nitrogen of Gly178, and imidazole ring of His177 (AtAEPγ numbering), of AtAEPγ (**Figure 18**).

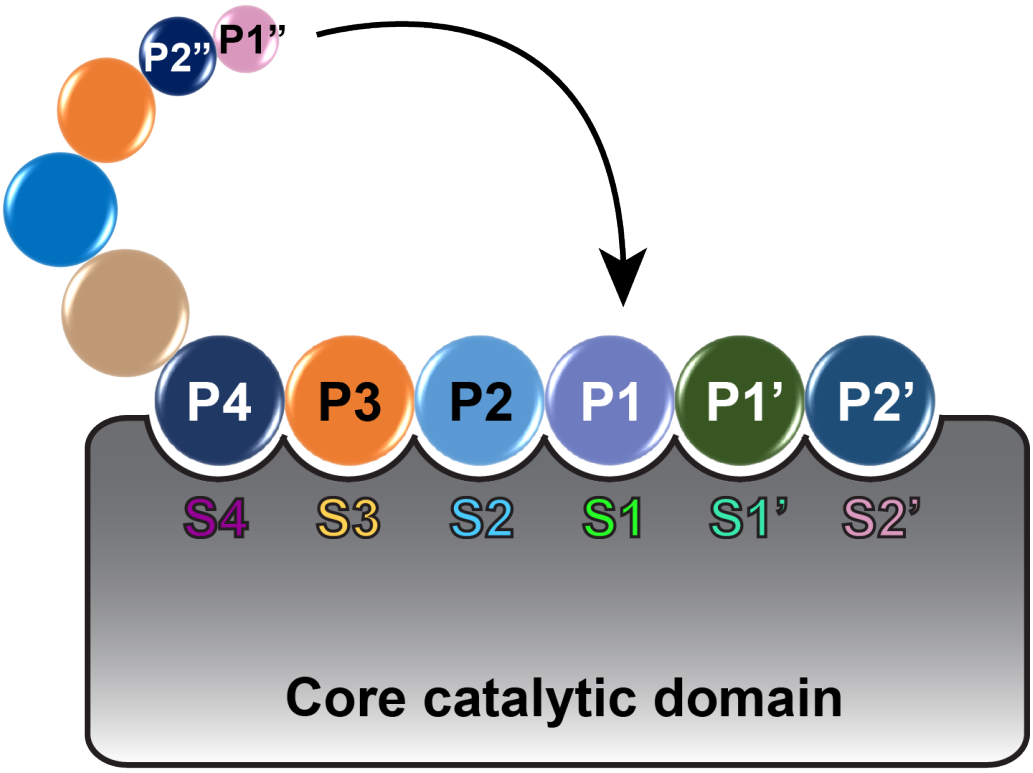


Figure 17. Substrate-binding pockets of PALs and AEPs. Nomenclature according to Schechter and Berger [36]. The enzyme cleaves between the P1 and P1’ position of the substrate, following the ligation of residues of the P1 position and P1” position. The residue at the P1 position of the substrate is an Asx for PALs and AEPs.

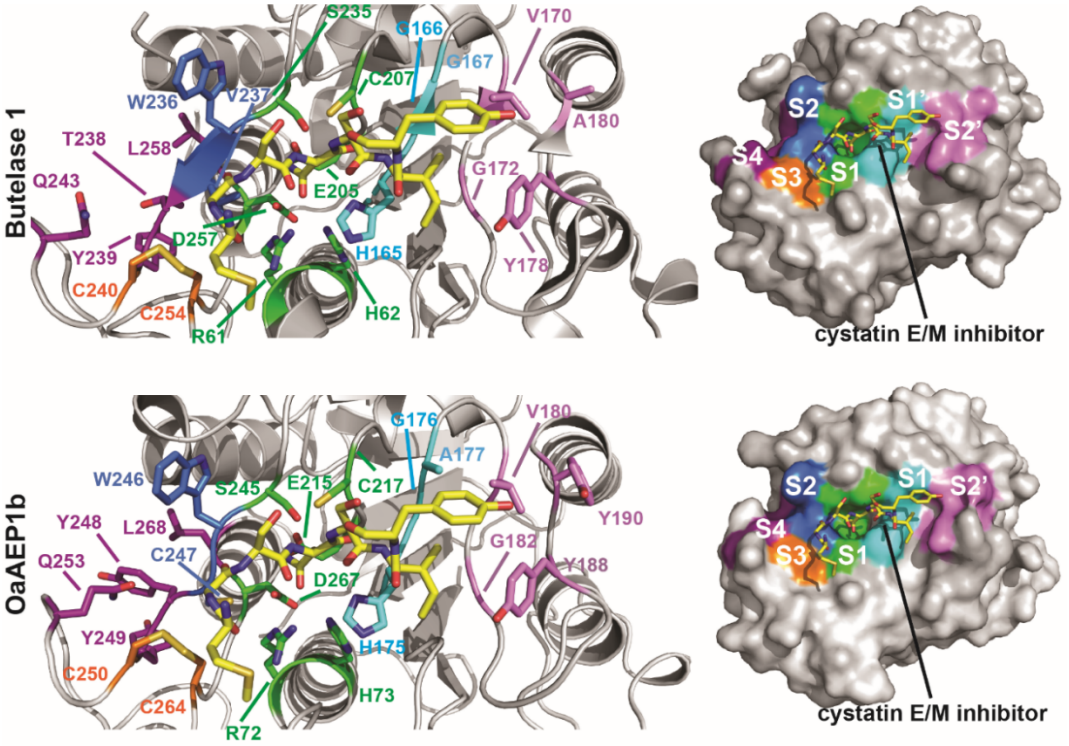


Figure 18. Superimposition of substrate binding pockets of butelase-1 and OaAEP1b. AtLEGγ with the cystatin E/M inhibitor (PDB accession code: 4N6N) [125] was superimposed with butelase-1 (PDB accession code: 6DHI) [132] and OaAEP1b (PDB accession code: 5H0I) [126] using PyMol Molecular Graphics System, Version 2.0 Schrödinger, LLC. The ribbon representations of the core catalytic domain of butelase-1 and OaAEP1b with the residues lining the substrate-binding pockets labeled are at the left. The surface view of the substrate-binding pockets using the same color code for each specificity pocket is shown on the right of the ribbon representations.

1.2.4 Autocatalytic Activation and Subsequent Conformational Change of PALs and AEPs

To unmask the catalytic residues of the surface and the substrate-binding pockets and allow the enzyme to function, the C-terminal cap domain shielding the catalytic sites of the enzyme has to be removed (**Figure 19**). Under neutral pH, the positively charged interface of the C-terminal cap domain neutralizes the negatively charged surface of the catalytic core domain [124, 127]. At pH 4.0, the catalytic surface and α6 helix of the C-terminal cap domain are protonated and electrostatic repulsion facilitates the dissociation of the catalytic core domain and cap domain [124, 127].

Frequently, the C-terminal cleavages sites, which can be heterogeneous, are located at the flexible linker and the beginning of the C-terminal cap domain [123, 124, 127, 133]. In contrast, although the N-terminal domain will be cleaved during activation as well, its removal is not essential for the activation of AEP [122].

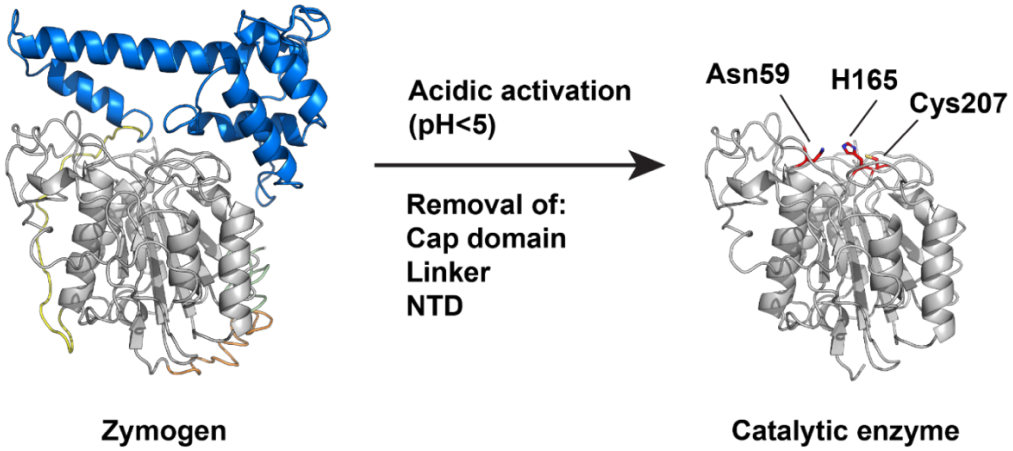


Figure 19. Activation of the enzyme zymogen. Lowering the pH results in the removal of cap domain and part of the linker, exposing the catalytic residues (colored in red, numbering according to butelase-1). The structure of butelase-1 (PDB accession code: 6HDI) was visualized by PyMol Molecular Graphics System, Version 2.0 Schrödinger, LLC.

The PALs and AEPs can be autoactivated in acidic pH and the activation process does not render a significant conformational change of the core catalytic form [123, 124, 127]. By incubating the recombinantly expressed hAEP (or human legumain) at 30 °C for 4 h at pH 4.5, the hAEP can be converted to the intermediate form of 46 kDa. Extending the incubation time to 24 h led to the conversion of the 46 kDa active intermediate into the 36 kDa mature form, and the activity of the two species was shown to be similar [133]. Sequencing results of the proenzyme (56 kDa) and the intermediate (46 kDa) showed that the two forms both contain the N-terminal domain, which is, however, found to be removed in the 35 kDa active AEP purified from pig kidney [105]. The results suggest that the removal of the N-terminal domain is not essential for the enzymatic activity of AEP.

It was demonstrated that the autolytic activation of hAEP is stepwise. Watts and colleagues reported that removal of part of the N-terminal domain and the C-terminal cap domain is essential for the enzymatic activity of hAEP [134]. Later, Dall *et al.* further characterized the pH-dependent AEP activation by activating hAEP at different pH, which resulted in different levels of enzymatic activity and different sizes of the intermediates [122]. Lowering the pH to 5.5 for 20 h resulted in the removal of part of the C-terminal cap domain and the intermediate showed no activity. Further lowering the pH to 5.0 for 20 h led to no additional cleavage but the enzyme intermediate started to show activity, which is proposed to be the result of conformational changes of the enzyme at pH 5.0. Incubating the hAEP at pH 5.0 to 40 h and at pH 4.5 for 20 h both led to the release of the N-terminal domain. Finally, lowering the pH to 4.0 led to an additional cleavage at the C-terminus cap domain and higher

enzymatic activity of hAEP. Mutating the C-terminal Asp303 and Asp309 to glutamine (D303E/D309E) hindered the pH 4.0-dependent conversion of hAEP. However, the mutations (D303E/D309E) did not influence the activity, suggesting that the enzymatic activity increment is unrelated to the additional cleavage and may be a consequence of conformational changes [122].

Notably, activation of AEPs is reversible and the catalytic domain can be linked back to the C-terminal cap domain through re-ligation. Zhao *et al.* discovered that the activated AEP can be reversed to the zymogen form during the crystallization at pH 8.5. By incubation the activated AEP in buffers from pH 4.5 to 7.5 overnight at 16 °C, it is revealed that the separated domains of AEPs are able to re-ligate themselves in a pH-dependent manner, and the interaction between the core catalytic domain, the C-terminal cap domain, and the N-terminal domain, is proposed to be non-covalent through hydrogen bonds and salt bridges [124].

Activation of the PALs and AEPs *in vitro* can occur both *in cis* and *in trans*. Hiraiwa *et al.* mutated the catalytic Cys222 of RcAEP from castor bean *Ricinus communis*, and the RcAEP-C222G dead mutant showed no protease or ligase activity, suggesting that the catalytic residues are essential for the enzymatic activity of PALs and AEPs. However, mixing the recombinantly expressed RcAEP-C222G mutant with extracted native RcAEP (37 kDa) resulted in the conversion of the inactive RcAEP-C222G mutant to the putative active form of approximately 42 kDa [135]. Dall *et al.* also *trans*-activated a dead mutant, hAEP-C189S, using activated wild-type hAEP. Interestingly, they discovered that the trimming after Asp303/Asp309 of the C-terminal cap domain of hAEP only occurs *in cis* [122].

1.2.5 The Broad Substrate Specificities and Ligase Activity of PALs and AEPs

PALs and AEPs both exhibit exquisite site-specificity toward an Asx residue at the P1 position of the substrate. It was demonstrated that PALs and AEPs both prefer substrates containing P1-Asn compared to that containing P1-Asp. The prototypic PAL, butelase-1, was shown to cyclize substrate containing P1-Asp (GLPVCGETCVGGTCNTPGCTCSWPVCTRDHV, the P1-Asp is underlined) with a 10,000 lower efficiency compared to its P1-Asn counterpart [6]. It was observed and reported by Brandstetter and coworkers that C-terminal Asp residues were preferably cleaved during autolytic activation of human legumain at pH 4.0 [122], indicating that optimum pH for reaction could be substrate-dependent. It was proposed that lower pH leads to the protonation of Asp, which makes it resembles an Asn, allowing the binding of Asp to the substrate-binding pocket S1 [109, 123].

Frequently, a minimum of two residues after the P1-Asx position is sufficient for PALs and AEPs to render ligation reactions efficiently. Butelase-1 was shown to cyclize the substrates with two to four residues after the P1-Asx as leaving group of the substrate with >95% yield. In contrast, for the substrate with only one or no residue after the P1-Asx, the butelase-1-mediated cyclization became sluggish and only yield less than 10% cyclic products [6]. Similarly, AtLEGβ was also able to carry out cyclization and hydrolysis with or without the primed side residues [130]. Notably, for P1-Asp-containing peptide substrates, the cyclization efficiency is higher with the presence of the leaving group (P1'-P2' residue), however, for the P1-Asn-containing peptides,

the presence of the leaving group does not influence the cyclization yield and substrate distribution [130].

Enzymes from different species usually possess different substrate preferences, which can be hinted by the presence of different cyclotide precursors in the plants. For example, the preferred tripeptide recognition signal of VyPAL2 from *Viola yedoensis* is Asn-Ser-Leu, which is commonly found in the cyclotides from the Violaceae family [38].

At the P1' position, VyPAL2 accepts almost all natural amino acids and prefers small residues, such as Gly and Ser at pH 6.5. A P1'-Pro is unfavored by VyPAL2-mediated cyclization. It was proposed that Ala located at the center of S1 pocket of VyPAL2, unlike the Gly at the central position of substrate-binding pocket S1 of butelase-1, may not tolerate larger amino acids at P1' position [136].

At the P2' position, hydrophobic and aromatic residues, such as Leu, Ile, and Phe, are favored by VyPAL2 [10], which may be a result of the van der Waals interactions and cation-π interactions. The S2'-Lys of VyPAL2 is positively charged and possesses a long aliphatic side chain, which may attract residues of large aliphatic side chains, such as Ile and Leu, at P2' position of the substrate through van der Waals interactions [136]. Meanwhile, the Lys residue may interact with the P2'-Phe through cation-π interactions, which may be the reason why a Phe is preferred by VyPAL2 at the P2' position of the substrate. For butelase-1, the residue at the center of the substrate-binding pocket S2' is a Val, which may make butelase-1 prefer amino acids with bulky aliphatic side chain for van der Waals interactions [136].

It is clearly demonstrated that butelase-1 and VyPAL2 have very different preferences towards the primed side of the peptide substrates. The cyclization efficiency of VyPAL2 against substrate containing the C-terminal Asn-Ser-Leu-Ala-Asn (P1-P1'-P2'-P3'-P4') motif is only 3.5-fold less than butelase-1. However, the catalytic efficiency of butelase-1 using peptide substrate containing a C-terminal Asn-His-Val tripeptide motif was shown to be 18.5 times more efficient than that of VyPAL2 [136]. VyPAL2 has a low activity towards the peptide substrates containing the Asn-His-Val tripeptide motif at the C-terminus compared to butelase-1 [10, 136], and this substrate preference disparity between butelase-1 and VyPAL2 makes them exquisite candidates for one-pot and sequential bio-orthogonal ligation.

The primed side of the substrate-binding pockets of PALs and AEPs is not only important for the substrate preference but also the substrate distribution. Through molecular dynamics simulations, Brandstetter and coworkers proposed that the interactions of the enzyme and substrate at the primed sides play key roles in preventing the pre-mature hydrolysis of the thioester bond [128]. The retention time of different leaving groups (P1'-P2' position) was tested, and it was found that the leaving group His-Val and Gly-Leu were still bound to the substrate-binding pocket S1'-S2' after 330-ns molecular dynamic simulation. On the contrary, leaving groups Gly-Gly and Gly-Ser were found to have a shorter retention time, less than 3 ns, at the S1'-S2' pockets of the enzyme, which allows the catalytic water to access the pre-mature thioester bond [128].

PALs and AEPs are both promiscuous for the incoming nucleophile (P1"-P2"-…). For the incoming peptides, butelase-1 accepts almost all natural

amino acids at the P1" position (the N-terminal end of the incoming peptide located next to P1 position and C-terminal to the cleavage site) when the P2" position (the position C-terminal to the P1" position) is an Ile, except Pro and acidic residues, such as aspartic acid and glutamic acid. At the P2" position, when the P1" position is a Leu, butelase-1 prefers Cys, Ile, Leu, and Val [6]. Similarly, OaAEP1b accepts all natural amino acids at the P2" position of the substrates.

Notably, although OaAEP1b ligated peptide substrates with Gly-Val (GVRL) and Gly-Leu (GLRL) at the P1"-P2" position to the peptide Fmoc-RWKGNGL with comparable efficiency, the ligated product Fmoc-RWKGNGVRL was poorly hydrolyzed by OaAEP1b. This result indicates that the dipeptide Gly-Val is a good incoming nucleophile (P1"-P2"), but not a good leaving group (P1'-P2') for OaAEP1b [137]. Rehm *et al.* then compared the efficiency of OaAEP1b-mediated ligation using substrates containing a C-terminal Asn-Gly-Val (P1-P1'-P2') and an Asn-Gly-Ile (P1-P1'-P2') motif. The catalytic efficiency of the substrate containing the Asn-Gly-Ile tripeptide (k_{cat}/K_m of 171 ± 96 $M^{-1}s^{-1}$) was 24.4 times more efficient than that containing the Asn-Gly-Val tripeptide (k_{cat}/K_m of 7.0 ± 6.5 $M^{-1}s^{-1}$), indicating that the low efficiency of OaAEP1b-mediated ligation is unique to substrates containing a P2'-Val [137]. Taking advantage of this discovery, a Gly-Val is preferred as the incoming group (P1"-P2") but not a leaving group (P1'-P2'), Rehm *et al.* were able to perform sequential ligation using OaAEP1b at both termini of a target protein, which is equipped with an N-terminal Gly-Val motif and a C-terminal Asn-Gly-Leu motif (**Figure 20**) [137].

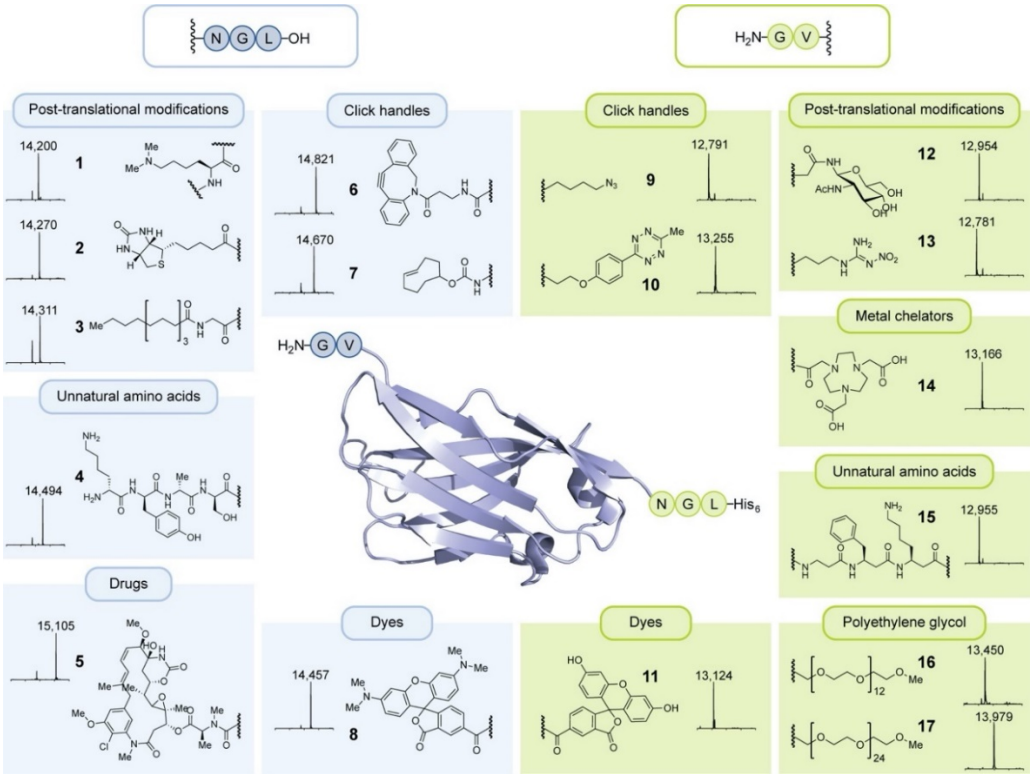


Figure 20. OaAEP1b-mediated N- and C-terminal dual-labelling of nanobody. The nanobody is equipped with an N-terminal Gly-Val motif and a C-terminal Asn-Gly-Leu motif. Reprinted with permission from [137]. Copyright © 2019 American Chemical Society.

1.2.6 The Sequence Motifs Indicating Ligase Activity of PALs and AEPs

As PALs and AEPs share similar sequences and structures, many attempts have been made to identify the structural and molecular basis or features that contribute to the ligase and protease activity of PALs and AEPs. Many motifs have been discovered hitherto and most of them located in close proximity to the substrate-binding pockets (**Figure 21**) [8, 10, 98, 126, 131].

Based on the crystal structure of OaAEP1b, Yang *et al.* proposed that the Cys247 of OaAEP1b is pivotal for the accommodation of the incoming nucleophile, and this residue was thus termed 'gatekeeper.' By mutating the gatekeeper residue to several amino acids, significantly increased ligation efficiency of the Cys247Ala mutant compared to the OaAEP1b was observed. It was proposed that Ala has a smaller side chain as compared to Cys and thus better accommodates the substrate, which resulted in a 160-fold increase of catalytic efficiency. In contrast, mutating the Cys247 to larger and more hydrophobic amino acids, such as Ile, Leu, Met, Val, led to significantly lower catalytic efficiency. However, mutating the Cys to even smaller Gly resulted in enhanced hydrolysis activity, which was proposed to be a result of the destabilization of local protein conformation [126].

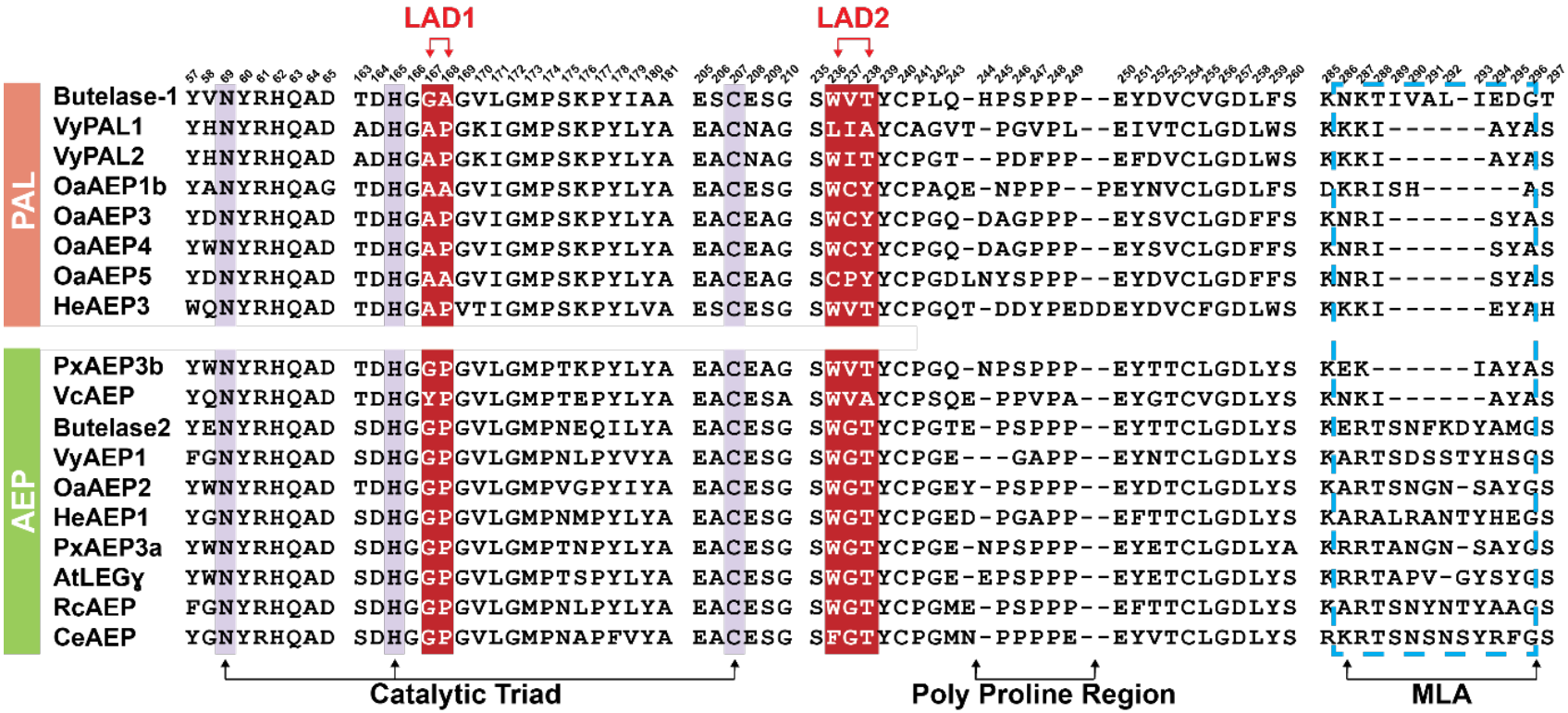


Figure 21. Multiple sequence alignment of reported butelase-1-like PALs and butelase-2-like AEPs. The sequences were aligned by Clustal Omega (available online at: https://www.ebi.ac.uk/Tools/msa/clustalo/) [138] using JalView [139, 140].

Later, the gatekeeper residue was found to be the middle residue of the substrate-binding pocket S2 [128] and the importance of the S2 pocket in modulating the ligase activity of PALs and AEPs was further demonstrated [10]. Hemu *et al.* noticed that among the isoforms of PALs and AEPs in the *Viola yedoensis*, those possessing a bulky and hydrophobic residue at the S2 pocket exhibit stronger ligase activity. In contrast, those isoforms having a Gly at the S2 pocket showed predominant hydrolysis activity [10]. Furthermore, by mutating the Gly at the S2 pocket of butelase-2, a prototypic protease, to a bulky and hydrophobic residue, such as Ile and Val, Hemu *et al.* were able to greatly enhance the ligase activity of butelase-2 [131]. The S2 substrate-binding pocket was termed 'ligase activity determinant 1 (LAD1)' for its potential role in binding and positioning the substrate and the enzymatic activity, which may affect the catalytic efficiency and the dual ligase-protease activity of PALs and AEPs [10].

Same as substrate-binding pocket S2, substrate-binding pocket S1' is also in proximity to the catalytic residues and the oxyanion hole, the S1' pocket was thus also proposed to play a role in modulating the ligase activity of PALs and AEPs. Unlike substrate-binding pocket S2, the interaction between residues of the S1' pocket and P1' residue of the substrate directly affects the access of the incoming nucleophile [10, 128, 131]. It was previously demonstrated by Brandstetter and coworkers that the retention time of the leaving group is important for displacing catalytic water at the primed side of the catalytic surface of PALs and AEPs. Hemu *et al.* later showed that replacing the bulky Tyr of the substrate-binding pocket S1' with Ala significantly suppresses the hydrolysis activity of VyPAL3 and VcAEP [10].

It was suggested that the Tyr-Ala dipeptide of VyPAL3 and Tyr-Pro dipeptide of VcAEP are bulky and may facilitate the departure of the leaving group, allowing the catalytic water to access the pre-mature thioester bond. Similarly, mutating the Gly-Pro dipeptide of butelase-2, a prototypic protease, to Ala-Ala also enhances the ligase activity of butelase-2 significantly [131]. In summary, it was suggested that the small and hydrophobic dipeptide at the S1' pocket, such as Gly-Ala of butelase-1, Ala-Pro of VyPAL2, and Ala-Ala of OaAEP1b, may increase the local hydrophobicity, retain the leaving group longer, and block the catalytic water from attacking the thio-enzyme intermediate. This dipeptide located at the substrate-binding pocket S1 was thus termed the 'ligase activity determinant 2 (LAD2)' [10, 131].

As the substrate-binding pockets flanking the substrate-binding pocket S1, S2 pocket, and S1' pocket, were both shown to be essential for modulating the ligase activity of PALs and AEPs, it is not surprising that attempts were also made to investigate the role of S1 pocket in ligase activity. It was reported that Haywood *et al.* mutating the Asn at the substrate-binding pocket S1 of HaAEP to Ala (Asn73Ala) increases the ligation activity of HaAEP. It was suggested that without the side chain of Asn73 the conformational flexibility of nearby catalytic His178 was increased, allowing the facilitation of deprotonation of incoming amine group as well as the reduction of steric hindrance of the incoming amine towards the thioester intermediate [141]. In contrast, mutating the Glu221, which is negatively charged, of the S1 pocket to Lys, which is positively charged, resulted in abolished ligase activity of HaAEP, suggesting that Glu221 plays a role in the deprotonation of the

incoming amine group or the nucleophilic attack of the thioester intermediate by the incoming amine group [141].

The search of structural and molecular features that affect ligase activity of PALs and AEPs is not restricted to the residues proximal to the catalytic residues and the oxyanion hole. A sequence motif located far away from the substrate-binding pocket S1, named 'marker of ligase activity (MLA),' was also reported to be essential for the ligase activity of PALs and AEPs [8]. By comparing the isoforms of PALs and AEPs in garden petunia (Petunia x hybrida E.Vilm.), Jackson et al. found a deletion of five amino acids at the MLA of PxAEP3b, a PAL. This deletion at MLA is not observed in the other isoforms that exhibit predominant hydrolysis activity, such as PxAEP3a. The same pattern was also observed in the isoforms in *Hybanthus enneaspermus* F.Muell. HeAEP3, which has a truncated MLA, showed higher ligase activity than its isoforms HeAEP1 and HeAEP2, which have a hydrophobic MLA. By deleting the residues at MLA of the proteases, PxAEP3a and OaAEP2, the ligase activity was largely increased [8].

1.3 Applications of Peptide Asparaginyl Ligases (PALs)

Ligation reaction allows peptide synthesis [142] and protein modification. However, it is restricted by arduous preparation of the protected segments, the size limitation of protein synthesis, and stringent conditions some chemical ligation techniques require [143-145]. Recombinant protein expression allows the production of peptides and proteins, however, it does not allow tailored post-translational modifications and the incorporation of unnatural amino acids.

PALs are thus attractive biochemical and biotechnological tools for peptide and protein ligation, which allows cyclization and installation of desired cargos to the targets. PAL-mediated ligation requires only a tripeptide recognition motif, which is very unlikely to disturb the overall structure of the recombinantly prepared peptide or protein substrates. Upon completion of the ligation by PALs, there is only one Asx left on the peptide or protein products, and the prototypic PAL, butelase-1 has been exploited in various applications [21]. Discovery of novel PALs with distinct substrate specificities and kinetics expands the catalog of available enzymes for applications, such as tandem ligation and one-pot ligation, which require the utilized enzymes to possess different substrate preferences.

1.3.1 Intramolecular Ligation

Cyclic peptides are attractive candidates of therapeutics that are able to target the nonadjacent protein binding sites (“PPis” protein-protein interactions). Cyclization of peptides and proteins enhances the stability towards proteolysis and heat, making them more feasible for therapeutical applications [146, 147].

Butelase-1 has been applied for the cyclization of peptides and proteins ranging from several amino acids to a few hundred amino acids with high yield and efficiency (**Table 3**) [6, 136, 148-152]. The versatility of butelase-1 as a cyclase is demonstrated in the cyclization of bacteriocins, the largest naturally occurring cyclic antimicrobial peptides [152]. The AS-48 and uberlolysin are both highly hydrophobic, containing more than half hydrophobic amino acids in their sequences. Butelase-1 is capable of cyclizing the AS-48 of 70 residues with >85% yield within 1 h (**Figure 22**).

Many D-antimicrobial peptides are more resistant to proteolytic degradation than their L-counterparts with similar bioactivity. With L-Asx and L-Gly at the P1 and P1" position, respectively, of the substrates, butelase-1 cyclized the all-D antimicrobial peptides with >95% yield within 1 h (**Table 3, example 2-4**) [151].

Table 3. Head-to-tail cyclization catalyzed by butelase-1. The underlined residues will be ligated to the N-terminal amino acid. Table updated and adapted from [21].

No	Substrate-HV	Sequence	Length (aa)	Time (min)	Yield (%)	Ref.
1	SFTI	GRCTKSIPPICFPN	14	45	>95	[6]
2	D-SFTI	Grctksippicfp**N**	14	60	>95	[151]
3	D-MrIA	Gvccgyklchpcag**N**	15	15	>95	[151]
4	D-θ-defensin analog	GvcrcicrrGfcrclcr**N**	18	60	>95	[151]
5	Thanatin	GSKKPVPIIYCNRRTGKCQRM**N**	22	240	59	[6]
6	Rat neuromedin	GIKYGVNEYQGPVAPSGGFFLFRPR**N**	26	5	>95	[153]
7	Kalata-B1	GLPVCGETCVGGTCNTPGCTCSWPVCTR**N**	29	45	>95	[6]
8	Apelin	GLVQPRGSRNGPGPWQGGRRKFRRQRPRLSHKGPMPF**N**	38	5	>95	[153]
9	Garvicin ML (GarML)	LVATGMAAGVAKTIV**N**AVSAGMDIATALSLFSGAFTAAGGIMALIKKYAQKKLWKLIAA	60	1440	>90	[152]
10	AS-48	MAKEFGIPAAVAGTVL**N**VVEAGGWVTTIVSILTAVGSGGLSLLAAAGRESIKAYLKKEIKKKGKRAVIAW	70	60	>85	[152]
11	UblA	LAGYTGIASGTAKKVVDAIDKGAAAFVIISIISTVISAGALGAVS	70	1440	>93	[152]

		ASADFIILTVKNYIS RNLKAQAVIW				
12	Z_{EGFR}	CGSSHHHHHHLQVDN KFNKEMWAAWEEIRN LPNLNGWQMTAFIAS LVDDPSQSANLLAEA KKLNDAQAPKVDGSG SNGFLGVKAN	87	30	>70	[136]
13	p53-binding domain (N-terminal domain) of murine double minute X (N-MdmX)	GLQINQVRPKLPLLK ILHAAGAQGEMFTVK EVMHYLGQYIMVKQL YDQQEQHMVYAGGDL LGELLGRQSFSVKDP SPLYDMLRKNLVTLA TN	92	40	>95	[148]
14	Human growth hormone (somatropin)	FPTIPLSRLFQNAML RAHRLHQLAFDTYEE FEEAYIPKEQKYSFL QAPQASLCFSESIPT PSNREQAQQKSNLQL LRISLLLIQSWLEPV GFLRSVFANSLVYGA SDSDVYDLLKDLEEG IQTLMGRLEDGSPRT GQAFKQTYAKFDANS HNDDALLKNYGLLYC FRKDMDKVETFLRIV QCRSVEGSCGFN	192	15	>85	[153]
15	IL-1Ra	GISYDYMEGGDIRVR RLFCRTQWYLRIDKR	143	15	>90	[153]

		GKVKGTQEMKNNYNI MEIRTVAVGIVAIKG VESEFYLAMNKEGKL YAKKECNEDCNFKEL ILENHYNTYASAKWT HNGGEMFVALNQKGI PVRGKKTKKEQKTAH FLPMAITN				
16	V44- DHFR	SIAGGVRPLNSIVAV SQNMGIGKNGDLPWP PLRNEFKYFQRMTTT SS-tag- EGKQNLVIMGRKTWF SIPEKNRPLKDRINI VLSRELKEPPRGAHF LAKSLDDALRLIEQP ELASKVDMVWIVGGS SVYQEAMNQPGHLRL FVTRMQEFESDTFFP EIDLGKYKLLPEYPG VLSEVQEEKGIKYKF EVYEKKGSRSGSGN	197	30	N.D.	[149]
17	GFP	GISMSKGEELFTGVV PILVELDGDVNGHKF SVSGEGEGDATYGKL TLKFICTTGKLPVPW PTLVTTLTYGVQCFS RYPDHMKQHDFFKSA MPEGYVQERTIFFKD DGNYKTRAEVKFEGD TLVNRIELKGIDFKE DGNILGHKLEYNYNS HNVYIMADKQKNGIK VNFKIRHNIEDGSVQ	242	15	>95	[153]

LADHYQQNTPIGDGP

VLLPDNHYLSTQSAL

SKDPNEKRDHMVLLE

FVTAAGITLGMDELY

K<u>N</u>

Underlined Asn indicates the cyclization site. Residues in lower-case are in the D-congifuration. N.D.: not determined.

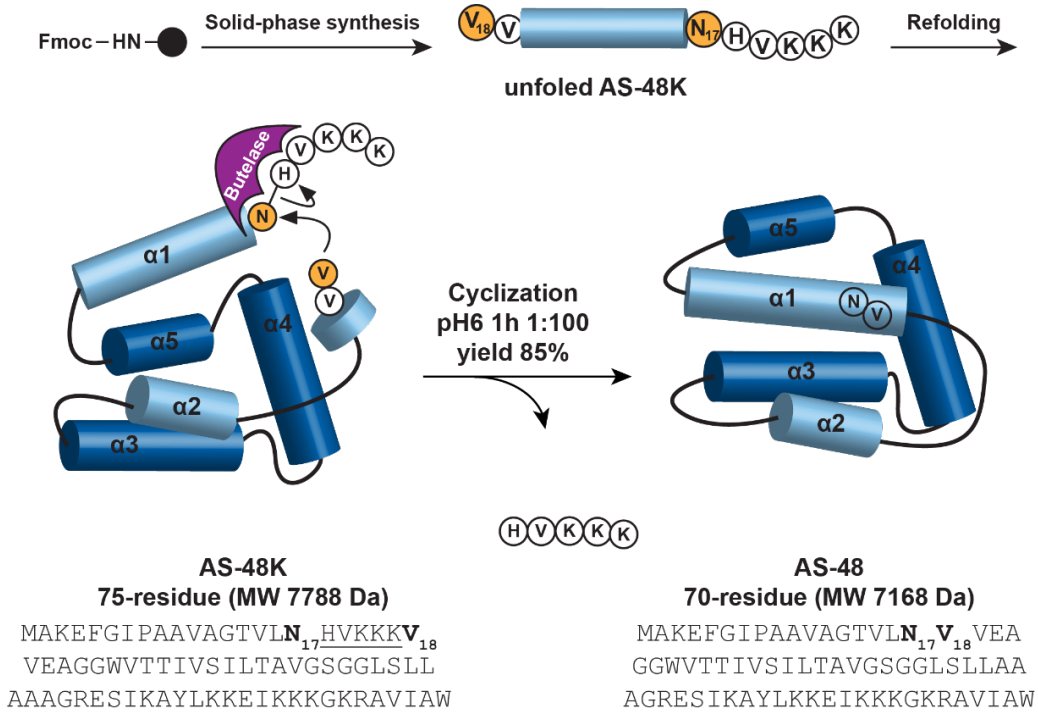


Figure 22. Total synthesis of the bacteriocin AS-48. The linear precursor, AS-48K, which contains triple Lys to enhance the solubility, was synthesized by Fmoc chemistry. The precursor was then cleaved, dissolved in 8 M urea, and purified by reverse-phase high-performance liquid chromatography (RP-HPLC). Lastly, the cyclization was mediated by butelase-1 within 1 h with 85% yield of the cyclic AS-48. Figure adapted from [152].

Frequently, PALs are applied for head-to-tail cyclization of peptides and proteins, they can also mediate side chain-to-tail cyclization. Yang et al. blocked the N-terminus of the peptide substrate, which allows proximity-driven attack of the thioester intermediate by the side chain of Lys (**Figure 23**) [126].

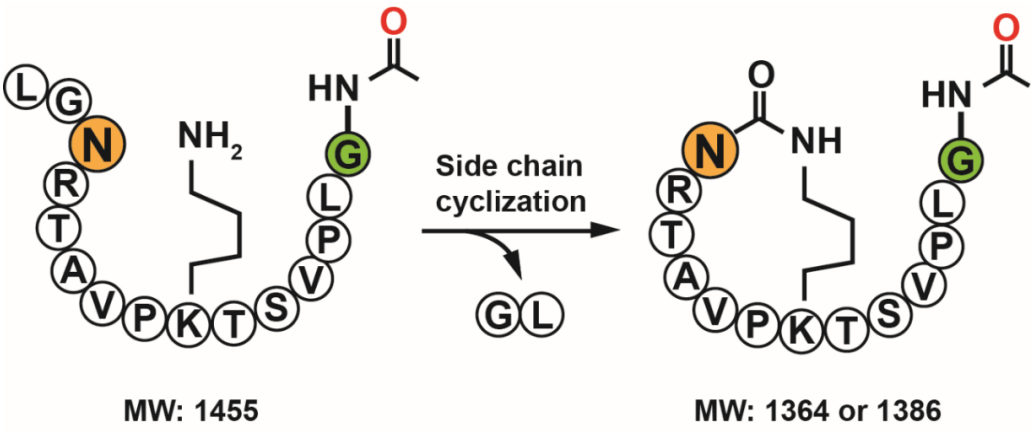


Figure 23. Side chain-to-tail cyclic peptide preparation by butelase-1. Butelase-1-mediated ligation of the Lys side chain and C-terminal Asn (colored in yellow) with the N-terminal amine of the substrate protected (colored in green). Figure taken from [21].

1.3.2 Precision Bioconjugation

Modifications of peptides and proteins expand the chemical and structural repertoire by the installation of the cargo of interest that confer desirable function, such as payload and functional groups, to the peptides and proteins with the canonical 20 amino acids at precise sites. By equipping the peptides and proteins of interest with recognition motifs (such as a tripeptide motif NHV for butelase-1), butelase-1 can catalyze N- (**Figure 24**) or C-terminal (**Figure 25**) modification with high efficiency (**Table 4**). Due to its high specificity, butelase-1 and other PALs can be applied for modification of one specific target in the mixture and live-cell labeling. On the other hand, using broad-spectrum ligases, such as subtiligase, may result in off-target modifications.

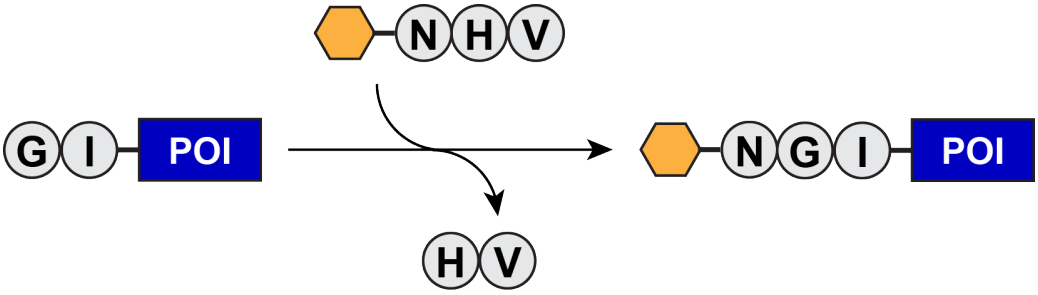


Figure 24. Schematic representation of N-terminal protein modification by butelase-1. Equipping the protein of interest and cargo of interest with an N-terminal Gly-Ile dipeptide and a C-terminal Asn-His-Val motif, respectively, allows the butelase-1-mediated N-terminal modification. The yellow hexagon represents the cargo of choice. The blue rectangle represents the protein of interest (POI). The grey circles represent the amino acids.

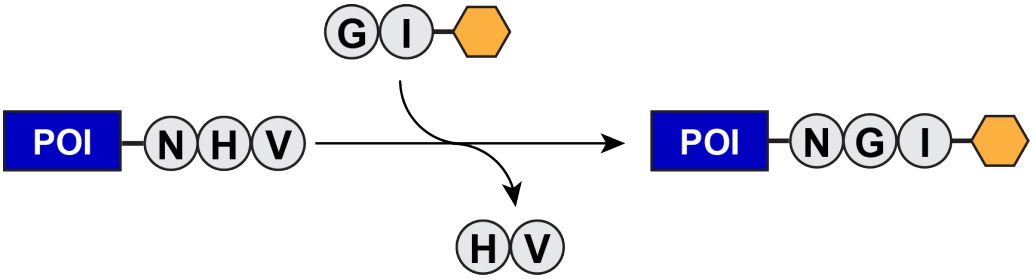


Figure 25. Schematic representation of C-terminal protein modification by butelase-1. Equipping the cargo of interest and protein of interest with an N-terminal Gly-Ile dipeptide and a C-terminal Asn-His-Val motif, respectively, allows the butelase-1-mediated C-terminal modification. The yellow hexagon represents the cargo of choice. The blue rectangle represents the protein of interest (POI). The grey circles represent the amino acids.

Table 4. Intermolecular ligation catalyzed by butelase-1. Table updated and adapted from [21].

	N-terminal Substrate	**C-terminal Substrate**	**Ref.**
N-terminal labeling	YKN-thioglc-V	GIGGIR	[154]
	Biotin-TYKN-thioglv-V	GI-Ubiquitin	[154]
	YKN-thioglv-V	MI-GFP	[154]
C-terminal labeling		GIGKFLHSAKKFG	[155]
	DARPin-NHV	KAFVGEIMNS	
		RIGK- fluorescein amidites	[155]
	DARPin-NHV		
Live-cell labeling	OmpA-NHV	GIGGIRK-fluorescein	[156]
		GIGGIRK-Biotin	[156]
		Monoglycated Muc1-like peptide with C-terminal Biotin	[156]
	TfR1-NHV	GI-FRET-based probes	[157]
Peptide/protein ligation	KALVINHV	XIGGIR (X = 20 natural amino acids)	[6]
		GXGGIR (X = 20 natural amino acids)	[6]
Antibody-fluorescence conjugate	IgG1 γ1-NHV	AL-Alexa (AlexaFluor 647 fluorescent dye)	[158]
Protein-oligonucleotide-protein conjugate	VHH-Enh-NHV	GV-Fmoc-DNA	[158]
Protein-protein conjugate	VHH-Enh-NHV	Two-headed PEG linker (NH2-GGG-PEG-AL-NH2)	[158]

Dendrimer conjugation	Ac-RYRLN-thioglc-V	(RIβA)2KY	[159]
		(RIβA)4K2KY	[159]
		(RIβA)8K4K2KY	[159]
Peptide/protein-thioester preparation	YXN-NHV	YXNG-COSR (X = V, L, S, F, Nle, d-A)	[160]
	Ubitquitin-NHV	GGMQIFVKTLTG KTITLEVEPSDTIE NVKAKIQDKEGIP PDQQRLIFAGKQL EDGRTLSDYNIQK ESTLHLVLRLRGG NXNG-COSR	[160]
	DARPin(ERK)-NHV	SMGSDLGKKLLE AARAGQDDEVRI LMANGADVNAH DDQGSTPLHLAA WIGHPEIVEVLLK HGADVNARDTDG WTPLHLAADNGH LEIVEVLLKYGAD VNAQDAYGLTPL HLAADRGHLEIVE VLLKHGADVNAQ DKFGKTAFDISID NGNEDLAEILQKL NKNXNG-COSR	[160]
	GFP-NHV	MSKGEELFTGVV PILVELDGDVNGH KFSVSGEGEGDA TYGKLTLKFICTT GKLPVPWPTLVT TLTYGVQCFSRYP DHMKQHDFFKSA MPEGYVQERTIFF	[160]

KDDGNYKTRAEV
KFEGDTLVNRIEL
KGIDFKEDGNILG
HKLEYNYNSHNV
YIMADKQKNGIK
VNFKIRHNIEDGS
VQLADHYQQNTP
IGDGPVLLPDNHY
LSTQSALSKDPNE
KRDHMVLLEFVT
AAGITLGMDELY
KNXNG-COSR

DARPins are synthetic small non-immunoglobulin proteins that are a promising alternative for monoclonal antibodies. Butelase-1-mediated bioconjugation has been applied to labeling breast cancer marker epidermal growth factor receptor HER2-specific designed ankyrin repeat proteins 926 (DARPin 926). By recombinantly expressing DARPin with a C-terminal NHV motif, Tam et al. successfully linked fluorescein and the cationic amphipathic cytolytic peptide magainin (GIGKFLHSAKKFGKAFVGEIMNS) to DARPin using butelase-1 with >90% yield within 30 min (**Figure 26**), allowing the bioimaging and precision bioconjugation, respectively [155]. Site-specifically linking cargo-of-interest to peptides and proteins by butelase-1 promotes precision biomanufacturing.

Thioesters are essential for various peptide synthesis strategies and peptide segment condensation [161-164] as well as subtiligase-mediated ligation. Butelase-1-mediated C-terminal modification of peptides and proteins allows the preparation of peptide and protein thioesters (**Figure 27**) [160]. Liu and colleagues produced peptide thioesters by ligating glycine thioester and peptide with C-terminal NHV motifs with high efficiency within 2 h. Through butelase-1-mediated thioester preparation, Liu and colleagues modified ubiquitin, designed ankyrin repeat proteins (DARPin), and green fluorescent protein (GFP), which are subjected to further modification by subtiligase, biorthogonal tandem ligation, or other strategies (**Figure 28**) [160].

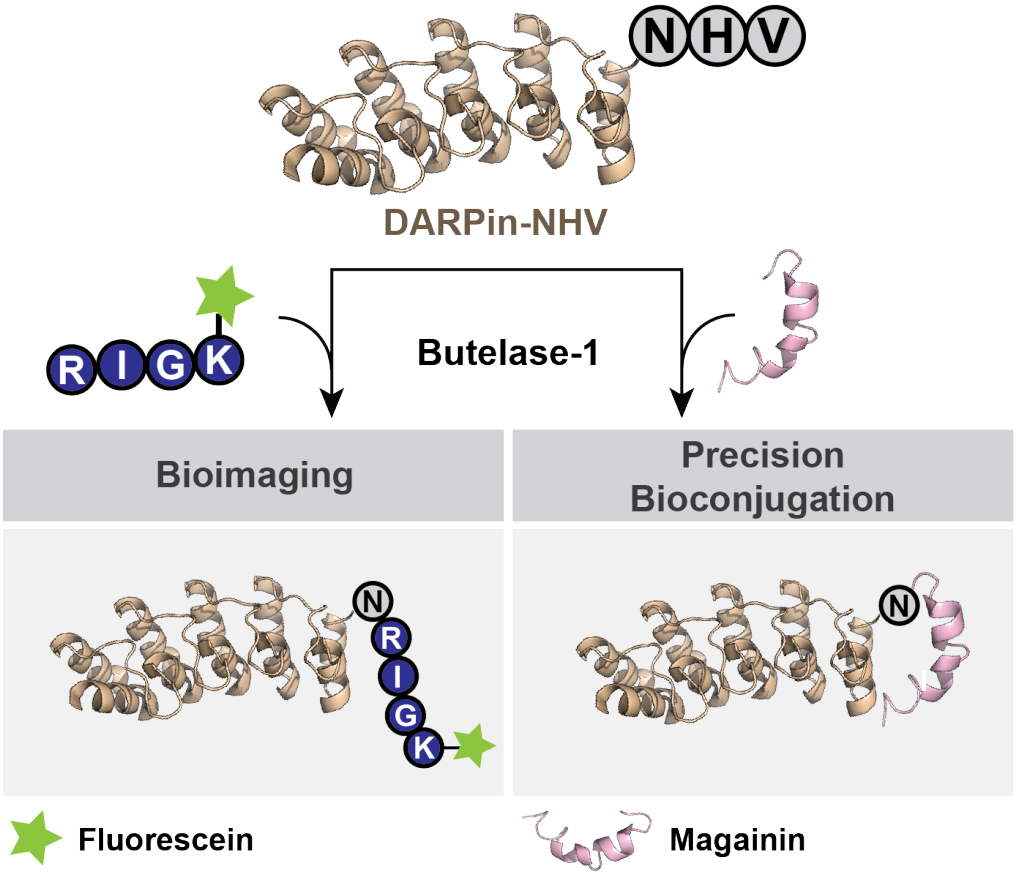


Figure 26. Butelase-1-mediated C-terminal modification of HER2-specific DARPin 926. The RIGK-fluorescein and cytolytic peptide magainin were linked to the C-terminus of DARPin 926 for bioimaging and targeted drug conjugate, respectively. The PDB accession codes of DARPin 926 [165] and Magainin are 5LW2 and 2LSA, respectively [166].

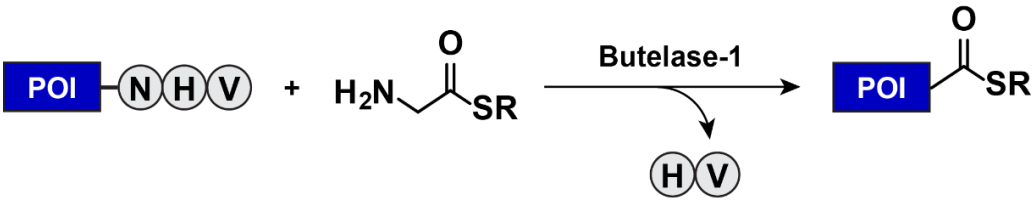


Figure 27. Thioester preparation by butelase-1-mediated C-terminal modification. The thioester group was linked to the protein of interest carrying a C-terminal Asn-His-Val motif. The blue rectangle represents the protein of interest (POI). The grey circles represent amino acids.

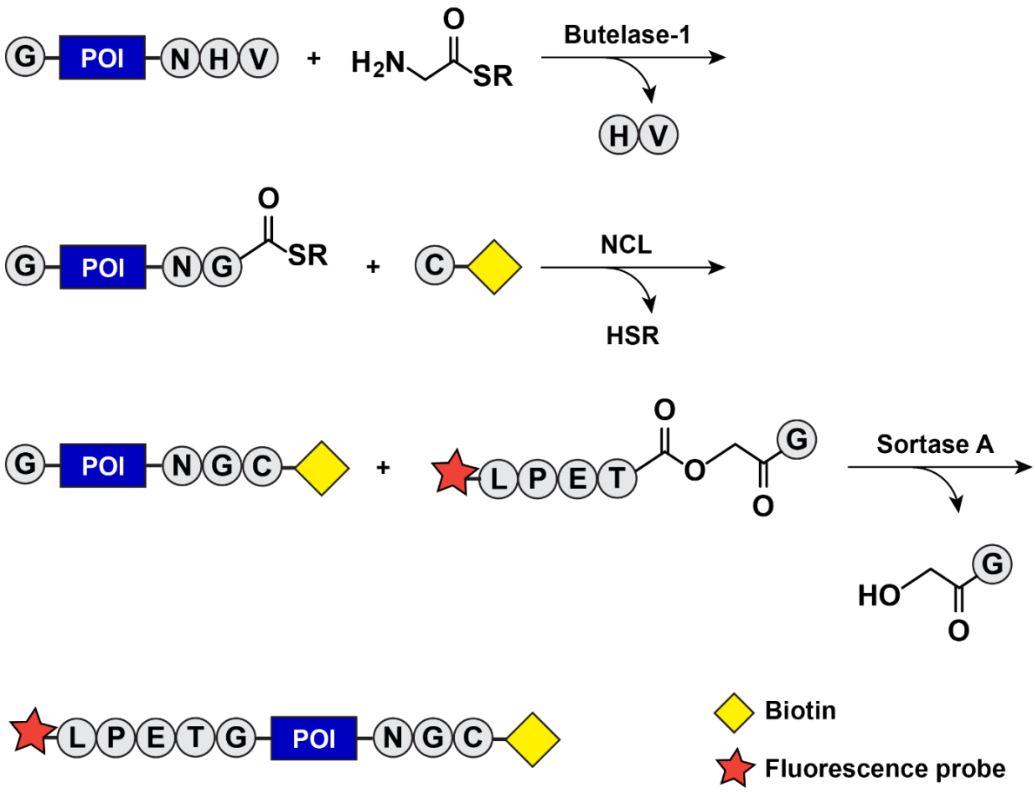


Figure 28. Tandem ligation by butelase-1, sortase A, and NCL [160]. The thioester was linked to the protein of interest (POI, blue rectangle) containing a C-terminal Asn-His-Val motif by butelase-1-mediated ligation, following the installation of biotin (yellow square) using NCL, and lastly the ligation of fluorescence probe (red star) to the POI.

1.3.3 Live-Cell Labeling

Labeling of living cells enables the tagging and tracking of the cells as well as interactions between cells and the surrounding environment. However, live-cell labeling requires very stringent conditions as the reaction has to be performed under physiological conditions. Bi *et al.* inserted the plasmid encoding Spp-OmpA protein with C-terminal NHV tripeptide motif, which can be recognized by butelase-1 specifically, into the *E. coli*. The induced cells display the Spp-OmpA carrying NHV at its C-terminus. By incubating butelase-1 and the fluorescein (GIGGIRK) with the washed cells at 37 °C for 30 min, the fluorescein was successfully ligated to the Spp-OmpA (**Figure 29**) [156].

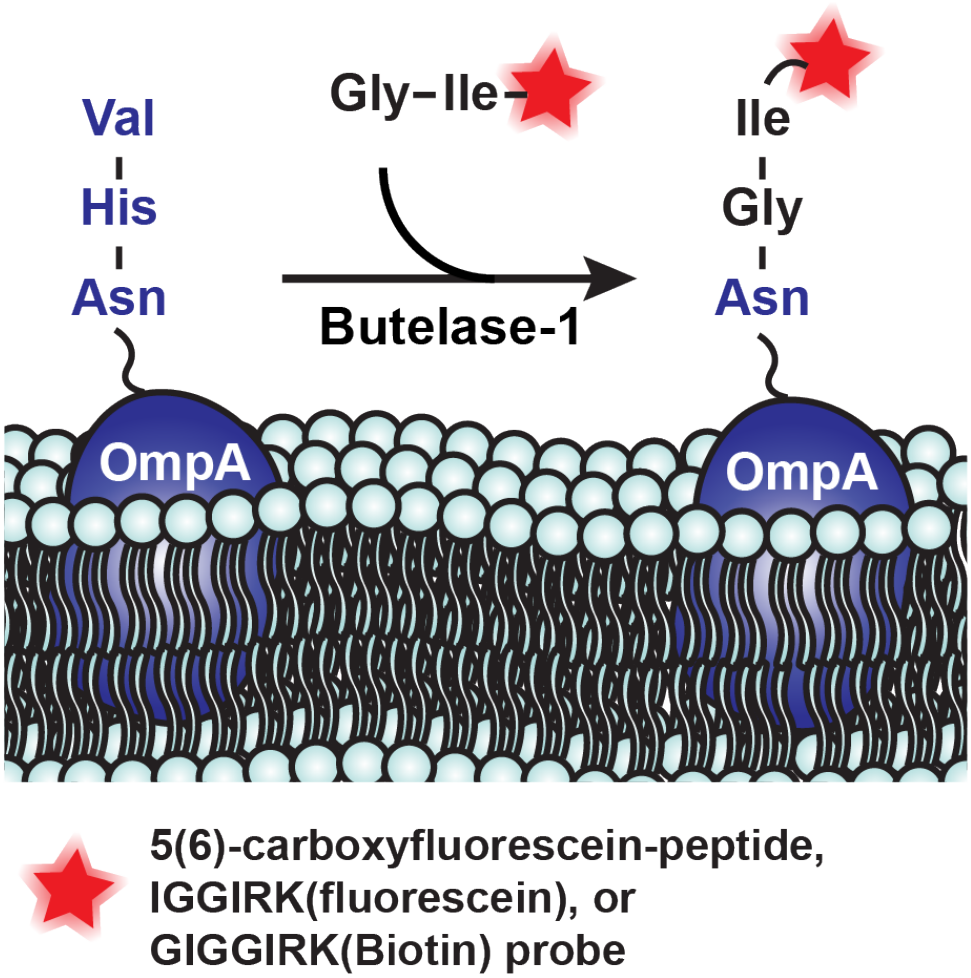


Figure 29. Live-cell labeling of the anchoring protein OmpA by butelase-1. By transforming the plasmid encoding the anchoring protein OmpA equipped with a C-terminal Asn-His-Val motif, the Asn-His-Val motif can be displayed on the cell surfaced and thus recognized by butelase-1 for labeling. The GIGGIRK(Biotin) probe was synthesized by solid-phase peptide synthesis (SPPS).

Receptor-mediated endocytosis (RME) is a coupled process during which the extracellular ligands are taken up by the cell through internalization of the cell surface receptors the ligands are bound to [167]. Elucidation of the RME promotes the utilization of RME as a potential drug delivery strategy [168-170]. Bi *et al.* designed a disulfide-based redox FRET probe to detect the redox state in endosomal compartments. The probe is first linked to the C-terminal recognition signal NHV of the human transferrin receptor 1 (TfR1) by butelase-1-mediated ligation, and the presence of holotransferrins (transferrins loaded with iron ions) lead to the internalization of the Holotransferrin-TfR1 (FRET) complex. Through the reduction of the disulfide bond linking the quencher DABCYL in the endosomal compartments, the fluorescence of BODIPY of the probes was released. The apotransferrin-TfR1 (reduced FRET) complex was then recycled to the cell surface after the irons are released (**Figure 30**) [157]. This strategy is not restricted to TfR1 trafficking and applies to other RME.

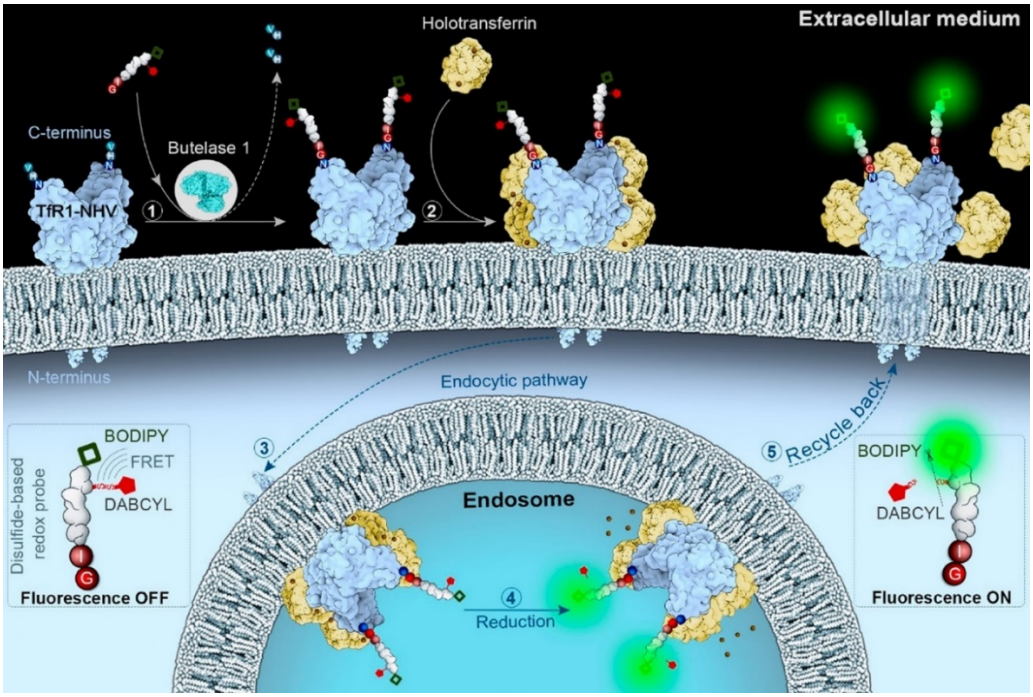


Figure 30. Strategy for real-time live-cell imaging of redox states by butelase-1. The TfR1-mediated endocytosis was studied by butelase-1-mediated transferrin labeling. Figure taken from [157].

1.3.4 One-Pot Ligation

Combining multiple enzymes at once to perform one-pot biorthogonal ligation saves reaction time and reagents. It requires the applied enzymes to be highly specific and to recognize different peptide motifs. Butelase-1 preferably recognizes an NHV tripeptide motif, while sortase A recognizes the LPXTG (where X stands for all amino acids) hexapeptide motif. Owing to their disparate recognition signals, butelase-1 and sortase A were applied for one-pot dual-labeling of proteins [158]. By installing the corresponding recognition signals to the proteins of interest, butelase-1 and sortase A were used in a one-pot condition to link the two proteins to a linker. Notably, it is not necessary for the linker to be peptide or protein. As long as the preferred incoming amino acids are equipped at the termini, the linker can be a synthetic polyethylene glycol (PEG)-based linker (**Figure 31**) or a double-stranded oligonucleotide-based linker, which can be recognized and cleaved by restriction enzyme (**Figure 32**).

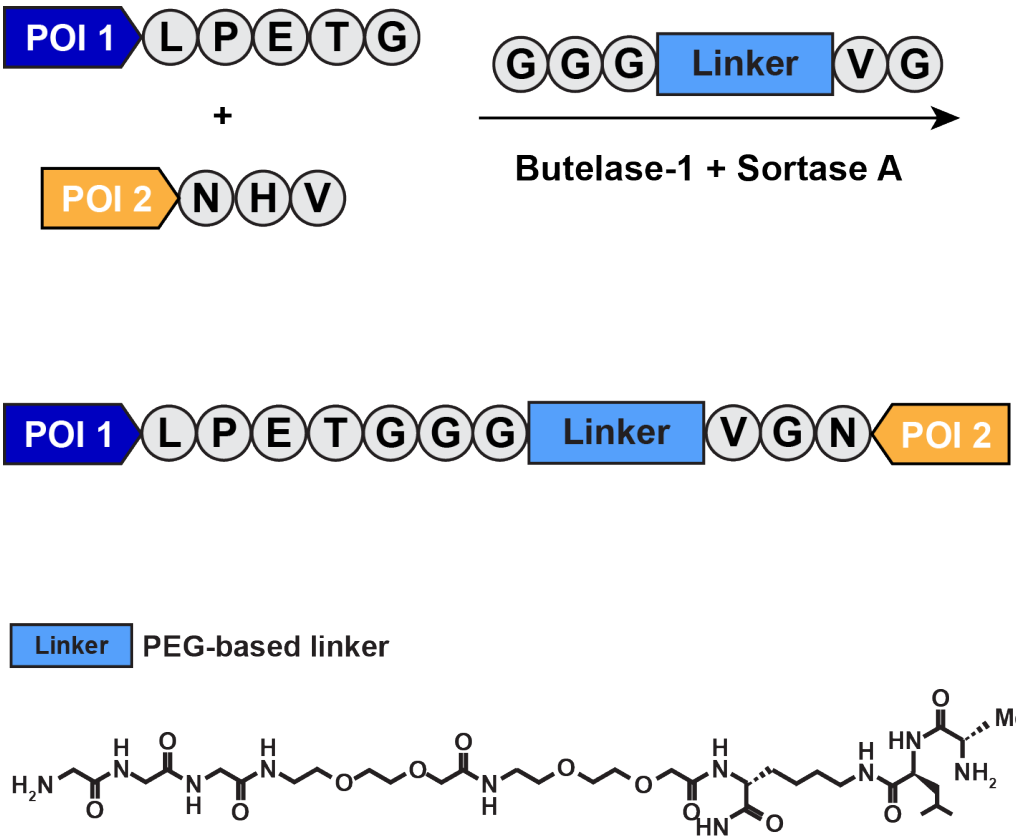


Figure 31. One-pot ligation using butelase-1, sortase A, and a two-headed PEG-based linker to prepare C-to-C fusion protein. The linker was equipped with the Gly-Gly motif, as the incoming nucleophile for sortase-A-mediated ligation, and a Val-Gly motif, as the incoming nucleophile for butelase-1-mediated ligation. The linker was synthesized by solid-phase peptide synthesis (SPPS).

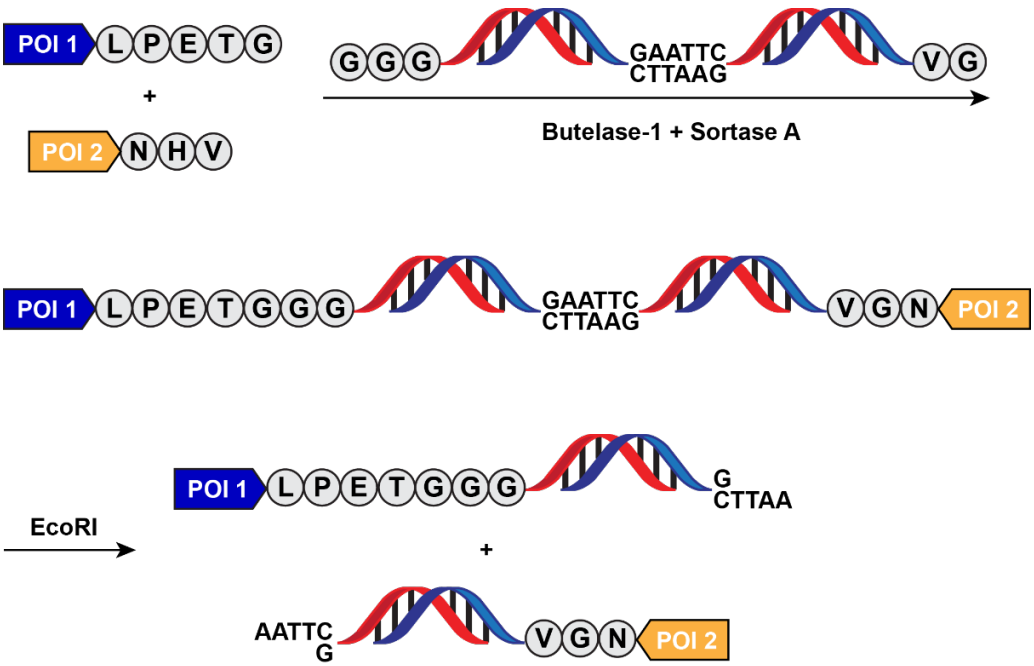


Figure 32. One-pot ligation using butelase-1, sortase A, and a double-stranded oligonucleotide linker. The oligonucleotide linker was equipped with a Gly-Gly-Gly tripeptide motif for sortase A-mediated ligation at one end and a Vla-Gly dipeptide motif at the other end for butelase-1-mediated ligation. The linker can later be cleaved by the restriction enzyme EcoRI.

1.3.5 Bio-Orthogonal Sequential Ligation

Butelase-1 and PAL exhibit different substrate preferences. An Asn-His-Val tripeptide motif is preferred by butelase-1, while VyPAL2 recognizes an Asn-Ser-Leu tripeptide motif better. Wang *et al.* showed that the ligation reaction catalyzed by butelase-1 (k_{cat}/K_m of 17265 ± 465 $M^{-1}s^{-1}$) was 18.5 times more efficient than VyPAL2 (k_{cat}/K_m of 932 ± 32 $M^{-1}s^{-1}$) using a nine-residue acyl peptide substrate Ac-KKLAVINHV and a heptapeptide GIGGIKA. Likewise, at the primed side, a P2'-Phe is favored by VyPAL2 but not butelase-1. The ligation efficiency of a hexapeptide substrate YKANGL and a heptapeptide substrate GFGGIKA by VyAPL2 (k_{cat}/K_m of 19559 ± 164 $M^{-1}s^{-1}$) is more than four times of that by butelase-1 (k_{cat}/K_m of 4256 ± 52 $M^{-1}s^{-1}$) [136]. Taking advantage of their high and disparate substrate specificity, butelase-1 and VyPAL2 can be used for site-specific tandem ligation.

An affibody Z_{EGFR} was equipped with an N-terminal GF dipeptide, as the incoming nucleophile for VyPAL-mediated ligation, and a C-terminal NHV tripeptide, as the acyl donor for butelase-1-mediated ligation [136]. Installation of the fluorescein peptide (with C-terminal tripeptide NGI) and an octapeptide with mitochondrial membrane-disrupting peptide KLA (GIGGFKGG-klaklaklaklaklak, the smaller case indicates D-amino acid) at the N- and C-terminus of the affibody Z_{EGFR} was performed by incubation of the peptide substrates with butelase-1 and VyPAL2 sequentially with about 70% yield. Using the same strategy, butelase-1 and VyPAL2 were also applied for chemoenzymatic tandem ligation to generate cyclic Z_{EGFR}-drug conjugate (**Figure 33**) [136].

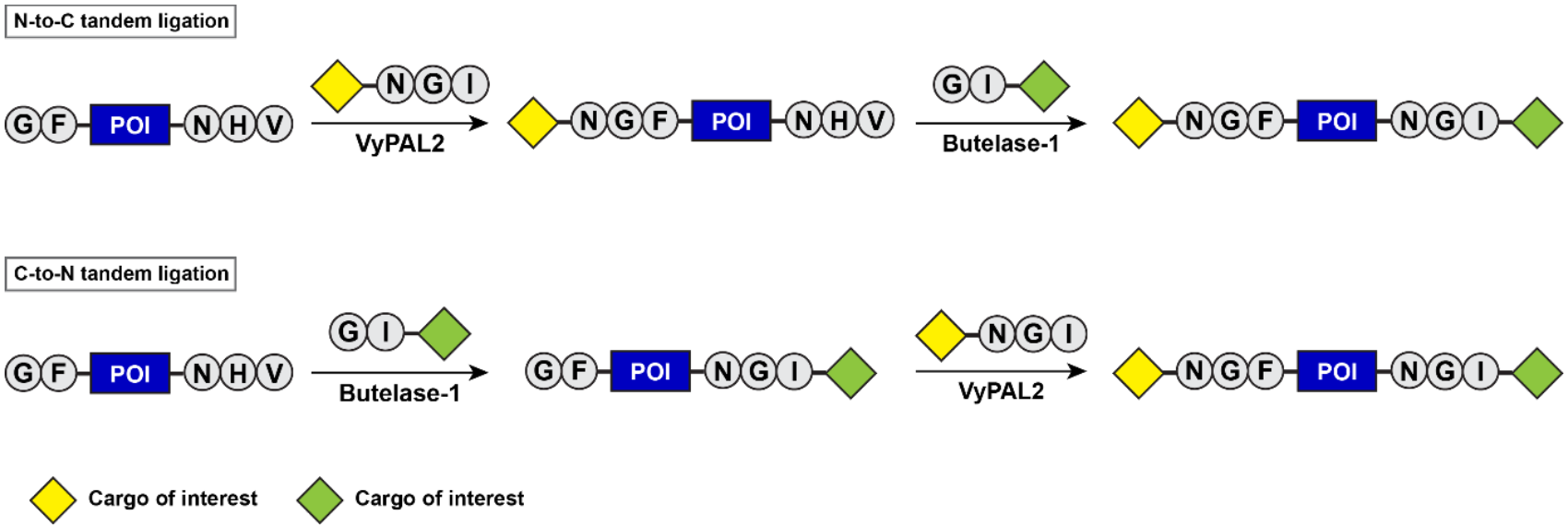


Figure 33. N-to-C and C-to-N tandem ligation by butelase-1 and VyPAL2. Taking advantage of the different substrate preferences of butelase-1 and VyPAL2, butelase-1 and VyPAL2 can be used in tandem for dual-labeling of protein of interest (POI, colored in blue). Cargos of interest are the yellow and green rectangle. Figure adapted from [136].

1.3.6 Synthesis of Peptides with Unusual Architectures

Peptide dendrimers are branched polypeptides containing a peptidyl multivalent core, such as a central lysine, and covalently attached functional units. The dendrimeric peptides were applied to the production of vaccines in the 1980s by Tam and coworkers [171, 172]. Later, the peptide dendrimers were used as protein mimetics, therapeutics, and antimicrobial agents [173].

An antimicrobial dendrimer containing several tetrapeptide RLYR motifs was designed in 2002 by Tam and coworkers. Tetrapeptide RLYR contains the common antimicrobial BHHB motif (B is basic amino acid, H is hydrophobic amino acid). Compared to the linear counterparts, the chemically-prepared tetravalent dendrimeric peptides possess higher stability toward proteolysis and lower toxicity [174]. In 2016, Cao *et al*. used lysyl dendron cores and N-acetylated thiodepsipeptides as acyl donors for butelase-1-mediated reaction to prepare peptide dendrimers (**Figure 34**). The RLYR-containing tetravalent dendrimeric peptides (Ac-RLYRNRIβA)$_4$K$_2$KY) had IC_{50} of 2.4 μM and 1.4 μM against *E. coli* and *S. aureus*, respectively. The RLYR-containing tetravalent dendrimeric peptides were shown to be more potent than its monomer counterpart, and it also showed antimicrobial activity against six drug-resistant strains [159].

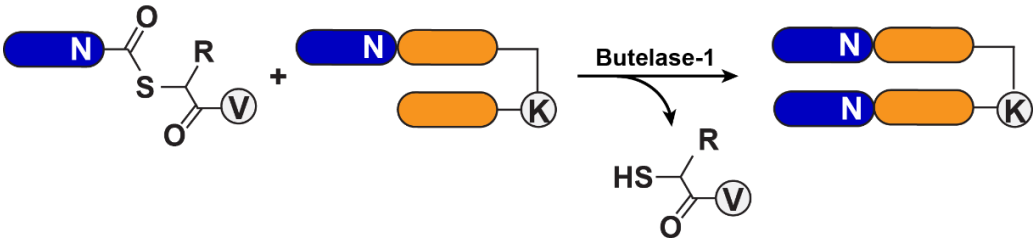


Figure 34. Schematic representation of dendrimer preparation by butelase-1-mediated bioconjugation. The N-acetylated thiodepsipeptide was linked to the bivalent lysyl dendron core by butelase-1-mediated ligation. Figure adapted from [159].

Later, to enhance the stability of RLYR motif-containing peptides, Hemu *et al.* successfully prepared cyclo-oligomeric antimicrobial peptides containing the RLYR motifs using butelase-1. It was also demonstrated that different conditions resulted in different product distributions by butelase-1. For example, precursors of nine or more residues were cyclized by butelase-1 rapidly, on the other hand, mixing butelase-1 with precursors containing five to eight amino acids resulted in the formation of cyclodimers as major products, and using tripeptides and tetrapeptides as precursors resulted in cyclotrimers and cyclotetramers as major products. It was also shown that a P2-Pro promotes the formation of cyclotrimers compared to the proline-free counterparts. By manipulating the reaction time, the substrate concentration, and the length and sequence of the precursors, butelase-1-mediated cyclo-oligomerization of the antimicrobial peptides containing the RLYR motifs was efficiently performed (**Figure 35**) [175].

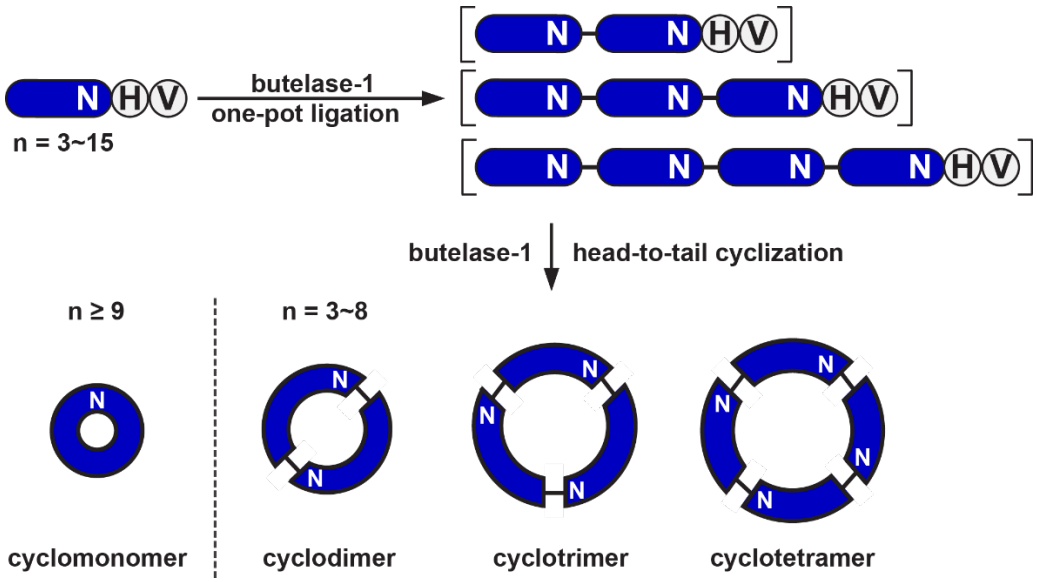


Figure 35. Schematic representation of butelase-1-mediated cyclo-oligomerization. The reaction can be controlled by conditions, such as the length of the substrates, the sequence of the substrate, the reaction time, and the substrate concentration. The yield of cyclodimer, cyclotrimer, and cyclotetramer using butelase-1-1mediated cyclo-oligomerization is higher when the length of the precursor is shorter than eight residues. n is the number of core residues of precursors. Figure adapted from [21].

Chapter 2 Hypothesis and Aim

The peptide asparaginyl ligases (PALs) have garnered attention for their broad substrate scope, simple recognition motif, as well as ability to catalyze traceless ligation at near-neutral pH and physiological conditions without ATP, cofactors, or ribosomal machinery. However, it is challenging to distinguish PALs from their protease homologs, asparaginyl endopeptidases (AEPs), which have highly similar structures and amino acid sequences. Several sequence motifs to identify PALs have been reported, however, they have not been utilized to fish out PALs from a pool of AEPs. My hypothesis is that the ligase activity determinants (LADs) can be used to facilitate the discovery of novel butelase-1-like PALs.

The major goal of this thesis is to discover and characterize the LADs, the sequence motifs that can be applied to distinguish PALs and AEPs. In this thesis, the specific aims are as follow:

(**1**) The classification of putative PALs and AEPs based on the amino acid compositions at the LAD sites. The results are summarized in chapter 4, which mainly describes the discovery of LADs by bioinformatics. A dataset containing 1570 unique sequences of putative PALs and AEPs from plants were generated and used for the identification of the evolutionary important residues by universal evolutionary trace (UET) analysis, the comparison of the amino acid compositions of the substrate-binding pockets, and visual Correlated Mutation Analysis Tool (visualCMAT). Several evolutionary important residues that vary among PALs and AEPs were discovered and used for the prediction of butelase-1-like PALs.

(**2**) The validation of the predicted PALs and AEPs of the dataset. The results are summarized in chapter 5, which includes the expression, purification, activation, and characterization of putative PALs and AEPs. The enzymatic activities of the selected sequences were investigated and determined by biochemical assays using various peptide substrates at different pH. Five novel butelase-1-like PALs, four partial ligases, and one butelase-2-like protease were discovered.

(**3**) The engineering of novel PALs and AEPs by mutagenesis at the LAD sites. The results are summarized in chapter 5, which includes the characterization of a butelase-1-like PAL engineered for higher catalytic efficiency and a butelase-2-like AEP engineered for higher ligase activity.

Chapter 3 Materials and Methods

3.1 Data Mining and Bioinformatic Analysis

3.1.1 Generation of the Sequence Dataset

Full-length amino acid sequences of butelase-1 (GenBank accession code: KF918345) and OaAEP1b (GenBank accession code: KR259377) proenzyme were used as a query individually to mine transcriptome in various databases. Blastp was performed in the 1000 Plant (OneKP) Transcriptome Database and National Center for Biotechnology Information (NCBI non-redundant protein sequence (nr) databases. TBlastn was performed against the NCBI transcriptome shotgun assembly (TSA) database and nucleotide collection (nr/nt) database. Access to 1000 plant transcriptomes was provided by the OneKP consortium [176]. In all BLAST searches, the threshold parameters were set to >60% identity and >90% sequence coverage as compared to the query. For non-annotated hits, nucleotide sequences were translated in all six open reading frames (ORFs) using the ExPASy translate tool (available online at: https://web.expasy.org/translate/) [177], and the longest ORFs were extracted. Data from various databases were pooled, and repeated homologs were removed to subsequently retrieve 1570 unique putative AEP homolog sequences.

3.1.2 Signal Peptide Prediction

The signal peptide cleavage sites were predicted using SignalP 5.0 [178], which is available online at: http://www.cbs.dtu.dk/services/SignalP/. The full-length amino acid sequences of PALs and AEPs were used as the inquiries on the online server, and ‘Eukarya’ was chosen for the ‘Organism group’ section.

3.1.3 Sequence Alignment and Comparison

The amino acid sequences were aligned using JalView 2.10.5 [139, 140] with Clustal Omega [138] with BLOSUM62 substitution matrix and default gap penalties. The conservation score for a column was computed according to the amino acid composition of each column [179] and visualized as a histogram on a scale of 11, where 11 (indicated by a *) indicates that the amino acids in the column are absolutely conserved and 10 indicates all amino acids properties are conserved. The decrease of the number indicates the decrease of the conservation.

The sequence identity and similarity were calculated by EMBL-EBI EMBOSS Water with default setting, which is available online at https://www.ebi.ac.uk/Tools/psa/emboss_water/. The native sequences of the enzymes listed in **Figure 36** were retrieved from OneKP database and GenBank. The accession codes for AtLEG, AtLEG, butelase-1, butelase-2, jack bean AEP, human legumain, mouse legumain, OaAEP1b, VcAEP, and VyPAL2 are NP_176458.1, NP_195020.1, KF918345.1, ALL55651.1, P49046.1, AAH03061.1, 001365804.1, KR259377, NJLF_2006210, and MK085231.1, respectively. The accession codes of the sequences of the ferns in this dataset retrieved from OneKP database are HEGQ_2009324, UWOD_2001395, QVMR_2008863, SKYV_2065121, POPJ_2007355, and MROH_2010625. The accession codes of the sequences of the ferns in this dataset retrieved from GenBank are FX958798.1, GBGN01122218.1, GEEJ01026098.1, GEEI01020256.1, and GBTV01008543.1.

3.1.4 Universal Evolutionary Trace Analysis

The aligned sequences were saved in Fasta format and the Fasta file was used as the inquiry for the Universal Evolutionary Trace (UET) analysis server

developed by Lichtarge's laboratory [180], which is available online at: http://lichtargelab.org/software/uet. The default setting was applied, and the fragments, identical sequences, and short sequences were not removed. The main results of the UET analysis are a sequence-based phylogenetic tree in the New Hampshire X (.nhx) format and real-value evolutionary trace (rvET) scores allocated to each residue based on the conservation among evolutionary close and distant sequences.

PyMol Molecular Graphics System, Version 2.0 Schrödinger, LLC. was used to visualize the UET analysis results. PDB file of the crystal structure selected as template for UET was opened in PyMol. Firstly, PyMol Plugin PyETV was launched, and the name of the template was entered in the section 'structure to use.' The ET rank file (.rank) generated from the UET analysis was chosen in the section 'ET ranks file path.' 'Prismatic (Gobstopper)' was selected in the section 'Show ET residue' to color each residue based on the UET analysis result. The colored structure was exported by simply entering 'ray 2000,2000' or using the following script:

```
load /tmp/thy_model/1l9l.pdb;
hide lines;
show cartoon;
set ray_trace_mode, 1; # color
bg_color white;
set antialias, 2;
remove resn HOH
remove resn HET
ray 3000,3000
```

```
png /tmp/1l9l.png
```

The numbers after ‘ray’ can be adjusted to generate figures with different resolutions and sizes.

3.1.5 Generation of Sequence-Based Phylogenetic Tree

A sequence-based phylogenetic tree containing 1570 sequences in the New Hampshire X (.nhx) format was automatically generated through UET analysis by the unweighted pair group method with arithmetic mean (UPGMA) method. The sequence-based phylogenetic tree containing 57 sequences of known and putative PALs and AEPs was generated through Simple Phylogeny, which is available online at https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/ [181]. The phylogenetic trees were visualized using iTOL Interactive Tree of Life (iTOL 6.1.2, Tree of Life version 1.0) [182, 183]. The classification of the PAL and AEP sequences was based on that by Yamada *et al.* [91]. The color strips displayed as rectangles were assigned to each branch using the script modified from the template provided by iTOL (available online at: https://itol.embl.de/help.cgi#ranges), which is shown below:

```
DATASET_COLORSTRIP

SEPARATOR SPACE

DATASET_LABEL color_strip2

COLOR #ff0000

STRIP_WIDTH 25

MARGIN 0

BORDER_WIDTH 1

BORDER_COLOR #000
```

```
SHOW_INTERNAL 0

DATA

GEUP01043581.1_Zenia_insignis #c1a9de COL#ff0000
```

The last line of the script is for assigning the color to one particular sequence based on the classification. Only one example (last line of the script) from the data is shown.

3.1.6 visualCMAT Analysis

The 1570 sequences from the dataset were aligned using CLUSTAL Omega; the aligned sequences were upload and analyzed by visual Correlated Mutation Analysis Tool (visualCMAT) [184], available online at: https://biokinet.belozersky.msu.ru/visualcmat. The default setting was used. The results were visualized by PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

3.1.7 WebLogo

The graphical representation of amino acids of multiple sequence alignment in Fasta format was generated by WebLogo 2.8.2, which is available online at: https://weblogo.berkeley.edu/logo.cgi) [185, 186]. The sequences of PALs and AEPs were aligned using JalView 2.10.5 with Clustal Omega as previously described. The sequences of the cyclotides of the Violaceae family retrieved from CyBase [38] were aligned using JalView 2.10.5 with the probabilistic consistency-based multiple alignments of amino acid sequences (Prob-Cons), which is available online at: http://probcons.stanford.edu/ [187], with the default setting.

3.1.8 Modeling of the Enzymes

To visualize the proteins that do not have the crystal structures solved, the amino acid sequences were submitted to the Iterative Threading ASSEmbly Refinement (I-TASSER) [117-119] for modeling, which is available online at: https://zhanglab.dcmb.med.umich.edu/I-TASSER/.

3.2 Recombinant Expression of Ligases

3.2.1 Plasmid Design

The DNA sequences encoding full-length proteins devoid of the signal peptide were inserted into the pET28a(+) vector with restriction sites Ndel/Xhol to generate a His6-ubiquitin-PAL/AEP fusion protein construct (Genscript, USA). The codons were optimized for bacterial expression. The histidines were added for the purificaton by immobilized metal affinity chromatography (IMAC) using the HisPur™ Ni-NTA Resin (see 3.3.2 for details). The ubiquitin was added to the construct for yield enhancement and will be removed after N-terminal cleavage during autocatalytic activation. The sequence of ubiquitin used was previously reported [7], which is shown below:

MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFA GKQLEDGRTLSDYNIQKESTLHLVLRLRGGA.

Mutations were performed by the Q5 mutagenesis kit (New England Biolabs, USA) through polymerase chain reaction (PCR). The forward primer used for deletion of part of the N-terminal domain of VoPAL1, and VvPAL1 were TTACGCCTGCCGTCTGAAGCG, and CTGCGCCTGCCGTCTGAA, respectively. The reverse primer used was GCCGCTGCTGTGATGATGATGATGATG. The pMJS9 plasmid containing a disulfide isomerase Erv1p gene and a protein disulfide isomerase (PDI) gene (a kind gift obtained from Professor Lloyd Ruddock, University of Oulu, Finland) was used. The DNA sequences of BmAEP1, VyPAL4, and VyPAL5 were retrieved from the NCBI nucleotide collection (nr/nt) database using NCBI reference sequence XM_022292359.1, GenBank accession number MK085233.1, and MK085234.1 from the NCBI nucleotide collection (nr/nt)

database. DNA sequences of PePAL1, PePAL2, SiPAL1, VaPAL1, VoPAL1, VuPAL1, and VvPAL1 were retrieved from the NCBI transcriptome shotgun assembly (TSA) database using the GenBank accession number GBRT01052954.1, GBRT01019050.1, JL339165.1, GFWC01037197.1, GFXR01024405.1, GCAB01004088.1, and GFWF01025417.1, respectively. The DNA sequence of PiPAL1 was retrieved from the OneKP database with the accession code BQEQ_2002574.

3.2.2 Plasmid Extraction and DNA Sequencing

Overnight culture of the TOP-10 or DH5-α *E. coli* cells containing desired plasmids was added to the tube and centrifuged at 12,000 rpm for 2 minutes. The plasmid was extracted from the drained pellet using EZ-10 Spin Column Plasmid DNA Miniprep Kit (Bio Basic, USA) following the manufacturer's instruction. The purified plasmid was sequenced by Bio Basic sequencing service.

3.2.3 Preparation of the Competent Cells

Approximately 50 µL of competent cells was cultured in 5 mL of LB broth at 30 °C overnight. Overnight culture was diluted 100 times in 100 mL of LB broth and cultured at 30 °C until the OD_{600} reached 0.5. The culture of OD_{600} 0.5 was aliquoted into tubes and placed on ice for 20 min then centrifuged at 4000 g for 5 min at 4 °C. The supernatants were discarded, and the cells were resuspended by cold 0.1M $CaCl_2$ on ice. The resuspended cells were incubated on ice for 20 min then centrifuged at 4000 g for 5 min at 4 °C. The supernatants were discarded, and the cells were resuspended to a density of 30 OD_{600} by cold solution containing 0.1M $CaCl_2$ and 10% glycerol on ice. The cells were stored at -80 °C.

3.2.4 Transformation of Plasmid into Bacterial Cells

The plasmids were transformed into SHuffle T7 *E. coli* cells (New England Biolabs, USA) via the electroporation method or heat shock method, which includes incubation on ice for 30 minutes, heat shock at 42 °C for 60 seconds, and incubation on ice for 20 minutes. Both transformation methods require recovery for 1 h in 200 µL super optimal broth with catabolite repression (SOC). For the transformation of plasmid with mutations generated through Q5 mutagenesis kit (New England Biolabs, USA) and PCR, the PCR product was directly mixed with KLD Enzyme Mix (New England Biolabs, USA) following the manufacturer's instruction. The SOC was prepared according to the recipe of Cold Spring Harbour Protocols (available online at: http://cshprotocols.cshlp.org/content/2018/3/pdb.rec098863.full?rss=1), which contains 20 g Tryptone (Bio Basic, Asia Pacific), 5 g yeast extract (Bio Basic, Asia Pacific), 0.5 g NaCl (Merck), and 20 mM glucose (Merck, USA). The transformed and recovered cells were spread on LB agar plates supplemented with 50 µg/mL kanamycin (Sigma-Aldrich, USA), 50 µg/mL streptomycin (Sigma-Aldrich, USA), and 25 µg/mL chloramphenicol (Sigma-Aldrich, USA), and incubated at 30 °C overnight.

3.2.5 Recombinant Expression of Target Proteins

The colony of transformed cells was selected by inoculation loop and incubated in 5-10 mL of LB broth supplemented with 50 µg/mL kanamycin, 50 µg/mL streptomycin, and 25 µg/mL chloramphenicol overnight at 30 °C. The liquid culture and 1% glucose were added to LB broth containing 50 µg/mL kanamycin, 50 µg/mL streptomycin, and 25 µg/mL chloramphenicol with a 1:500 ratio. 0.1% (w/v) L-Arabinose and 0.4% (v/v) ethanol were added to the

culture at 30 °C when OD_{600} reached 0.4. The temperature was reduced to 16 °C after one hour and the protein expression was induced by 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 20 h.

3.2.6 Harvesting the Bacterial Pellets by Centrifuge

The *E. coli* cells were harvested by centrifugation at 8,000 g for 15 min at 4 °C using JA-10 rotor (Beckman Coulter, USA) and Avanti J-25 centrifuge (Beckman Coulter, USA). The pellets were stored at -80 °C refrigerator if not subjected to purification immediately.

3.3 Purification of Recombinant Ligases

3.3.1 Homogenization of Cells by Sonication

The purification of proteins is comprised of multiple steps. The pellets of cells containing target proteins were lysed by 20 mM pH 7.0 phosphate buffer containing 100 mM NaCl, 1 mM Ethylenediaminetetraacetic acid (EDTA), 5% (v/v) glycerol, 1 mM Phenylmethane sulfonyl (PMSF), 5 mM β-mercaptoethanol (β-ME), and 0.1 % Triton-X100 (v/v). The solution in which the pellet resuspended was then sonicated at 130 W for 10-40 min, depends on the volume of the solution, with a frequency of 3 s pulse in every 13 s on ice. The homogenized solution was centrifuged at 10,000 g for 30 min at 4 °C using Avanti J-25 centrifuge (Beckman Coulter, USA) to remove the cell debris. The supernatant was filtered using a 50 mm polyethersulfone bottle-top vacuum filter (ThermoFisher, USA).

3.3.2 Immobilized Metal Affinity Chromatography (IMAC)

The HisPur™ Ni-NTA Resin (ThermoFisher, USA) was equilibrated by equilibration buffer, which contained 50 mM pH 7.0 HEPES buffer with 10 mM imidazole, 100 mM NaCl, and 1 mM EDTA, before use. The equilibrated beads were then added to the filtered solution at 4 °C for a minimum of 30 min to allow the target proteins with His-tag to bind to the resin. After washed by washing buffer containing pH 7.0 20 mM HEPES buffer with 50 mM imidazole, 100 mM NaCl, and 1 mM EDTA, the target proteins were eluted by the same solution with 500 mM imidazole.

3.3.3 Fast Protein Liquid Chromatography (FPLC) – Anion Exchange

The eluent was then purified using a 5 mL HiTrap Q Sepharose column (GE Healthcare, USA) coupled to an AKTA system (GE Healthcare, USA). The

column was equilibrated using a buffer consisting of 50 mM HEPES, 1 mM EDTA, and 1 mM dithiothreitol (DTT) at pH 7.0 for at least two column volumes. Bound proteins were eluted (flow rate <5 mL/min) using a continuous salt gradient of 0−30% (v/v) of a buffer consisting of 50 mM HEPES, 1 mM EDTA and 1 mM DTT and 1 M NaCl at pH 7.0.

For each run, corresponding fractions were analyzed by Western blot analyses, Coomassie staining, and cyclization assay. Fractions with positive cyclization results were pooled and purified with the subsequent purification steps.

3.3.4 Fast Protein Liquid Chromatography (FPLC) – Size Exclusion

The protein was subsequently purified by a HiLoad Superdex 75 prep grade column (GE Healthcare, USA) using 50 mM Na HEPES, 1 mM EDTA, 1 mM DTT, 100 mM NaCl and 5% glycerol at pH 7.0. Prior to the purification, the column was equilibrated using the same buffer for at least two column volumes. The flow rate was set to <1.5 mL/min.

The fractions were subjected to Western Blotting analyses, Coomassie staining, and cyclization assay to pool out the fractions with proteins of correct sizes (about 50-60 kDa) and ligase activity detected.

3.3.5 Activation of PALs and AEPs

To obtain the catalytic enzymes, the pH of the solution containing the enzymes was lowered to 4.0. The solution containing the enzymes was incubated at 4°C, 25 °C, or 37 °C for 10 min to overnight, the condition differs for different enzymes.

3.3.6 Fast Protein Liquid Chromatography (FPLC) – Size Exclusion to Obtain Active Enzymes

To separate the catalytic enzymes from the cleaved domains, such as the C-terminal cap domain and N-terminal domain, the activated enzymes were concentrated by a 10 kDa cut-off Vivaspin® Turbo 15 (Sartorius, Germany) at 4 °C. The catalytic enzymes were purified using HiLoad Superdex 200 pg preparative size exclusion chromatography columns (GE healthcare, USA) with buffer containing 20 mM sodium citrate buffer, 1 mM EDTA, 1 mM DTT, 100 mM NaCl, and 5% (w/v) glycerol at pH 4.5. The flow rate was set to < 1.5 mL/min. The eluents were neutralized to pH 6.0 by buffer exchange using 50 mM pH 6.0 phosphate buffer containing 1 mM EDTA, and stored at 4 °C or −20 °C after the addition of 20% sucrose prior to use.

3.4 Ligase Characterization Assays

3.4.1 Cyclization Assay Using Peptide Substrates

The intramolecular cyclization assay to probe the ligase activity of catalytic enzymes was performed in a reaction mixture of 20 to 50 μL containing 20 mM phosphate buffer, 1 mM EDTA, the catalytic enzymes, and the peptide substrates. The peptide substrates were synthesized by GL Biochem (Shanghai) Ltd., or kindly prepared by solid-phase peptide synthesis (SPPS) by Dr. Xiaohong Zhang of Professor Chuan-Fa Liu's laboratory and Mr. Heng-Tai Liew of Professor James P. Tam's laboratory. The reaction mixture was incubated in 37 °C then qualified by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS or MS) to monitor the presence of cyclic and linear peptides.

3.4.2 Qualification of Cyclization Assay by Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI/TOF MS)

Mass spectrometry was performed on the Refurbished AB Sciex 5800 MALDI-TOF/TOF (Applied Biosystems, Framingham, MA, USA) operated in positive ion reflector mode to detect the peaks of the peptide substrates and products. The laser intensity was set between 3000-5000. The matrix solution contains saturated α-cyano-4-hydroxycinnamic acid, 90% acetonitrile (ACN), and 0.05% Trifluoroacetic acid (TFA). Samples were mixed with the matrix solution at the ratio of 1:1 (v/v) and 0.5 μL of the mixture was spotted onto the Applied Biosystems MDS SCIEX Opti-TOF 384 Well plate.

3.4.3 Qualification and Quantification of Cyclization Efficiency by Reverse-Phase High-Performance Liquid Chromatography (RP-HPLC)

The reaction mixtures containing the catalytic enzymes, the peptide substrates, and the reaction buffer were quenched by incubation at 95 °C for 5 min and the addition of 200 μL quench buffer. The quench buffer is the Buffer A for RP-HPLC and contains 0.1% TFA in Milli-Q water.

The analytical RP-HPLC was run on the Shimadzu system equipped with UV detector at 220, 254, and 280 nm attached to a C18 Aeris widepore analytical column (Phenomenex, USA) using linear gradient between 20-40% of Buffer B (0.1% TFA in ACN) in 20 min.

3.4.4 Fluorescence Resonance Energy Transfer (FRET)-Based Kinetic Assay

The catalytic efficiency of the enzyme was measured using a FRET peptide substrate equipped with the C-terminal recognition motif Asn-Ser-Leu-Lys, the quencher DABCYL, and the N-terminal fluorophore EDANS. Through VuPAL-mediated cyclization the quencher DABCYL was cleaved, and the fluorescence was released, accompanying the shift in molecular weight from 2401.8 (GISTKSIPPIE(EDANS)YRNSLK(DABCYL)) Da to 1805 Da (GISTKSIPPIE-(EDANS)TYN). The peptide substrates were kindly provided by Dr. Xiaohong Zhang of Professor James P Tam's group. Butelase-1 used as explicit comparison and positive control for FRET-based kinetic assays was extracted and purified by Dr. Xiaohong Zhang of Professor James P Tam's group using the previously described protocol [150].

The reaction mixtures were prepared in the Greiner 96 Black Flat Bottom Fluotrac 96-well plate. The reactions were performed at 37 °C with orbital

shaking every 10 s. The fluorescence was read at excitation of 360/20, emission of 490/20, and gain of 70 using Cytation 5 Cell Imaging Multi-Mode Reader (BioTek, USA).

3.4.5 In-Gel Digestion and *de novo* Sequencing by Liquid Chromatography with Tandem Mass Spectrometry (LC-MS/MS)

The gels of bands of activated catalytic enzymes were subjected to in-gel digestion, which started with washing the gel with Milli-Q water and cutting the gels into pieces (about 2 mm by 2 mm). The gel cubes were placed in 100 µL Milli-Q water for 5 min, following the replacement of Milli-Q water with 100 µL solution of 100 mM pH 7.8 Ambic and ACN mixed in a 1:1 ratio. After 10 min, 50 µL ACN was added and let sit for 5 min. After discarding ACN, the gel cubes were shrunken and opaque, otherwise, the above steps can be repeated multiple times. Dry the gel cubes in a vacuum centrifuge or ventilated fume hood for 5 to 10 min.

Following removal of the Coomassie blue dye, the washed and dried gel cubes were subjected to reduction and alkylation. A one-pot reduction and alkylation were achieved by adding 5 mM DTT and 10 mM 2-bromoethylamine (BrEA) in 200 mM Tris-HCl (pH 8.6) at 55 °C for 30 min as described in [188]. The solution was discarded, and the gel cubes were dried. The gel cubes were then rehydrated by 10 ng/µL trypsin solution (dissolved in 50 mM ammonia bicarbonate at 4 °C) and incubated on ice for 30 min, following incubation at 30 °C overnight.

The tryptic digestion was quenched by adding formic acid to a final concentration of 1.0%. The supernatant was removed and saved in a new tube. 30 µL 50% ACN with 5% formic acid was added to the gel cubes and let sit for

45 min, following 5 min sonication for peptide extraction. The solution was removed and saved to the same tube containing the supernatant. After repeating the steps 3 times, 30 µL of 90% ACN with 5% formic acid was added and let sit for 5 min. The solution was removed and saved to the same tube. The tube with extracted peptides was then placed in a vacuum centrifuge, the peptides in the dried tube were then resuspended in 20 µL 0.1% formic acid. To dissolve the peptides thoroughly, the peptides and solution were vortexed for 20 min, sonicated for 20 min, then centrifuged for 20 min. The solutions containing peptides were subjected to LC-MS/MS and *de novo* sequencing, which were kindly done by Dr. Aida Serra, a previous member of Professor Newman Sze's laboratory. The results were analyzed and visualized by PEAKS Studio 7.5 (Bioinformatics Solutions, Waterloo, ON, USA)

3.4.6 Protein Visualization

Proteins samples were subjected to Sodium Dodecyl Sulphate–Polyacrylamide Gel Electrophoresis (SDS-PAGE) for visualization by Coomassie staining and Western Blot analyses.

The 12.5% resolving gels were prepared by mixing 4.2 mL Milli-Q water (Merck Milli-Q® IQ 7000 Ultrapure Water System), 3.1 mL 40% acrylamide bis 19:1 (Bio-Rad, USA), 2.5 mL 1.5M Tris-HCl, pH 8.8, 100 µL 10% SDS, 50 µL 10% ammonium persulfate (APS), and 10 µL Tetramethylethylenediamine (TEMED) for two gels. The 6% stacking gels were prepared by mixing 2.91 mL Milli-Q water, 750 µL 40% acrylamide bis 19:1, 1.26 mL 0.5M Tris-HCl, pH 6.8, 50 µL 10% SDS, 25 µL 10% APS, and 5 µL TEMED for two gels. The concentration of acrylamide bis was subjected to adjustment to better visualize the proteins.

The protein ladders used were PageRuler™ Plus Prestained Protein Ladder (Thermo Fisher, USA) and Precision Plus Protein™ All Blue Prestained Protein Standards (Bio-Rad, USA). The six-times protein loading dye was prepared by mixing 1.2 g SDS, 6 mg bromophenol blue, 4.7 mL glycerol, 1.2 mL 0.5 M pH 6.8 Tris buffer, 2.1 mL Milli-Q water, and 0.93 g DTT (stored in -20 °C). The 10 times running buffer for SDS-PAGE was prepared by dissolving 288 g glycine, 60.4 g Tris base, 20 g SDS, in 2 L Milli-Q water. The 10 times buffer is diluted to the one-time SDS running buffer before use. The SDS-PAGE was performed using CBS Scientific Company Electrophoresis Power Supply EPS-250 Series II.

The Coomassie staining buffer contains 1 mL 37% HCl and 500 mg Coomassie brilliant blue in 1 L Milli-Q water. The SDS-PAGE gels were heated in Coomassie staining buffer by microwave for less than 3 min then destained in room temperature by destain buffer (10% acetic acid, 40% ethanol, 50% Milli-Q water) on the shaker.

For Western blotting, the proteins were transferred onto the ImmunBlot® Polyvinylidene fluoride (PVDF) (Bio-Rad, USA) in wet conditions. The PVDF membrane was moistened with 100% methanol for activation then transfer buffer prior to use. The SDS-PAGE gel was placed in close contact with the PVDF membrane and sandwiched between three layers of filter papers by the cassette. The transfer buffer contains 28.8 g glycine, 6.04 g Tris base, 200 mL methanol, and 1.6 L Milli-Q water. The blocking buffer contains 5% bovine serum albumin (BSA).

To prevent non-specific binding, the PVDF membrane with transferred proteins was incubated in the blocking buffer containing 5% bovine serum

albumin (BSA) solution on a shaker at room temperature for 1 h. The proteins with His-tag were later labeled by horseradish peroxidase (HRP)-conjugated anti-His antibody diluted at the ratio of 1:5000 (v/v). After incubating the HRP-His antibody with the PDVF membrane for 45 min, the HRP-His antibody was removed, and the membrane was washed by Tris Buffered Saline with Tween (TBST) for 15 minutes three times.

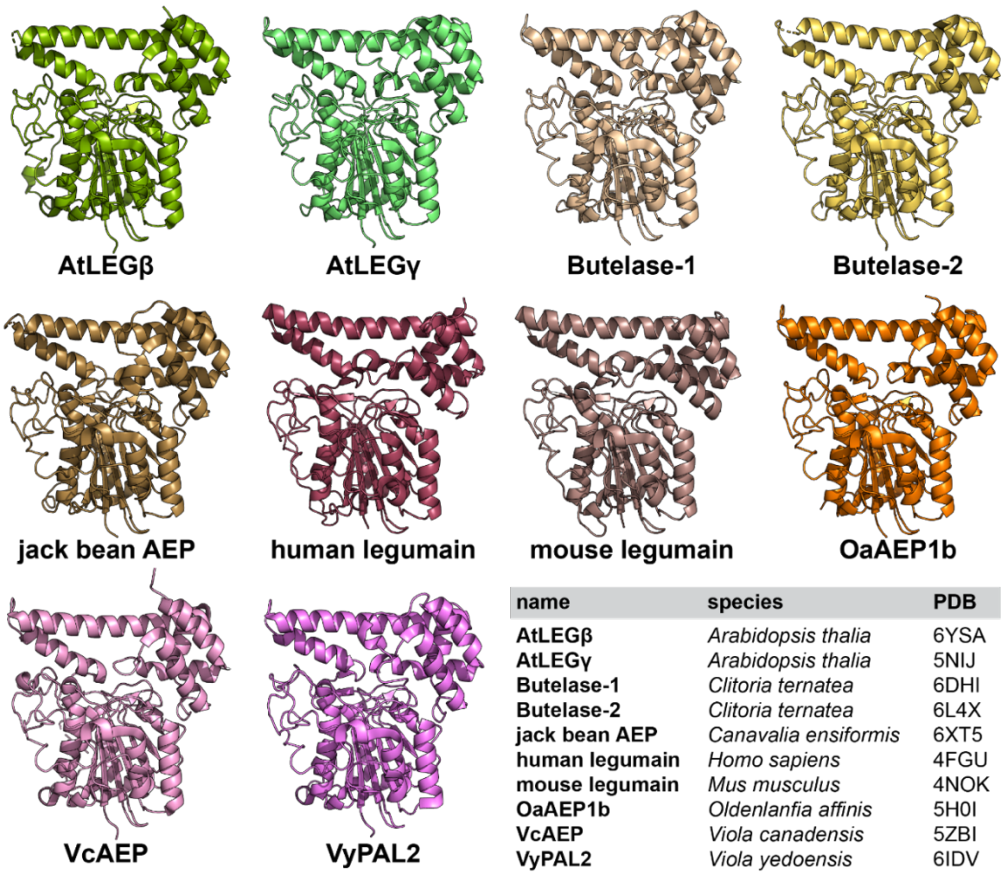
The labeled proteins were visualized by ChemiDoc Imaging Systems (Bio-Rad, USA) or Enhanced chemiluminescence (ECL). The ECL reagent A contains 50 mL 1M Tris-HCl, pH 8.5, 1.1 mL 90 mM p-coumaric acid in dimethyl sulfoxide (DMSO), 2.5 mL 250 mM luminol in DMSO, and 450 mL Milli-Q water. The ECL reagent B contains 1 mL 30% H_2O_2 and 9 mL Milli-Q water. 1 The membranes were incubated in the mixture of 1 mL ECL reagent and 3 μL reagent B, and visualized by medical X-ray film (Kodak, USA) using the Kodak X-OMAT 2000 processor.

Chapter 4 Discovery of Ligase Activity Determinants by Data Mining and Bioinformatics Analysis

4.1 Introduction

Thus far, several unique crystal structures of AEPs and PALs have been reported with similar α-carbon backbone structures [10, 122-131] (**Figure 36**), and the substrate-binding pockets of the catalytic domain of PALs and AEPs have been identified from crystal structures of activated AtLEGγ bound to the Ac-YVAD-CMK inhibitor (PDB accession code: 5OBT) [128] and human cystatin E/M (PDB accession code: 4N6N) [125]. Structurally, examination of the non-primed substrate-binding pockets S4-S1 offers no clear clue pertaining to the molecular basis to distinguish a PAL from an AEP.

The sequence identities of PALs and AEPs frequently do not reflect the similarity of the enzymatic activity preferences of the PALs and AEPs. For example, the sequence identities of the enzymes listed in **Figure 36** range from 34.4% to 82.8% (**Table 5**), however, there is no significant difference shown by the crystal structures. Amino acid sequences from the same plant family or species usually share higher sequence identities and similarities and are located close on the sequence-based phylogenetic tree, such as OaAEP1b, a ligase, and OaAEP2, a protease [7]. However, these enzymes of similar sequences frequently possess very different enzymatic activities.



name	species	PDB
AtLEGβ	*Arabidopsis thalia*	6YSA
AtLEGγ	*Arabidopsis thalia*	5NIJ
Butelase-1	*Clitoria ternatea*	6DHI
Butelase-2	*Clitoria ternatea*	6L4X
jack bean AEP	*Canavalia ensiformis*	6XT5
human legumain	*Homo sapiens*	4FGU
mouse legumain	*Mus musculus*	4NOK
OaAEP1b	*Oldenlanfia affinis*	5H0I
VcAEP	*Viola canadensis*	5ZBI
VyPAL2	*Viola yedoensis*	6IDV

Figure 36. Crystal structures of PALs and AEPs. The crystal structures were obtained by multiple groups [10, 123, 126, 128-131, 189] and are available on the Protein Data Bank (PDB). The details of the crystal structures are shown on the right bottom corner of the figure. The crystal structures were aligned using PyMol Molecular Graphics System, Version 2.0 Schrödinger, LLC.

Table 5. Sequence homology chart of PALs and AEPs. Sequence identity (%) of the enzymes was calculated by EMBL-EBI EMBOSS Water Pairwise Sequence Alignment (available online at https://www.ebi.ac.uk/Tools/psa/emboss_water/) with default setting. See section 3.1.3 for the accession codes used. See **Appendix B** for the sequence similarity.

Enzyme	AtLEGβ	AtLEGγ	Butelase-1	Butelase-2	Jack bean AEP	Huamn legumain	Mouse legumain	OaAEP1b	VcAEP	VyPAL2
AtLEGβ	100.0									
AtLEGγ	61.8	100.0								
Butelase-1	54.3	64.9	100.0							
Butelase-2	66.3	58.3	54.0	100.0						
Jack bean AEP	68.5	58.4	54.2	82.7	100.0					
Huamn legumain	40.8	38.9	35.4	39.1	39.5	100.0				
Mouse legumain	40.3	39.2	36.2	37.8	39.2	82.8	100.0			
OaAEP1b	54.6	65.3	66.0	52.7	53.5	39.4	39.2	100.0		
VcAEP	51.6	61.2	61.6	52.8	50.6	36.9	37.5	57.9	100.0	

VyPAL2	49.0	63.2	65.8	51.7	50.5	35.2	34.4	62.7	65.3	100.0

The evolutionary trace (ET) analysis identifies functionally and structurally important residues of protein homologs by scalable computational methods [180]. ET analysis was designed to eschew the exhaustive mutational studies by predicting the evolutionary important residues for ligand binding, catalysis, or other important functions of a protein [190]. In a sequence-based phylogenetic tree, a residue of the sequence is considered evolutionary important when its variations are correlated with the divergences [191, 192]. For example, the residues conserved among a cluster of evolutionary close species, but not conserved among other clusters, would be ranked as important residues.

Previously, universal evolutionary trace (UET) analysis and computational studies on the structural dynamics of 1627 sequences of ATPases revealed that evolutionary conservation correlates with structural mobility. The more conserved residues are usually less mobile [193]. The correlation between sequence variability and conformational mobility was also reported in 34 enzymes, such as uracil-DNA glycosylase [194]. Similarly, catalytic sites and the substrate-binding pockets of AEPs may be highly conserved and enjoy less mobility compared to other variable residues. We hypothesized that the evolutionary important residues that are not invariant and allow flexibility at the substrate-binding pockets may be the determinants of ligase activity of PALs.

This chapter describes the discovery and characterization of the evolutionary important sequence motifs that may indicate the ligase activity of PALs by UET analysis, sequence conservation analysis, and visual Correlated Mutation Analysis Tool (visualCMAT). The sequence motifs identified were located at the substrate-binding pocket S2 and S1', and termed, ligase activity determinants 1 and 2 (LAD1 and 2), respectively. By applying the ligase activity

determinants (LADs), we could predict the butelase-1-like PALs from a vast pool of AEP sequences and enhance the ligase activity of AEPs with predominant protease activity.

4.2 Result

4.2.1 Generation of Dataset of 1570 Sequences

The amino acid sequences of butelase-1 (GenBank accession code: KF918345) and OaAEP1b (GenBank accession code: KR259377) were used as queries to mine the transcriptome of plants in 1000 Plant (OneKP) Transcriptome Database, the non-redundant protein sequence (nr) databases, transcriptome shotgun assembly (TAS) database, and nucleotide collection (nr/nt) database of National Center for Biotechnology Information (NCBI). To exclude the truncated and incomplete sequences from the databases, the threshold parameters were set to higher than 60% identity and higher than 90% sequence coverage as compared to the query in all BLAST searches. Blastp, using the amino acid sequences of butelase-1 and OaAEP1b to mine the putative amino acid sequences of PALs and AEPs, was performed against the OneKP Transcriptome Database [176] and NCBI non-redundant protein sequence (nr) databases. TBlastn was performed against the NCBI transcriptome shotgun assembly (TAS) database and nucleotide collection (nr/nt) database to retrieve the nucleic acid sequences of putative PALs and AEPs. The retrieved nucleotide sequences from the NCBI transcriptome shotgun assembly (TAS) database and nucleotide collection (nr/nt) database were translated in all six open reading frames (ORFs) using the ExPASy translate tool [177]. The longest ORF among all six ORFs was extracted for each sequence. Data from various databases were pooled and summarized in an excel sheet. Repeated homologs and sequences without the catalytic Cys and His, the characteristics of the C13 family of cysteine proteases (clan CD) (EC 3.4.22.34) [1, 22, 23] were removed. A dataset of 1570 unique sequences of putative PALs

and AEPs retrieved from the NCBI database and OneKP databases was subsequently generated.

The 1570 sequences were distributed among 898 species in 259 plant families (**Figure 37**). They include the Poaceae family, which is the family containing the largest number of sequences of this dataset, comprising 12.5% (197/1570) of the sequences from 75 species, such as rice (*Oryza sativa*), wheat (*Triticum aestivum*), and barley (*Hordeum vulgare*). The second-largest family in this dataset is the Fabaceae family, or the legume family, which comprises 8.5% (134/1570) sequences, such as the pea (*Pisum sativum*), common bean (*Phaseolus vulgaris*), and butterfly pea (*Clitoria ternatea*), where the prototypic PAL butelase-1 was discovered. The third-largest family in this dataset is the Asteraceae family, or the daisy family, which contains 80 sequences from 39 species, including sunflower (*Helianthus annuus*), lettuce (*Lactuca sativa*), and milk thistle (*Silybum marianum*).

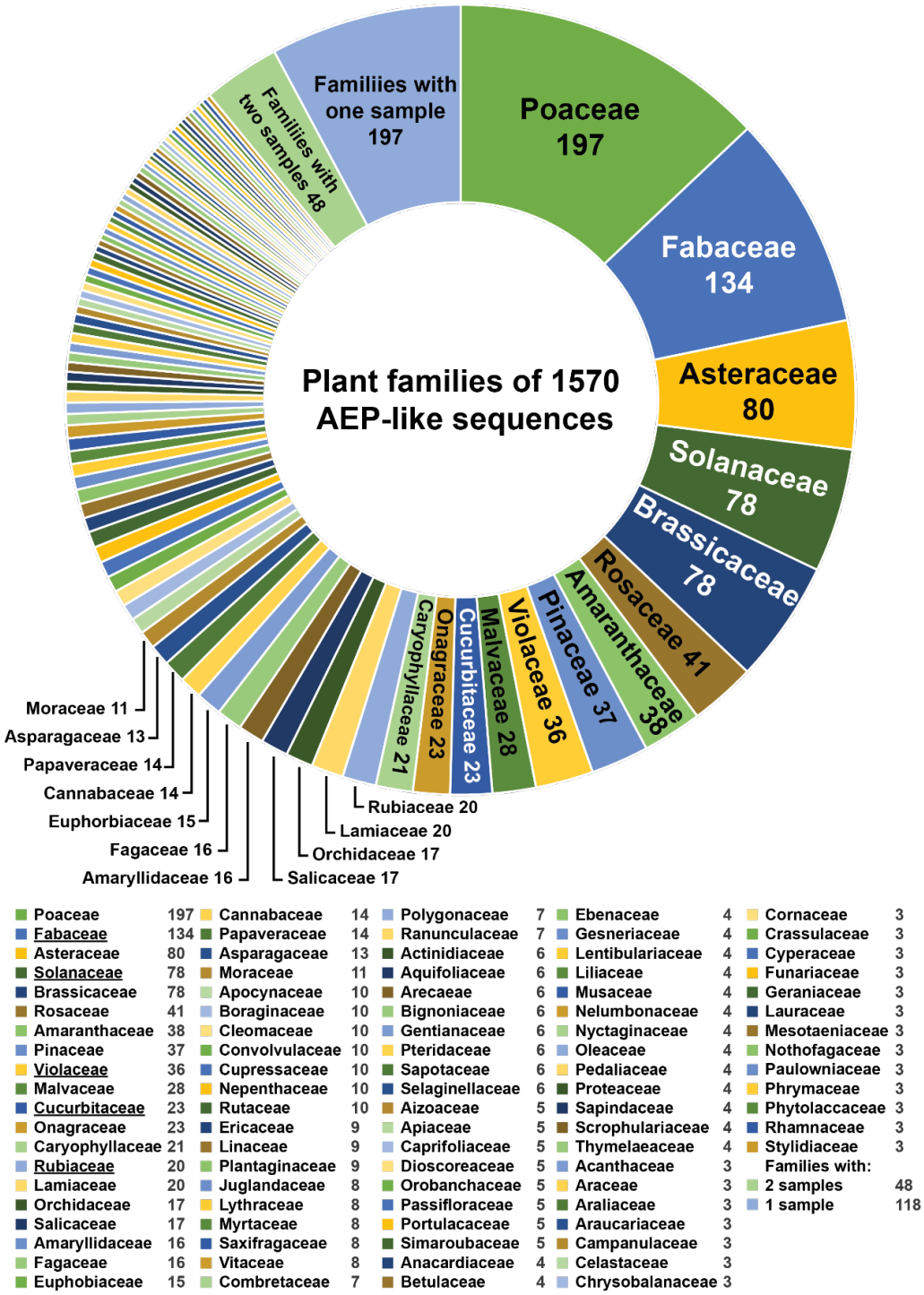


Figure 37. The distribution of families of the 1570 sequences.

Only 296 sequences (18.9%) from 120 species are from the five plant families that are known to produce cyclotides. 23 sequences are from nine species of the Cucurbitaceae family, including watermelon (*Citrullus lanatus*), muskmelon (*Cucumis melo*), and bitter melon (*Momordica charantia*). There are 134 sequences from 60 species of the Fabaceae family, the second-largest family in this dataset and the largest family among five cyclotide-producing families. 20 sequences of this dataset are from nine species of the Rubiaceae family, including Arabian coffee (*Coffea arabica*), Robusta coffee (*Coffea canephora*), and *Oldenlandia affinis*, from where the PAL OaAEP1b was discovered. Solanaceae family includes many common plants, such as eggplant (*Solanum melongena*), pepper, potato (*Solanum tuberosum*), tobacco, and tomato (*Solanum lycopersicum*), and AEP sequences of these common plants are included in this dataset. 78 sequences of the Solanaceae family are from 32 species, and they include three species of peppers of the genus *Capsicum*, *Capsicum annuum*, *Capsicum baccatum*, and *Capsicum chinense*, as well as five species of the genus *Nicotiana*, wild tobacco (*Nicotiana attenuata*), important plant model (Nicotiana benthamiana), desert tobacco (*Nicotiana obtusifolia*), woodland tobacco (*Nicotiana sylvestris*), cultivated tobacco commonly used for smoking (*Nicotiana tabacum*), and *Nicotiana Tomentosiformis*. *Viola* is the genus of most of the 36 sequences from ten species of the Violaceae family, and only three sequences from *Hybanthus enneaspermus* are not members of the genus *Viola*.

Importantly, there are 509 plant species of 197 under-represented plant families in this dataset that contain only one sequence. 72.4% (1137/1570) of the sequences in this dataset are eudicots, and 71.3% (363/509) of the sequences of species containing only one sequence also belong to eudicots. Similarly, 18.7%

(293/1570) of the sequences in this dataset are monocots, and 11.8% (60/59) of the species containing only one sequence are monocots. This result indicates that the distributions of eudicots and monocots in the dataset and in the under-represented groups, the plant species that contain only one sequence in this dataset, are similar. Notably, all the species of ferns as well as most of the species of green algae (Chlorophyta and Charophytes) and Bryophytes, including liverworts, hornworts, and mosses, in this dataset contains only one sequence. Notably, sequence identity and similarity of evolutionary distant PALs and AEPs do not differ significantly. For example, the sequence identity and similarity of butelase-1 and 11 ferns of this dataset range from 49.5% to 55.3% and 66.5% to 73.3%, respectively. While the sequence identity and similarity of butelase-1 and butelase-2 from same species are 54.0% and 72.0% (**Table 5**, **Appendix B**).

4.2.2 Universal Evolutionary Trace Analysis of 1570 AEP Sequences

The universal evolutionary trace (UET) analysis identifies evolutionarily important residues based on the diversity of amino acid composition of one particular position in the sequence. A position is considered evolutionary important when the amino acid composition is diverse among evolutionary distant sequences, while a position is considered evolutionary unimportant when it is diverse among evolutionary close sequences. [180].

To determine the pivotal residues that regulate ligase activity from other residues, UET was used to analyze the 1570 sequences of the dataset. The sequences were aligned using JalView 2.10.5 [139, 140] with Clustal Omega [138] with BLOSUM62 substitution matrix and default gap penalties. The aligned sequences were then submitted for UET analysis. A sequence-based phylogenic tree containing 1570 sequences was generated based on multiple

sequence alignment and visualized using iTOL Interactive Tree of Life (iTOL 6.1.2, Tree of Life version 1.0) (**Figure 38**) [182, 183]. The classification of the sequences on the phylogenic tree was based on that by Yamada *et al.* [91]. The majority of the sequences are from eudicots, which comprises 72.4% (1137/1570) of the sequences. 18.7% (293/1570) of the sequences is from monocots, which formed two major clusters on the phylogenic tree. The remaining 140 sequences were classified into four groups, basal angiosperms, Tracheophytes, including gymnosperm, Lycopodiophyta, and ferns, Bryophytes, including liverworts, hornworts, and mosses, and green algae, including Charophyta and Chlorophyta. Sequences from Bryophytes, ferns, and gymnosperm formed only one major cluster separately, and a cluster of sequence from the Violaceae family, including VyPAL1-5 and VcAEP, located between the major cluster of sequences from Bryophytes and the cluster of sequences from gymnosperm. Interestingly, together with seven other sequences from eudicots, the sequence of butelase-1 located far away from other sequences from Fabaceae family and between sequences from monocots (**Figure 38**).

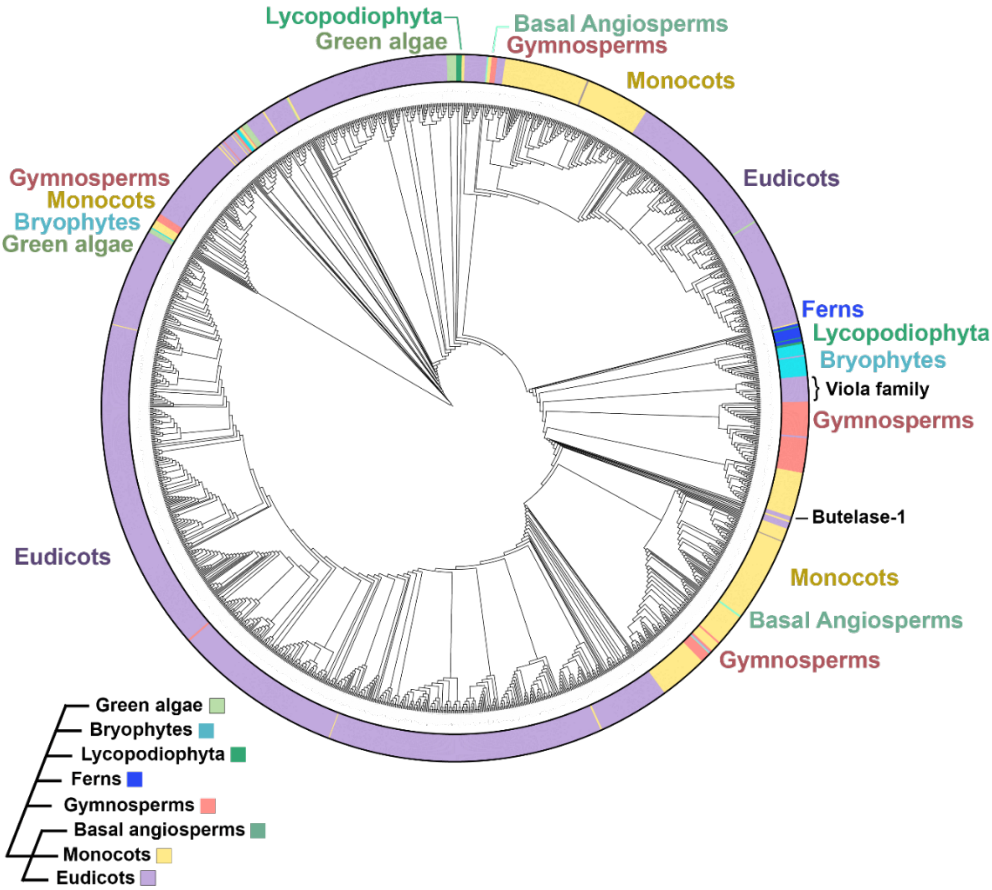


Figure 38. Sequence-based phylogenetic tree of 1570 AEP-like sequences generated for UET analysis. The protein sequences were aligned by MUSCLE algorithm [195] using JalView 2.10.5 [139, 140]. The phylogenetic tree was generated automatically by the UET server using the UPGMA method. The phylogenetic tree was visualized using iTOL Interactive Tree of Life [182, 183].

The real-value Evolutionary Trace (rvET) scores measure the phylogenetic divergence associated with substitution at a particular position based on the sequence-based phylogenetic tree [196]. The rvET scores were allocated to each residue based on the level of diversity of particular residues among evolutionary distant sequences. A completely conserved residue is scored as 1.0, indicating that the residue is likely to possess a critical function of the protein, and substitution or mutation of this position might result in major change or loss of function of the protein. In contrast, a high rvET score at a particular position indicates the presence of a diverse residue among evolutionarily close analogs. The sequence variability, quantified by rvET score, across 1570 AEP sequences was mapped onto the crystal structure of OaAEP1b zymogen (PDB accession code: 5H0I) [126] using PyMol Molecular Graphics System, Version 2.0 Schrödinger, LLC (**Figure 39**).

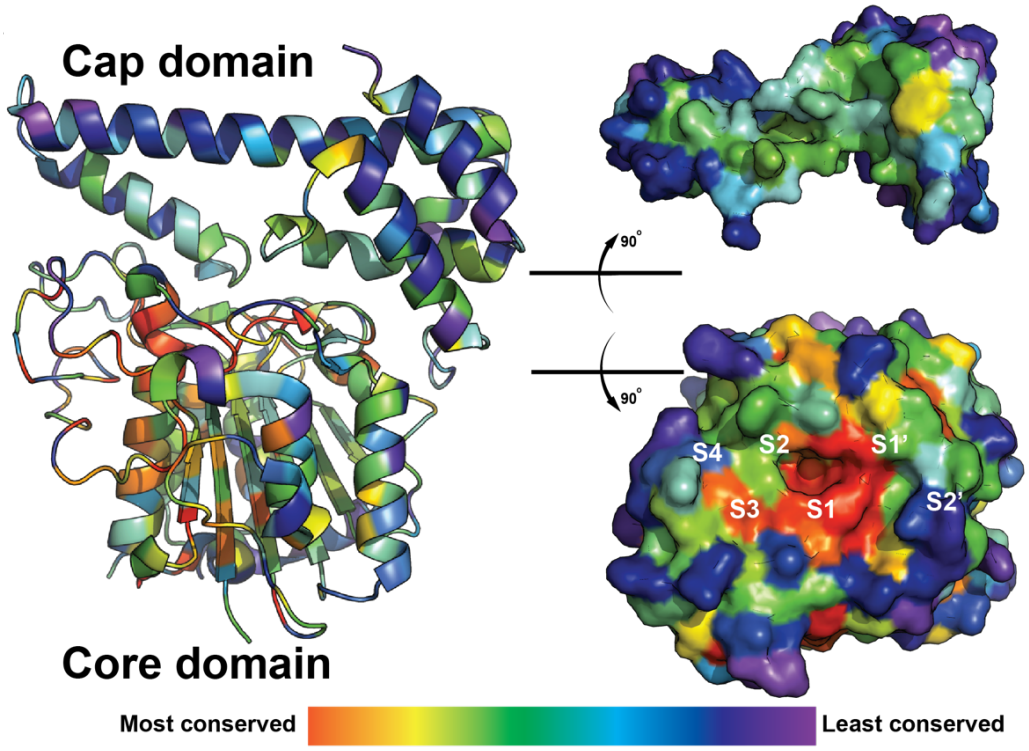


Figure 39. Structure of OaAEP1b colored according to the rvET scores among 1570 sequences. Residues highlighted in red and purple refer to the most and least conserved positions, respectively. The majority of the residues near the oxyanion hole and catalytic sites are highly conserved (highlighted in red and orange). PDB accession code of OaAEP1b: 5H0I.

Figure 40 shows the residues of the six substrate-binding pockets, the colors are based on the rvET scores and the overall height of the stacks of amino acids at one position generated by WebLogo 2.8.2 suggests the conservation of the substrate-binding pockets based on multiple sequence alignment. The height of the symbols of amino acids within a position indicates the relative frequency of the amino acids at this position. For example, position Tyr188 (OaAEP1b numbering) of the substrate-binding pocket S2' and position Gln253 of substrate-binding pocket S4 are colored in dark blue and exhibited a short stack, indicating that residues at these two positions are not evolutionary important and not conserved, respectively. Most residues at or near the oxyanion hole are evolutionary important with a low rvET score ranging from 1.0 to 4.73 (highlighted in red and orange in **Figure 40**). These residues include Arg72, His73, Glu215, Ser245, Asp267 of substrate-binding pocket S1 (OaAEP1b numbering), and the catalytic residues Asn70, Cys217, and His175, the characteristic of the C13 family, clan CD, of cysteine proteases (EC 3.4.22.34) [1, 22, 23]. The rvET scores decrease radially (colored from red to blue) away from the catalytic sites at the center of the surface, suggesting that the evolutionary importance of the residues decreases gradually when moving away from the catalytic sides.

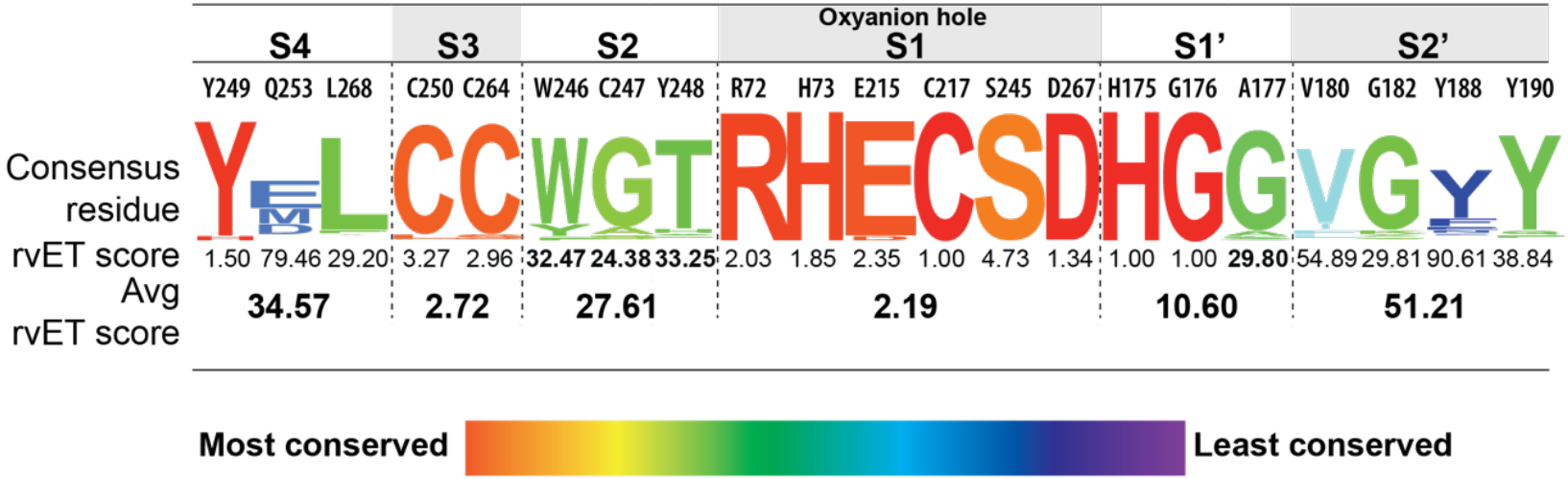


Figure 40. The rvET scores of residues of the six substrate-binding pockets. The numbering is according to OaAEP1b. Completely conserved residues were scored as 1 (red), while higher rvET scores indicate less evolutionary important residues (colored in green, light blue, and dark blue). The consensus sequence was generated by WebLogo 2.8.2, which is available online at: https://weblogo.berkeley.edu/logo.cgi [185, 186].

Compared to the invariant residues (highlighted in red in **Figure 39** and **Figure 40**) at the substrate-binding pocket S1, residues of moderate variations (highlighted in green in **Figure 39** and **Figure 40**) are found in the substrate-binding pocket S4, S2, S1', and S2'. They include Trp246, Val247, and Thr248 of the S2 pocket, Leu268 of the S4 pocket, Ala177 of the S1' pocket, Ala178 near the S1' pocket, Gly182 and Tyr190 of the S2' pocket (OaAEP1b numbering). Previously studies by Liu *et al.* show that sequence evolution correlates with structural dynamics [193, 194]. It was found that the evolutionary conservation of 1627 ATPases correlates with structural mobility, and the more conserved residues are usually less mobile [193]. On the other hand, the conservation and potential mobility of residues of the substrate-binding sites tend to be more variable [193, 197]. Similarly, catalytic sites of PALs and AEPs may enjoy less mobility compared to other residues located further away from the oxyanion hole, and residues of moderate evolutionary importance may allow flexibility for the binding of variable substrates. We thus propose that these residues with moderate evolutionary important may play important roles in the substrate preferences and enzymatic directionality of PALs and AEPs.

Notably, the UET analysis using only 40 selected sequences (**Table 6**), which include 27 reported PALs and AEPs, revealed no significant residue (colored in red and orange) (**Figure 41**). The UET analysis using 291 sequences from the known cyclotide-producing family resulted in similar pattern observed in that using 1570 sequences. For all three UET analyses using 40, 291, and 1570 sequences, there were no evolutionary important residues identified in the C-terminal cap domain, suggesting that the cap domain may not be essential for the enzyme to function.

Table 6. List of selected 40 sequences of putative PALs and AEPs for UET analysis.

Family	Species	Name	Accession no.	Activity reported by
Brassicaceae	*Arabidopsis thaliana*	AtLEGα	NM_128154.5	-
Brassicaceae	*Arabidopsis thaliana*	AtLEGβ	NM_104948.4	[130]
Brassicaceae	*Arabidopsis thaliana*	AtLEGγ	NP_195020.1	[128]
Fabaceae	*Canavalia ensiformis*	CeAEP	P49046	[141]
Fabaceae	*Clitoria ternatea*	Butelase-1	KF918345.1	[6]
Fabaceae	*Clitoria ternatea*	Butelase-2	ALL55651.1	[131]
Malvaceae	*Gossypium raimondii*	GrAEP	XP_012448355.1	[8]
Asteraceae	*Helianthus annuus*	HaAEP	KJ147147.1	[141]
Violaceae	*Hybanthus enneaspermus*	HeAEP1	AWD84473.1	[8]
Violaceae	*Hybanthus enneaspermus*	HeAEP2	AWD84470.1	[8]
Violaceae	*Hybanthus enneaspermus*	HeAEP3	AWD84474.1	[8]
Cucurbitaceae	*Momordica charantia*	-	XP_022148051.1	-
Cucurbitaceae	*Momordica charantia*	-	XP_022131350.1	-
Cucurbitaceae	*Momordica charantia*	-	XP_022156460.1	-
Cucurbitaceae	*Momordica charantia*	-	XP_022148043.1	-
Rubiaceae	*Oldenlandia affinis*	OaAEP1b	KR259377.1	[7]
Rubiaceae	*Oldenlandia affinis*	OaAEP2	KR259378.1	[7]
Rubiaceae	*Oldenlandia affinis*	OaAEP3	KR259379.1	[9]
Rubiaceae	*Oldenlandia affinis*	OaAEP4	LQ854853.1	[9]
Rubiaceae	*Oldenlandia affinis*	OaAEP5	LQ854855.1	[9]

Solanaceae	*Petunia exserta*	-	GBRT01052954.1	-
Solanaceae	*Petunia exserta*	-	GBRT01019050.1	-
Solanaceae	*Petunia x hybrid*	PxAEP1	AWD84471.1	[8]
Solanaceae	*Petunia x hybrid*	PxAEP2	AWD84475.1	[8]
Solanaceae	*Petunia x hybrida*	PxAEP3a	AWD84472.1	[8]
Solanaceae	*Petunia x hybrida*	PxAEP3b	MG720076.1	[8]
Rubiaceae	*Psychotria ipecacuanha*	-	XP_022148051.1	-
Euphorbiaceae	*Ricinus communis*	RcAEP	D17401.1	
Pedaliaceae	*Sesamum iicum*	-	JL339165.1	-
Violaceae	*Viola canadensis*	VcAEP	NJLF_2006210	[10]
Violaceae	*Viola uliginosa*	-	GCAB01004088.1	-
Violaceae	*Viola yedoensis*	VyPAL1	MK085230.1	[10]
Violaceae	*Viola yedoensis*	VyPAL2	MK085231.1	[10]
Violaceae	*Viola yedoensis*	VyPAL3	MK085232.1	[10]
Violaceae	*Viola yedoensis*	VyPAL4	MK085233.1	-
Violaceae	*Viola yedoensis*	VyPAL5	MK085234.1	-
Violaceae	*Viola yedoensis*	VyAEP1	MK085226.1	[10]
Violaceae	*Viola yedoensis*	VyAEP2	MK085227.1	-
Violaceae	*Viola yedoensis*	VyAEP3	MK085228.1	-
Violaceae	*Viola yedoensis*	VyAEP4	MK085229.1	-

A dash (-) indicates that the sequence has not been characterized and the enzymatic activity has not been reported by biochemical assay or *in planta* assay. The sequence of VcAEP was retrieved from OneKP, the rest were from NCBI.

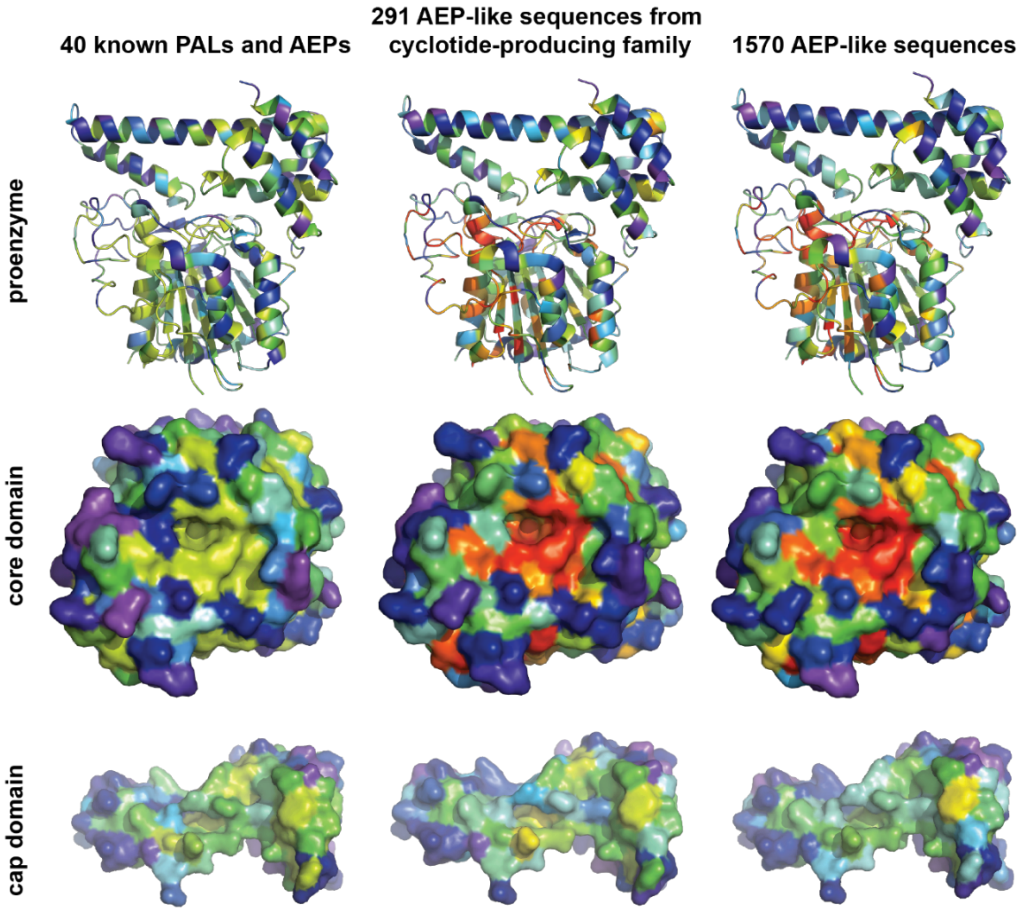


Figure 41. The AEP structures colored based on the rvET scores allocated to each residue by UET analysis. The importance of the substrate-binding pockets was not indicated using UET analysis based on 40 selected sequences of known and putative PALs and AEPs.

4.2.3 Amino Acid Composition of the Substrate-Binding Pockets

To further investigate the abovementioned eight residues of moderate evolutionary importance, sequences of known PALs and AEPs were aligned, and it was found that Val247 of the S2 pocket and the dipeptide Ala-Ala 177-178 around the S1' pocket are distinguishable among PALs and AEPs (**Table 7**). For example, at the position Val247 (OaAEP1b numbering), PALs usually contain a bulky and hydrophobic residue, such as Cys, Ile, or Val, while a Gly is conserved among the AEPs. Meanwhile, other residues with moderate variation (labeled in green in **Figure 39** and **Figure 40**) are not able to distinguish PALs and AEPs. For example, Leu268 is conserved among all the sequences in **Table 7**.

The amino acid composition of the residues of all six substrate-binding pockets of 1570 sequences of the dataset, which includes the known PALs and AEPs, was next analyzed. At the substrate-binding pocket S4, Tyr249 (OaAEP1b numbering) is the most conserved among three residues of the S4 pocket, more than 95% (1494/1570) of the sequences possessing a Tyr at this position. 89.1% (1399/1570) of the sequences harbor a Leu at the position Leu268 (labeled in green in **Figure 39** and **Figure 40**) and 7.4% (117/1570) of the sequences contain a large bulky residue, such as Phe, Trp, Tyr, and Val. At Gln253, the amino acid composition is more diverse, Glu is the most abundant amino acid and only 44.8% (704/1570) of the sequence contains a Glu at this position. However, more than 93.5% (1469/1570) of the sequences at the position Gln253 contain a polar amino acid at this position, such as Asp, Gln, Glu, and Met (**Figure 42**).

Table 7. Amino acid composition of 20 PALs, AEPs, and partial ligases.

	S4 pocket	S3 pocket	S2 pocket (LAD1)	S1 pocket	S1' pocket (LAD2)	S2' pocket
Ligase-type						
Butelase-1	Y---Q.L	C.C	WVT	RH.E-C.S.D	HGGA	V-G-----Y-A
VyPAL1	Y---V.L	C.C	LIA	RH.E-C.S.D	HGAP	K-G-----Y-Y
VyPAL2	Y---T.L	C.C	WIT	RH.E-C.S.D	HGAP	K-G-----Y-Y
OaAEP 1b	Y---Q.L	C.C	WCY	RH.E-C.S.D	HGAA	V-G-----Y-Y
OaAEP 3	Y---Q.L	C.C	WCY	RH.E-C.S.D	HGAP	V-G-----Y-Y
OaAEP 4	Y---Q.L	C.C	WCY	RH.E-C.S.D	HGAP	V-G-----Y-Y
HeAEP 3	Y---Q.L	C.C	WVT	RH.E-C.S.D	HGAP	T-G-----Y-V
VuPAL1	Y---A.L	C.C	WIT	RH.E-C.S.D	HGAP	K-G-----Y-Y
Dual functional						
PiPAL1	Y---Q.L	C.C	WAA	RH.E-C.S.D	HGAP	V-G-----Y-Y
PePAL1	Y---D.L	C.C	WAT	RH.E-C.S.D	HGAP	M-S-----I-A
PePAL2	Y---Q.L	C.C	WVT	RH.E-C.S.D	HGGP	V-G-----Y-Y
SiPAL1	Y---G.L	C.C	YAI	RH.E-C.S.D	HGAA	V-G-----Y-Y
VcAEP1	Y---Q.L	C.C	WVA	RH.E-C.S.D	HGYP	V-G-----Y-Y
PxAEP 3b	Y---Q.L	C.C	WVT	RH.E-C.S.D	HGGP	V-G-----Y-Y
Protease-type						
BmAEP1	Y---Q.L	C.C	WAT	RH.E-C.S.D	HGSA	L-G-----Y-Y
AtLEG γ	Y---E.L	C.C	WGT	RH.E-C.S.D	HGGP	V-G-----S-Y
PxAEP 3a	Y---Q.L	C.C	WGT	RH.E-C.S.D	HGGP	V-G-----N-Y
HeAEP 1	Y---E.L	C.C	WGT	RH.E-C.S.D	HGGP	V-G-----M-Y
Butelase 2	Y---T.L	C.C	WGT	RH.E-C.S.D	HGGP	V-G-----I-Y
OaAEP 2	Y---E.L	C.C	WGT	RH.E-C.S.D	HGGP	V-G-----Y-Y
consensus	Y---E.L	C.C	WGT	RH.E-C.S.D	HGGP	V-G-----Y-Y

Residues in blue are conserved with known PALs, residues in red are conserved with known proteases. Residues highlighted in grey are not at the substrate-binding pockets but located nearby the pockets. Bold enzymes are tested in this study. Period between residue indicates that the residues located more than one residue away from each other on the sequence. Each dash between residue indicates one residue that is not in the substrate-binding pockets.

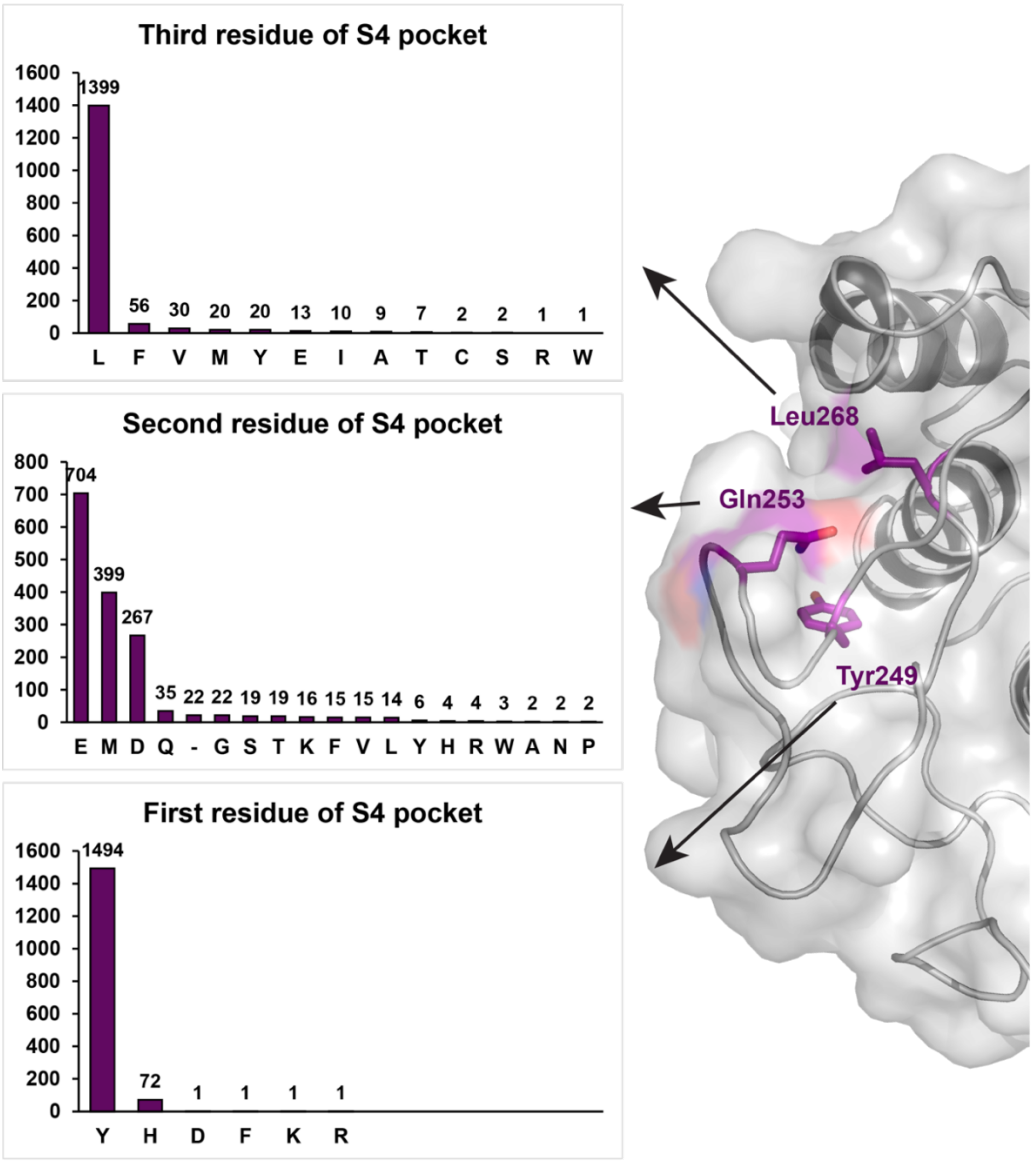


Figure 42. Amino acid composition of the substrate-binding pocket S4. From the N-terminus to the C-terminus of the sequence, they are Tyr249, Gln253, and Leu268 (OaAEP1b-numbering).

Both Cys residues at the substrate-binding pocket S3 are highly conserved, with more than 94% of the sequences in the dataset containing a Cys at both positions (**Figure 43**). The two Cys may be highly conserved because of the formation of the disulfide bridge that stabilizes the enzyme.

At Trp246 (OaAEP1b numbering) of the S2 pocket, more than 99% (1562/1570) of the sequences contain an amino acid with an aromatic side chain, such as Trp, Tyr, and Phe, with Trp as the most abundant amino acid (1299/1570) at this position. At the Tyr248 position, the other side of the S2 pocket, a Thr is the most common residue among 1570 sequences, which is found in more than 87% (1367/1570) of the sequences. At the middle of the S2 pocket (Cys247), a Gly is predominantly present in 85.9% (1348/1570) of the sequences (**Figure 44**). In contrast, only 152 (<15%) sequences comprise hydrophobic residues, including Ala (122/1570), Ser (75/1570), Val (13/1570), Ile (11/1570), Cys (2/1570), and Pro (1/1570). The rare amino acids in the S2 pocket include Val and Ile, which account for only 1.5% (24/1570), as well as Cys, which is only found in two of 1570 AEP sequences from *Oldenlandia affinis*, and Pro from oat (*Avena sativa*) of Poaceae family. The hydrophobic amino acids are commonly found in the PALs, such as Val237 of butelase-1, Ile244 in VyPAL1-2, and Cys247 in OaAEP1b. In contrast, Gly is highly conserved in all AEP homologs with protease activity, such as butelase-2, OaAEP2, CeAEP, and human legumain (**Figure 45**).

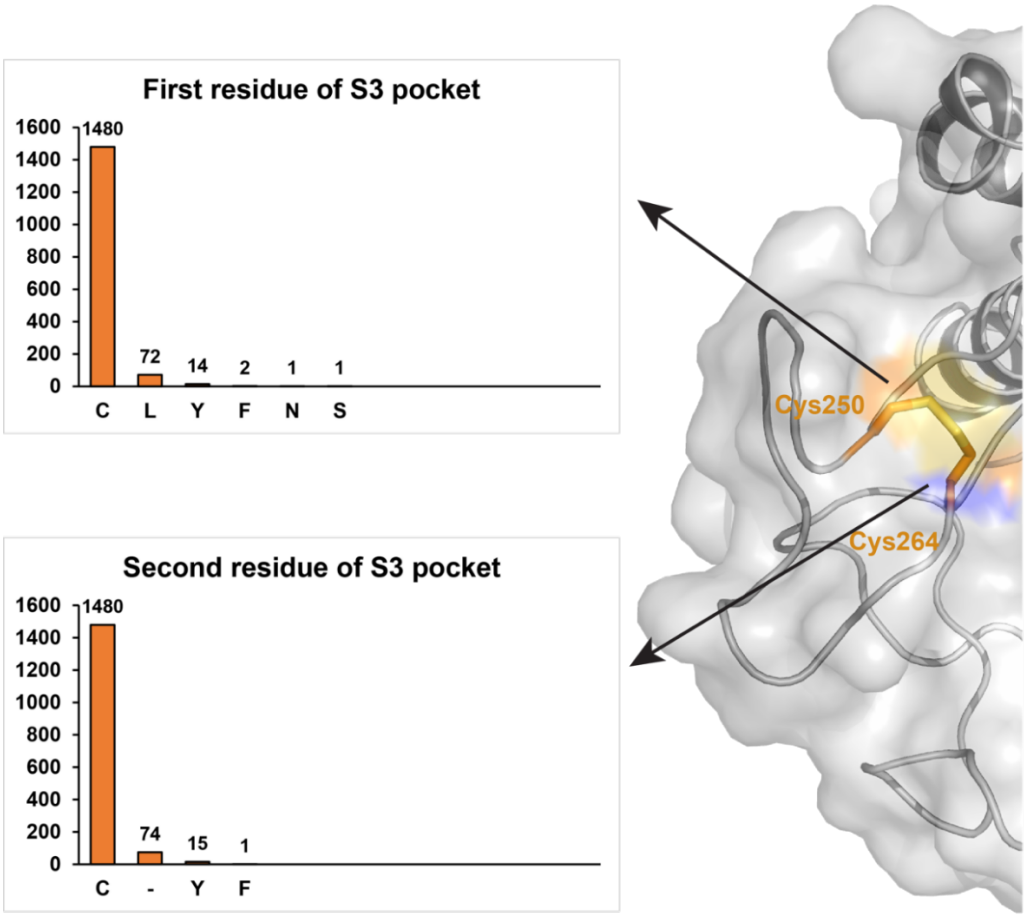


Figure 43. Amino acid composition of the substrate-binding pocket S3. From the N-terminus to the C-terminus of the sequence, they are Cys250 and Cys264 (OaAEP1b-numbering).

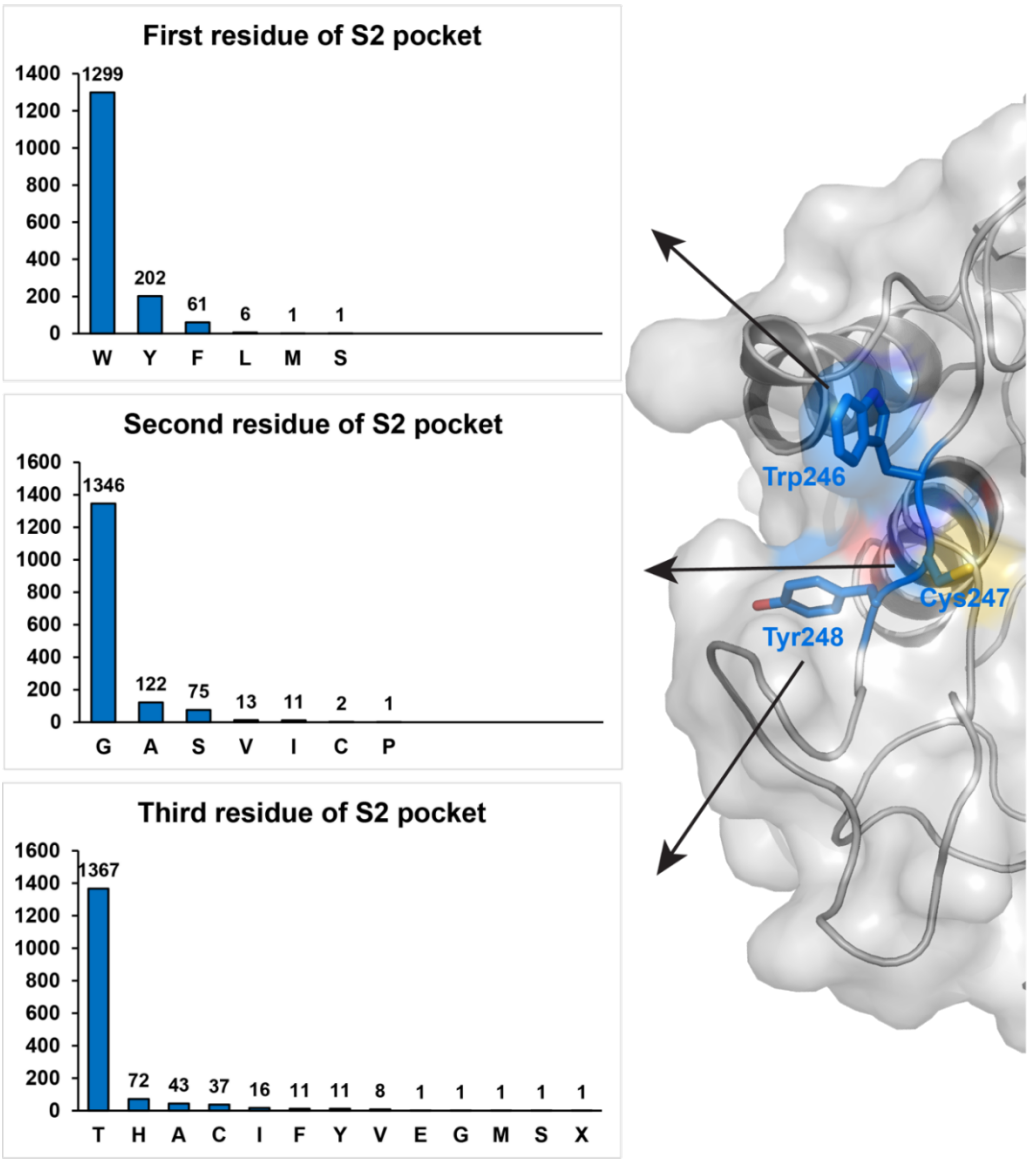


Figure 44. Amino acid composition of the substrate-binding pocket S2. From the N-terminus to the C-terminus of the sequence, they are Trp246, Cys247, and Tyr248 (OaAEP1b-numbering). X of the bar chart at the bottom indicates the amino acid composition of the sequence of *Cynara cardunculus* var. scolymus (NCBI reference sequence: XP_024996221.1), which was derived from the whole genome shotgun sequence (NCBI Reference Sequence: NC_037543.1) and predicted by The NCBI eukaryotic gene prediction tool. X indicates all 20 native amino acids.

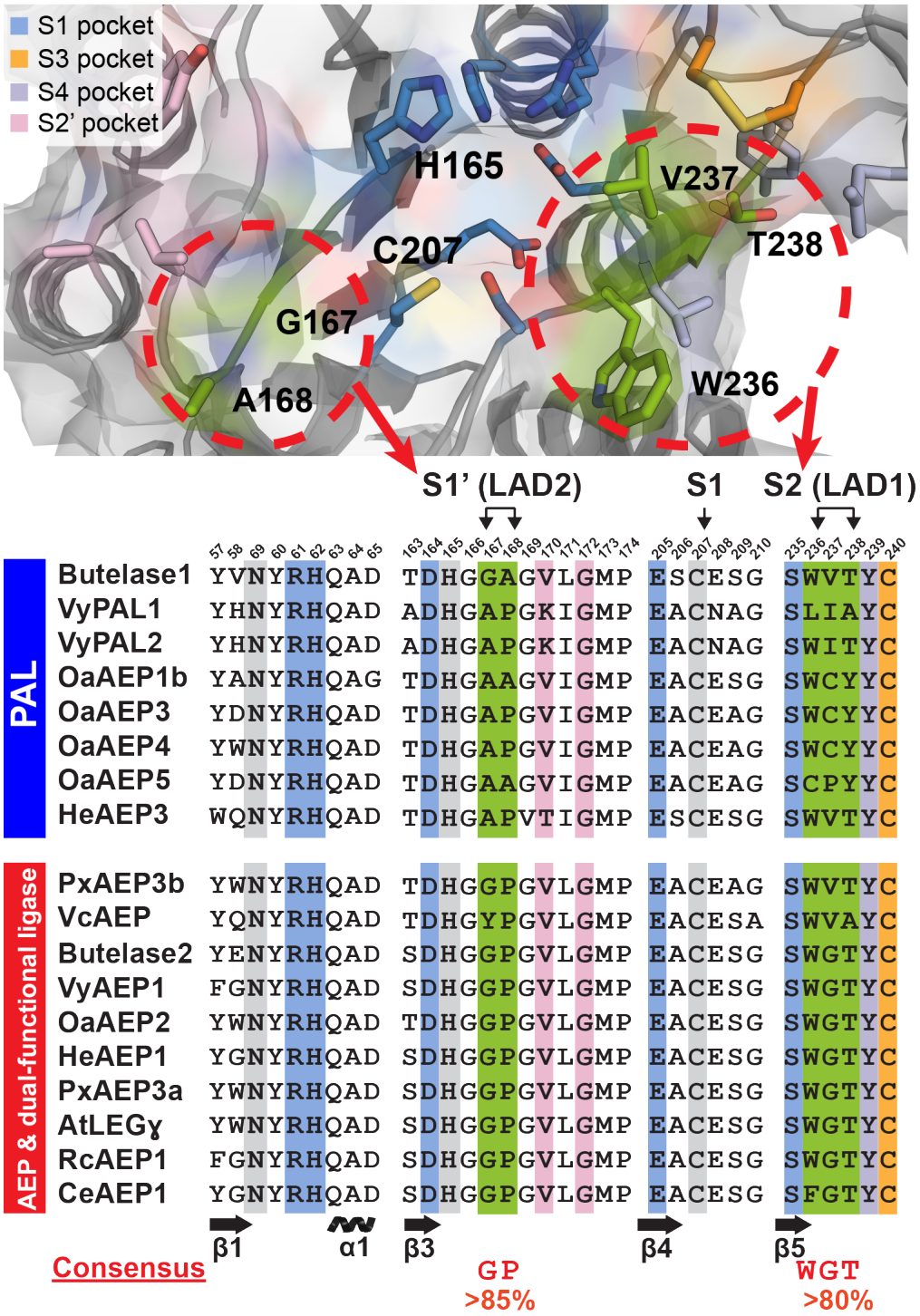


Figure 45. Sequence comparison of PALs and AEPs. Residues at the substrate-binding pockets were extracted from the aligned sequences and compared between PALs and AEPs. In the S2 pocket, Trp-Gly-Thr with Gly in the middle accounts for more than 85% of all sequences. Similarly, Gly-Pro dipeptide accounts for more than 80% of all AEP-like sequences. The sequences were aligned by Clustal Omega (available online at: https://www.ebi.ac.uk/Tools/msa/clustalo/) [138] using JalView [139, 140].

Due to the high conservation and evolutionary importance, all residues of the S1 pocket are nearly invariant and the most abundant amino acid of each position was found in more than 95% of the sequences. The Cys217, one of the characteristics of the C13 family, clan CD, of cysteine proteases, was absolutely conserved among all 1570 sequences, following the Arg72 and His73, which are both found in 1568 out of 1570 sequences. 1561 sequences contain a Ser at the position Ser245 (OaAEP1b numbering). The least conserved residues of the S1 pockets are Glu215 and Asp267, which are found in 1496 and 1495 sequences, respectively (**Figure 46**).

At the primed side of the substrate-binding surface, the characteristic His175 and Gly176 of the S1' pocket are absolutely conserved among all 1570 sequences (**Figure 47**). At position A177 (OaAEP1b numbering) of the S1' pocket, a Gly is conserved in 1418 sequences (90.3%). At the position next to the A177, A178 (OaAEP1b numbering), a Pro is conserved in 1346 sequences (85.7%). Combining these two positions, a Gly-Pro dipeptide is found in 80.3% (1260/1570 of the sequences, whereas less than 20% of the sequences contain non-Gly-Pro motifs, such as Gly-Asp (75/1570), Gly-Ala (73/1570), Ala-Pro (43/1570), Ser-Pro (33/1570), Ala-Ala (13/1570), and Ser-Ala (13/1570) (**Table 8**). A distinguishing feature of PALs is the presence of a dipeptide of small hydrophobic amino acids at S1' pocket, they include Gly-Ala 167-168 in butelase-1, Ala-Pro 174-175 in VyPAL2, and Ala-Ala 177-178 in OaAEP1b, while the dipeptide motif Gly-Pro is conserved in proteases, such as butelase-2 and OaAEP2 (**Figure 45**).

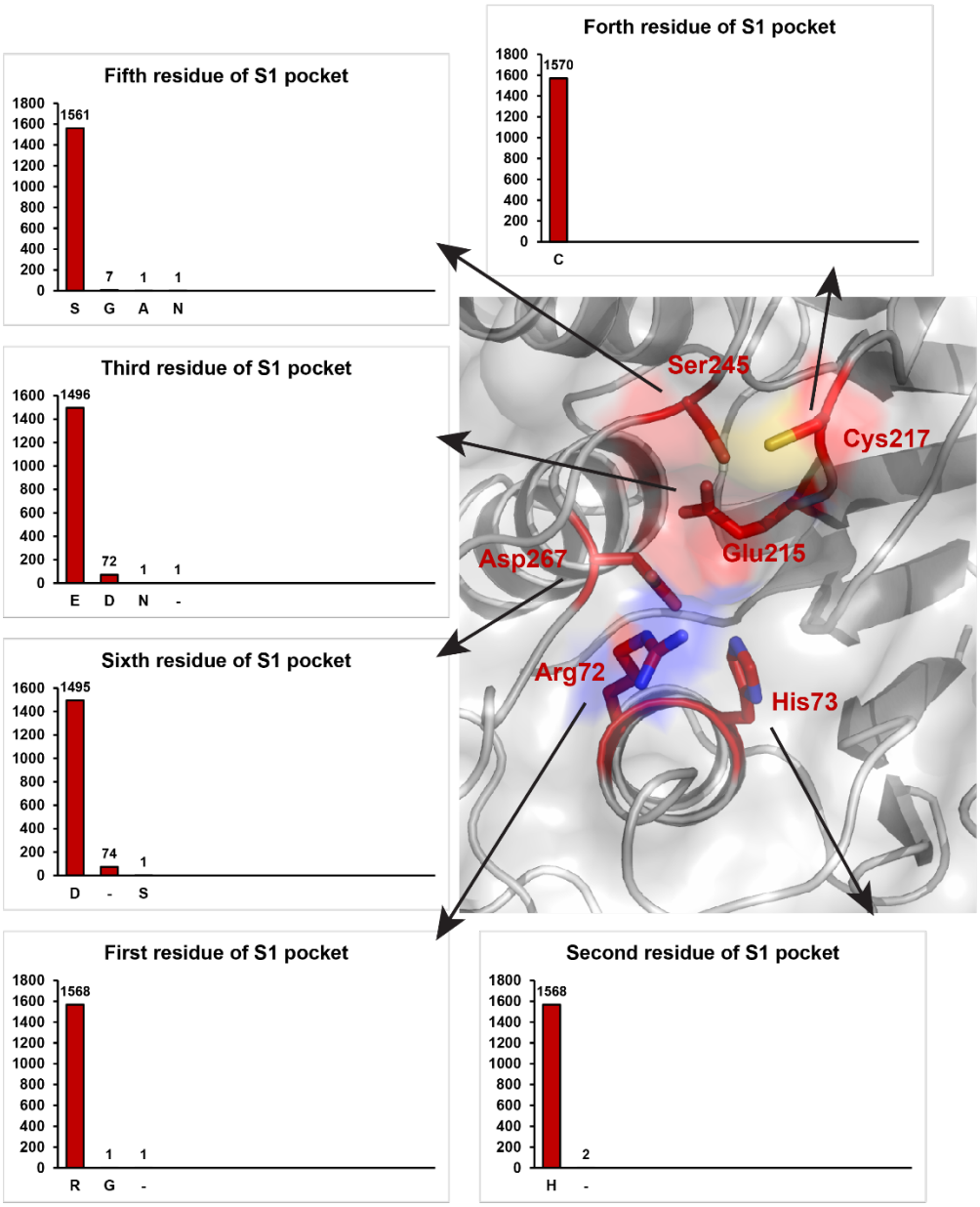


Figure 46. Amino acid composition of the substrate-binding pocket S1. From the N-terminus to the C-terminus of the sequence, they are Arg72, His73, Glu215, Cys217, Ser245, and Asp267 (OaAEP1b-numbering). Among them, Cys217 is the catalytic Cys and the characteristic of the C13 family, clan CD, of cysteine proteases (EC 3.4.22.34) [1, 22, 23].

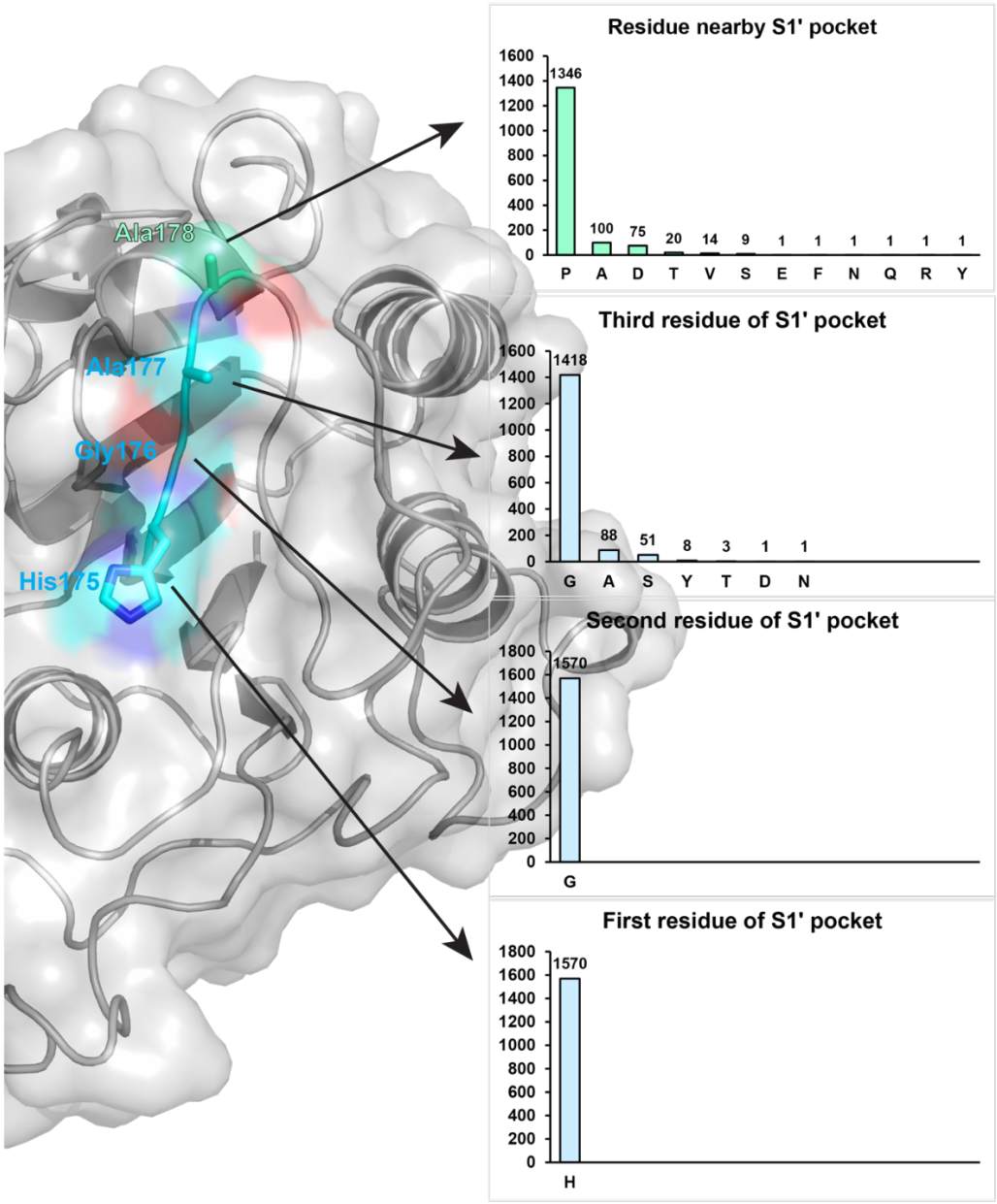


Figure 47. Amino acid composition of the substrate-binding pocket S1'. From N-terminus to C-terminus of the sequence, they are His175, Gly176, and Ala177 (OaAEP1b-numbering). Ala178 is located in close proximity to the S1' pocket and one of the residues of the LAD2.

Table 8. Amino acid composition at the LAD2.

Residue	Number of sequences	Percentage (%)
GP	1258	80.1
GD	75	4.8
GA	73	4.7
AP	48	3.1
SP	33	2.1
AA	13	0.8
SA	13	0.8
AT	12	0.8
AV	9	0.6
YP	7	0.5
AS	5	0.3
GS	4	0.3
GT	3	0.2
GV	3	0.2
TT	3	0.2
AY	1	<0.1
DT	1	<0.1
GF	1	<0.1
GQ	1	<0.1
NV	1	<0.1
SE	1	<0.1
SN	1	<0.1
SR	1	<0.1
ST	1	<0.1
SV	1	<0.1
YA	1	<0.1

At the S2' pocket, there are two residues considered to have moderate variation based on UET analysis, Gly 182 and Tyr 190 (labeled in green in **Figure 48**). 1388 sequences contain a Gly at the position Gly182 (OaAEP1b numbering), and the second and third commonly found amino acids are both non-polar and aliphatic; they are Lys (74/1570) and Ala (40/1570). Tyr190 is the most conserved residue, which is found in 88.9% (1397/1570) of the sequences, following the polar Gln (71/1570) and aromatic Phe (48/1570). A Val and a Tyr are found in 79.7% (1252/1570) and 66.7% (148/1570) of the sequences at position Val180 and Tyr188, respectively (**Figure 48**). These two residues are less conserved and considered not evolutionary important (labeled in blue in **Figure 39** and **Figure 40**).

To summarize, based on this dataset of 1570 sequences including reported PALs and AEPs, only two sequence motifs, the C247 and dipeptide Ala-Ala 177-178, can be used to distinguish PALs from AEPs. This result is in agreement with the previous reports which termed the two sequence motifs at S2 and S1' pocket as 'ligase activity determinant (LAD)' 1 and 2, respectively [10, 131].

The combination of Gly and Gly-Pro at LAD1 and LAD2, respectively, are highly conserved among AEPs. In contrast, a combination of hydrophobic reside (Val/Cys/Ile) and Gly-Ala/Ala-Ala/Ala-Pro dipeptide at LAD1 and LAD2, respectively, are only found in 15 out of 1570 AEP sequences (**Figure 49**). Such a combination of LAD sequences is conserved in the known PALs (LAD1: Val/Ile/Cys; LAD2: Gly-Ala/Ala-Ala/Ala-Pro). These differences could serve as indicators for distinguishing PALs from AEPs.

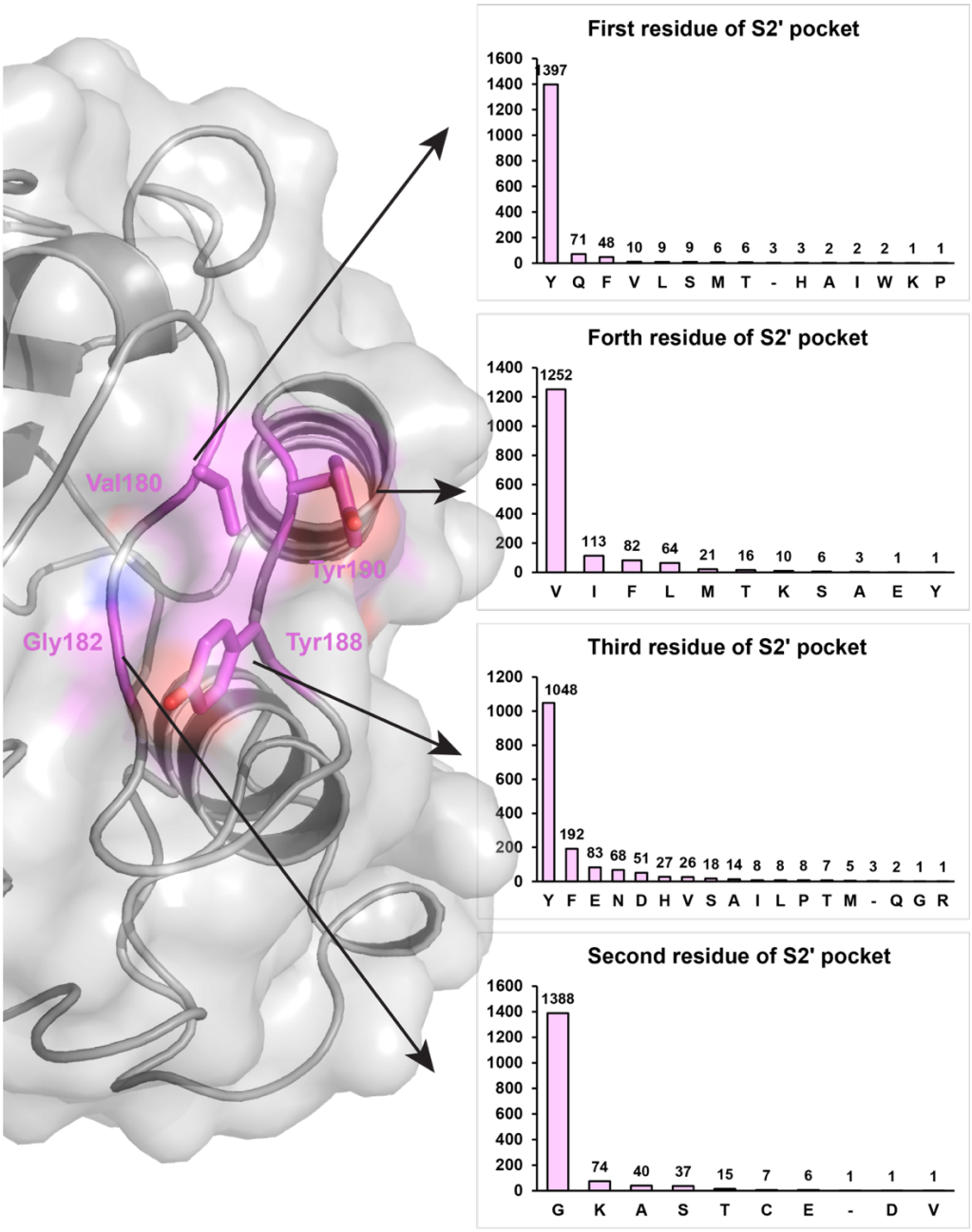


Figure 48. Amino acid composition of the substrate-binding pocket S2'. From the N-terminus to the C-terminus of the sequence, they are Val180, Gly182, Tyr188, and Tyr190 (OaAEP1b-numbering).

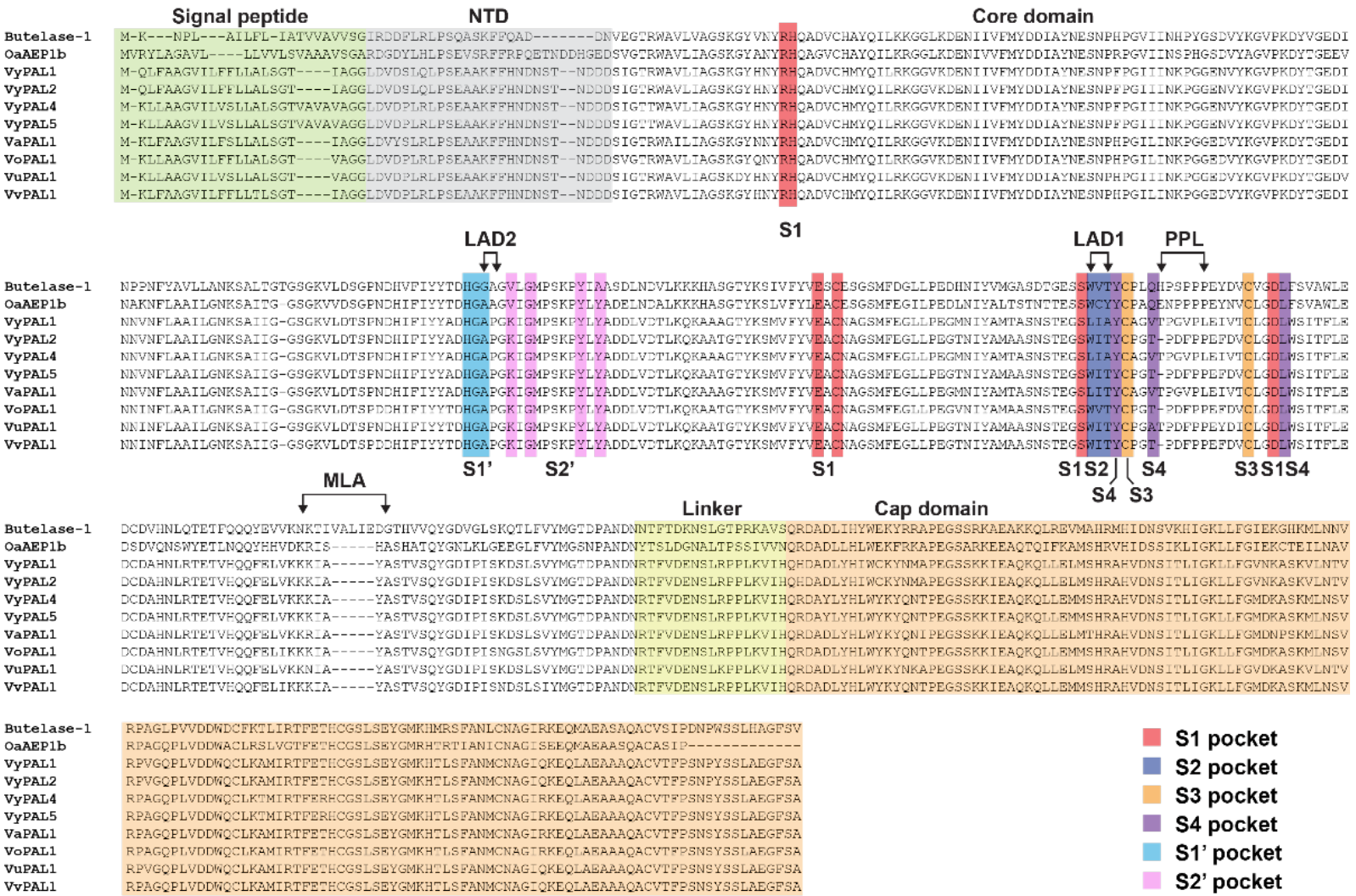

Signal peptide
NTD
Core domain
S1
LAD2
LAD1
PPL
S1'
S2'
S1
S1S2
S4
S3
S3S1S4
MLA
Linker
Cap domain
S1 pocket
S2 pocket
S3 pocket
S4 pocket
S1' pocket
S2' pocket

Figure 49. Multiple sequence alignment of four known PALs and six putative PALs. The reported sequence motifs that indicate ligase activity of PALs were labeled, they are ligase activity determinants 1 and 2 (LAD1 and LAD2), poly-proline loop (PPL), and marker of ligase activity (MLA). The residues of substrate-binding pockets were colored, the S1, S2, S3, S4, S1', and S2' pockets were colored in red, blue, orange, purple, light blue, and pink, respectively. The signal peptide, N-terminal domain, core catalytic domain, linker, and the C-terminal cap domain were colored in green, grey, white, yellow, and orange. The sequences were aligned by Clustal Omega (available online at: https://www.ebi.ac.uk/Tools/msa/clustalo/) using JalView [139, 140].

4.2.4 VisualCMAT Analysis of 1570 AEP Sequences

To further characterize the LADs and the substrate-binding pockets, we performed the visual Correlated Mutation Analysis Tool (visualCMAT) on the aligned 1570 sequences. The visualCMAT predicts the correlated amino acid substitutions in a multiple sequence alignment based on mutual information, the measure of the mutual dependence. The results were visualized using PyMol Molecular Graphics System, Version 2.0 Schrödinger, LLC. The residues labeled in red and linked by dashed lines are the co-evolving residues that are strongly correlated among the 1570 sequences, indicating that substitution of one of the red residues is strongly correlated to the substitution of the other red residue linked by the dashed line. These correlated residues include Lys85 with Asp91, Typ283 with Ser323, Val345 with Asp349, Gln373 with Phe377, Gly394 with Gly399, Lys395 with Ile400, Ser458 with Ala463, Glu459 with Ala463, Ala463 with Ser458, and Ser467 with Ala471 (**Figure 50**). While the residues in grey are only weakly correlated. The other non-labeled residues in blue were analyzed as not correlated.

There is no co-evolving residue predicted at the six substrate-binding pockets and catalytic sites. The residues of both LADs and the substrate-binding pockets were found to be not correlated by visualCMAT analysis, indicating that LADs may not form physical contact, and substitution of residue at one LAD does not influence the other LAD. To sum up, visualCMAT analysis suggests that LAD1 and LAD2 do not have direct physical contact and are not co-evolving residues [184].

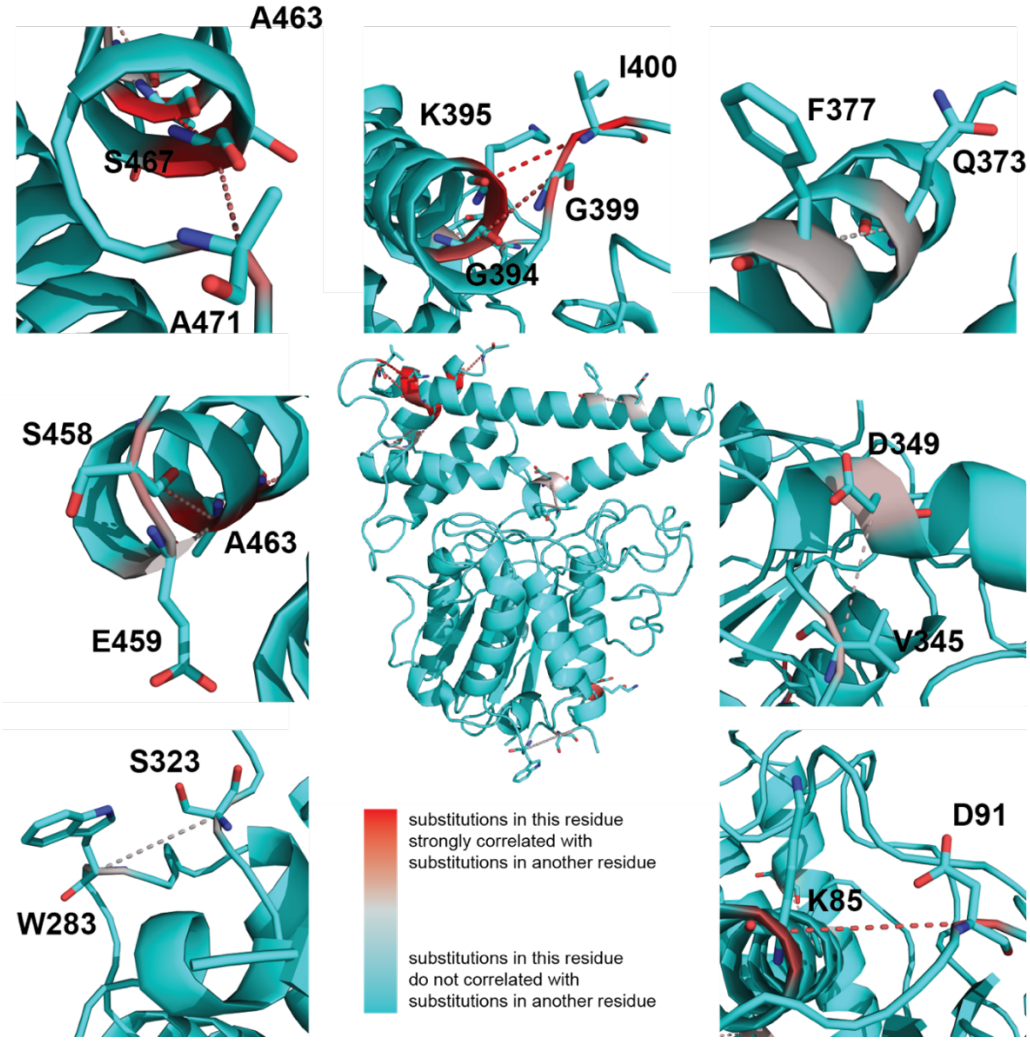


Figure 50. The visualCMAT analysis of correlated residues [184]. The residues in red and linked by dashed lines are strongly correlated among 1570 sequences, the residues in grey are weakly correlated. Substitution of one of the two paired residues labeled in red is predicted to be strongly correlated with another residue. Likewise, the paired residues in grey are predicted to be weakly correlated. The paired residues are predicted to have direct physical contact or interact with the same ligand. In contrast, substituting the blue residues does not correlate with other residues.

4.2.5 Distribution of Putative PALs and AEPs in the Dataset

For clarity, the 1570 sequences of the dataset were categorized into four types based on their amino acid composition at the LADs. They include (1) 15 sequences classified as 'LAD+/+' and predicted to be butelase-1-like ligases, with a Val/Cys/Ile residue at the LAD1 (middle of substrate-binding pocket S2) and a dipeptide Gly-Ala/Ala-Ala/Ala-Pro motif at the LAD2 (substrate-binding pocket S1'), (2) 11 'LAD+/-' sequences of predicted partial ligases that show both ligase and protease activity, with only Val/Cys/Ile at the LAD1, (3) 119 'LAD-/+' sequences, predicted partial ligases with a Gly-Ala/Ala-Ala/Ala-Pro dipeptide at the LAD2, and lastly (4) 1425 'LAD-/-' sequences of predicted butelase-2-like proteases that show predominant protease activity. They do not contain Val/Cys/Ile at the LAD1 nor Gly-Ala/Ala-Ala/Ala-Pro at the LAD2. In total, 145 out of 1570 sequences from 39 plant families and 124 species are predicted to contain ligase activity (**Figure 51**). All of the sequences classified as LAD+/+ and LAD(+/-) are from eudicots, and most sequences are from the Violaceae family, which is in agreement with the abundance of cyclotides found in the Violaceae family (**Appendix A**) [38]. The LAD+/+ include several previously reported PALs, including the prototypic butelase-1 [6], VyPAL1-2 [10], OaAEP1b [7], and HeAEP3 [8]. The largest and second-largest family of sequences labeled as LAD-/+ are Solanaceae (15/119) and Amaranthaceae family (14/119), respectively. More than 75.6% (90/119) of the sequences of the group LAD-/+ are from eudicots, following 16 out of 119 sequences from monocots, eight sequences from gymnosperm, two sequences from green algae, and one sequence from each of basal angiosperm, Bryophytes, and fern.

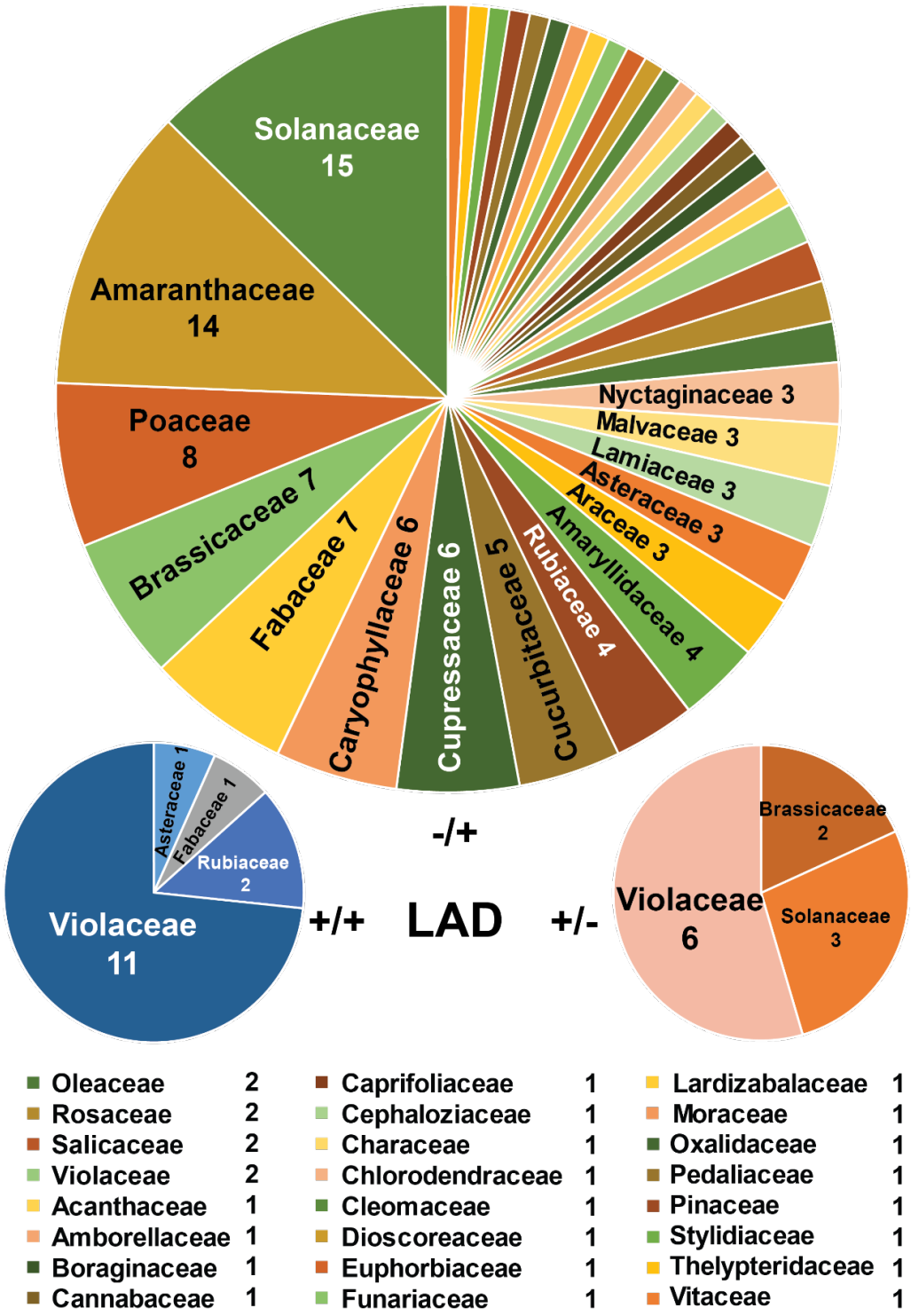


Figure 51. The distribution of 145 sequences from the 39 families and 124 species with one or both LADs conserved with known PALs. Most of the sequences with LAD +/+ or +/- are from the Violaceae family, while most sequences of LAD -/+ are from the Solanaceae and Amaranthaceae family. Notably, the Amaranthaceae family is not known to produce cyclic peptides.

4.3 Discussion

An approach to identify PALs from AEPs by comparison and analysis of 1570 AEP sequences representing 898 plant species and 259 plant families from diverse clades was described. The dataset and analysis included various economically important species, such as rice (*Oryza sativa*), common wheat (*Triticum aestivum*), and potato (*Solanum tuberosum*), facilitating future agricultural studies. The phylogenetic tree based on 1570 sequences of putative PALs and AEPs revealed clusters of eudicots, monocots, ferns, and green algae. These separated clusters suggest the presence of unknown motifs on the AEP sequences that distinguish the AEPs of the less-evolved plants from more-evolved plants. Some of the sequences of the same clades formed multiple clusters, which may be a result of the presence of different types of AEPs, such as the β-type AEPs that are specifically expressed during the seed maturation and embryo development, and the γ-type AEPs frequently expressed during the plant programmed cell death (PCD) [91]. The phylogenetic tree could be combined with evolutionary studies on how AEPs and PALs evolved with the plants from green algae to flowering plants that contain several cyclic peptides.

Analysis of the dataset through universal evolutionary trace (UET) revealed a specific pattern of variations of the amino acid compositions in the substrate-binding pockets. At the S4 pocket, located further away from the oxyanion hole, residues were the least conserved and inferred unimportant by UET, suggesting that its role in modulating enzyme activity is relatively minor. Similarly, the S2’ pocket does not differ between ligases and proteases, and residues of the S2’ pocket were generally not conserved and not evolutionary important. However, it was reported that the residue of the P2’ position of the

peptide substrates could influence the enzymatic kinetics of PALs and AEPs, suggesting that the S2 pocket may affect the catalytic efficiency of the enzymes, but not the activity preferences [9]. The S4 and S2' pockets located far away from the catalytic residues and the oxyanion hole (the S1 pocket), which accommodates the thio-enzyme intermediate, and may play minor roles in the enzymatic activity and directionality. In contrast, the S1 pocket and the S3 pocket are highly conserved and evolutionarily important. The S3 pocket contains two Cys residues that form a disulfide bridge that may confer stability of the enzyme. Mutating these two cysteines in AtLEGγ led to the loss of protein expression completely [130], confirming the importance of the two Cys residues.

The combination of Gly and Gly-Pro at LAD1 (S2 pocket) and LAD2 (S1' pocket) respectively, are found in 77% (1213/1570) of the AEP sequences from 259 plant families analyzed in this study. These sequences are predicted to be butelase-2-like proteases, indicating the predominant presence of proteases in the plants. Many studies have reported the pivotal role that AEPs play in the maturation and activation of seed storage protein and thus germination [18, 29-31, 198, 199]. Previous reports support our results that indicated the ubiquitous presence of protease-AEPs in various plant species.

Most predicted PALs were found in eudicots, which is in agreement with the previous report of the discovery of cyclic peptides in eudicots [38]. Cyclotide-producing families are the largest families in the group of predicted butelase-1-like ligases and dual-functional ligases (**Figure 51**). This result suggests that more PALs exist in cyclic peptide-producing plants, supporting the critical role of PALs in the biosynthesis of cyclic peptides [200]. Many sequences predicted to be ligases were found in various plant families with no cyclic peptides reported

hitherto, such as the Amaranthaceae and the Brassicaceae family. This result indicates that novel ligases and cyclic peptides might be found in these plant families. For example, isoforms of the Brassicaceae family have been reported to be partial ligases [128, 130].

Chapter 5 Expression, Purification, and Characterization of the Novel Peptide Asparaginyl Ligases

5.1 Introduction

Proteases, the enzymes which cleave peptide bonds, are ubiquitous and well-characterized [1]. In contrast, the occurrences of naturally occurring peptide ligases, the enzymes which catalyze the reverse reactions of proteases to form peptide bonds, are rare [21, 56, 68, 109]. Stand-alone peptide ligase is a useful and versatile biochemical tool because it does not require an ATP or other cofactor [201]. These ATP-independent ligases can be found in plants [5-10], cyanobacteria [13], and fungi [11]. In general, the naturally occurring ligases were discovered because of their ability to act as cyclases to process the mature domains in the biosynthesis of cyclic peptides [3]. Plant peptide asparaginyl ligases (PALs), the asparaginyl endopeptidase (AEPs) showing predominant ligase activity, are also known for splicing and post-translational modifications of proteins, such as concanavalin A in the legume family [96, 97, 102-104].

PALs constitute a promising group of stand-alone, ATP-independent, and Asx-specific ligases. They display broad substrate specificity, short recognition motif in the form of a tripeptide, fast kinetics, and catalyze traceless ligation at near-neutral pH and physiological conditions [21]. Consequently, they have been utilized in various applications, including macrocyclization of proteins and peptides [150-153], live-cell labeling [156, 157], protein modifications [149, 154, 158], synthesis of peptides and proteins with unusual architectures [126, 159, 160, 175]. In addition, PALs are compatible with the conventional chemical and enzymatic ligation methods. They have been applied in bioorthogonal or

chemoenzymatic ligation in tandem or under one-pot conditions for site-specific modification as well as total and semi-synthesis of proteins [136, 158, 160].

In this chapter, the LADs-guided prediction of PALs in the dataset containing 1570 sequences is described. The selected sequences were expressed, purified, activated, and characterized to validate the predictions. Five novel butelase-1-like PALs, four partial ligases, and one protease from eight plant species and four plant families were reported and characterized. Among them, VuPAL1 and BmAEP1 were engineered based on the LADs to enhance the catalytic efficiency and ligase activity, respectively. It is demonstrated in this chapter that LADs are useful sequence motifs for the identification of butelase-1-like PALs from a pool of AEPs and modulation of the enzymatic activities of PALs and AEPs.

5.2 Result

5.2.1 Ligase Activity Determinants (LADs)-Guided Selection of PALs and AEPs

Ten sequences from eight plant species of five different plant families in the dataset were selected as representative examples to fish out PALs from a sea of AEPs. Examples of all four types of classification based on LADs were selected. They include VoPAL1 (LAD+/+), VuPAL1 (LAD+/+), VvPAL1 (LAD+/+), VyPAL4 (LAD+/+), and VyPAL5 (LAD+/+) from the Violaceae family, PePAL1 (LAD-/+), PePAL2 (LAD+/-), and PiPAL1 (LAD-/+), from the Rubiaceae family, SiPAL1 (LAD-/+) from the Pedaliaceae family, and BmAEP1 (LAD-/-) from Cucurbitaceae family (**Table 9**).

Table 9. Selected butelase-1-like PALs, partial ligases, and butelase-2-like proteases in this study.

Enzyme	Family	Species	LAD1	LAD2
Predicted butelase-1-like PAL				
VoPAL1	Violaceae	*Viola orientalis*	WVT	AP
VyPAL4	Violaceae	*Viola yedoensis*	LIA	AP
VyPAL5	Violaceae	*Viola yedoensis*	WIT	AP
VuPAL1	Violaceae	*Viola uliginosa*	WIT	AP
VvPAL1	Violaceae	*Viola orientalis*	WIT	AP
Predicted partial ligase				
PePAL1	Solanaceae	*Petunia exserta*	WAT	AP
PePAL2	Solanaceae	*Petunia exserta*	WVT	GP
PiPAL1	Rubiaceae	*Psychotria ipecacuanha*	WAA	AP
SiPAL1	Pedaliaceae	*Sesamum indicum*	YAI	AA
Predicted butelase-2-like protease				
BmAEP1	Cucurbitacaea	*Momordica charantia*	WAT	SA

Those belong to the group of LAD+/+, such as the selected example VoPAL1, VuPAL1, VvPAL1, and VyPAL4-5, were predicted to be butelase-1-like ligases with predominant ligase activity and negligible hydrolysis activity. Those from the groups of LAD+/- (PePAL2) and LAD-/+ (PePAL1, PiPAL1, and SiPAL1) were predicted to be dual-functional ligases, or partial ligases, that exhibit both ligase and protease activity. BmAEP1 (LAD-/-) was predicted to be a butelase-2-like protease that preferably cleaves than ligates. It was selected as an example of protease of the group of LAD-/-.

A sequence-based phylogenetic tree containing 57 sequences, including the selected ten putative PALs and AEPs and previously reported PALs and AEPs, was generated and visualized using Simple Phylogeny [181] and Interactive Tree of Life [182, 183], respectively (**Figure 52**). Two color strips were generated on the right of the phylogenetic tree to indicate the LAD compositions and families of the sequences. There are three major branches shown in the phylogenetic tree, two branches containing mostly predicted and reported AEPs, and one branch of mostly predicted and reported PALs, including butelase-1 [6], VyPAL2 [10], and OaAEP1b [7]. All seven sequences of the Rubiaceae family cluster together, including the known PALs OaAEP1b and predicted PAL PiPAL1 as well as the known protease OaAEP2. Similarly, six out of seven sequences (except PePAL1) of the Solanaceae family cluster together on the phylogenetic tree.

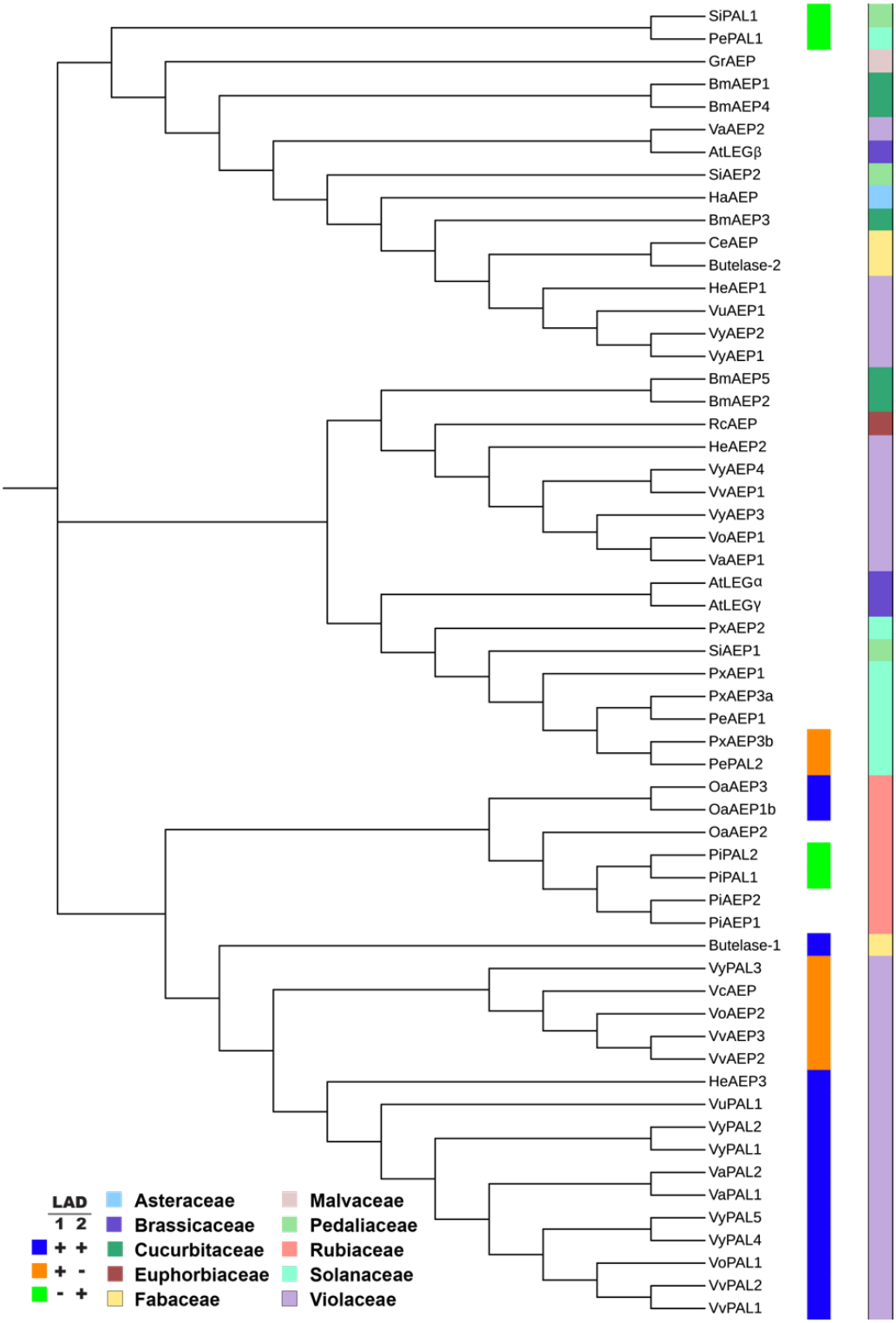


Figure 52. Sequence-based phylogenetic tree of 57 known and putative PALs and AEPs. The color strips on the right indicated the amino acid compositions at the LAD2 and the plant family the sequences belong to. Sequences of the same species do not necessarily cluster together, both PALs and AEPs cluster with other PALs and AEPs with similar LADs, respectively. For example, butelase-1 and butelase-2 locate far away on the phylogenetic tree although they are both

from the plant *Clitoria ternatea*. Similarly, sequences from *Viola yedonesis* form four clusters on the phylogentic tree, and all the PALs locate in close proximity together with other predicted PALs from the Violaceae family in the same cluster. The sequence-based phylogeny tree was generated by Simple Phylogeny using the neighbor-joining method with default setting [181]. The phylogenetic tree is visualized using iTOL Interactive Tree of Life [182, 183].

However, the sequences of the same species and family do not necessarily cluster together in the phylogenetic tree. Frequently, PALs and AEPs cluster with other PALs and AEPs with similar LAD compositions. For example, butelase-1 is located far away from butelase-2, although they are from the same species, *Clitoria ternatea*. Similarly, sequences of the *Viola yedoensis* form four distant clusters on the tree. VyPAL1-2 (known PALs) and VyPAL4-5 (predicted PALs) cluster with the known PAL HeAEP3 [8] and the predicted PALs of the Violaceae family, including VuPAL1, VaPAL1-2, VoPAL1, and VvPAL1-2. While VyPAL3 (LAD-/+), a predicted partial ligase, is located slightly away from other PALs from the Violaceae family and cluster with VcAEP (LAD+/-) and other sequences which are also classified as 'LAD+/-,' including VoAEP2 and VvAEP2-3. Notably, the sequence identities of the core domains of VyPAL1-2 and VyPAL4-5 are higher (95.0%-99.7%) compared to VyPAL3 (69.8%-69.9%) (**Table 10**). Similarly, the previously reported protease HeAEP1 [8] is located far away from its ligase isoform HeAEP3 and clusters with another previously reported protease VyAEP1 [10] and other predicted protease. It is important to note that classifications of PALs and AEPs were based on only three residues at the LADs, but the phylogenetic tree is generated based on multiple sequence alignments of more than 400 residues. This pattern is also observed in the phylogenetic tree comprised of 1570 sequences (**Figure 38**).

Table 10. Sequence homology chart of isoforms in *Viola yedoensis*. Sequence similarity (%) of the core domains of the isoforms was calculated by EMBL-EBI EMBOSS Water Pairwise Sequence Alignment (available online at https://www.ebi.ac.uk/Tools/psa/emboss_water/) with default setting.

Enzyme	VyAEP1	VyAEP2	VyAEP3	VyAEP4	VyPAL1	VyPAL2	VyPAL3	VyPAL4	VyPAL5
VyAEP1	100.0								
VyAEP2	98.9	100.0							
VyAEP3	66.1	65.7	100.0						
VyAEP4	66.1	65.7	96.2	100.0					
VyPAL1	57.2	57.6	69.5	68.1	100.0				
VyPAL2	57.2	57.6	70.1	69.0	95.4	100.0			
VyPAL3	57.6	57.6	70.2	69.9	69.9	69.8	100.0		
VyPAL4	56.8	57.2	69.1	67.7	99.7	95.0	69.9	100.0	
VyPAL5	56.8	57.2	69.8	68.7	95.0	99.6	69.8	95.4	100.0

5.2.2 Expression, Purification, and Activation of Recombinant PALs and AEPs

The DNA sequences of the ten selected putative PALs and AEPs were retrieved from the NCBI nucleotide collection (nr/nt) database, NCBI transcriptome shotgun assembly (TSA) database, and OneKP database. They were designed as 6His-ubiquitin-enzyme fusion protein constructs. All constructs do not contain the signal peptide, which is replaced by a hexahistidine tag followed by ubiquitin (**Figure 53**).

The presence and location of the cleavage sites of signal peptides of the ten selected sequences were predicted by SignalP 5.0 [178]. SignalP 5.0 can detect three types of signal peptides. They include signal peptides translocated through the secretory (Sec) pathway by the Sec translocase in an unfolded state and removed by Signal Peptidase I (Sec/SPI), (2) lipoprotein signal peptides transported by secretory (Sec) pathway in an unfolded state by the Sec translocase and removed by Signal Peptidase II (Sec/SPII), and (3) Tat signal peptides, which is subjected to Tat translocation and Signal Peptidase I cleavage (Tat/SPI). Signal peptides of all ten selected sequences from five plant families were all predicted to be transported by the Sec pathway and removed by Signal Peptidase I (Sec/SPI) (**Figure 54**).

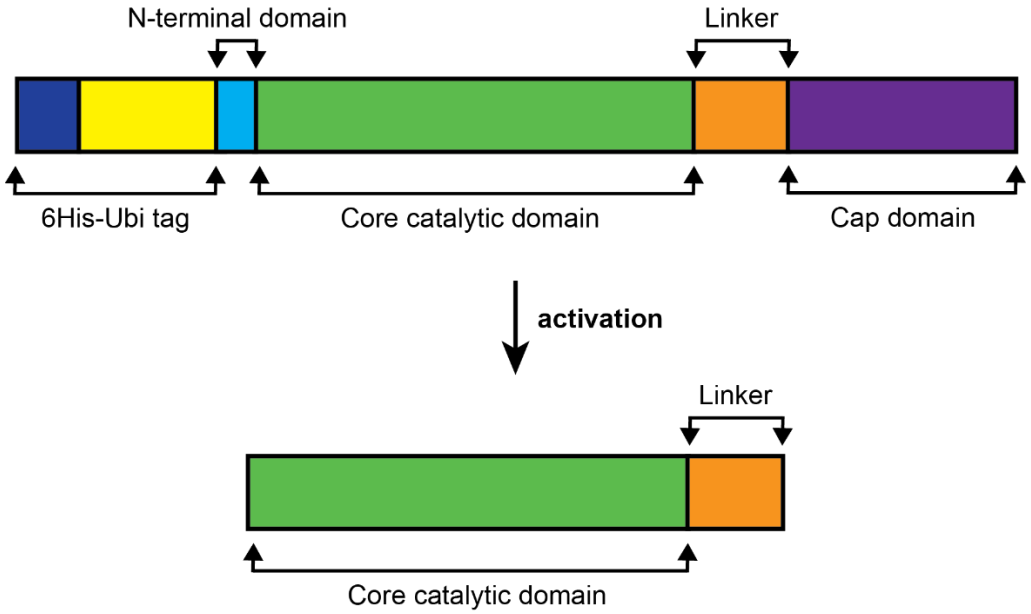


Figure 53. Construct design, purification, and activation of selected PALs and AEPs. Schematic representation of the His-Ub-enzyme construct used in this study and the activated catalytic enzyme containing core catalytic domain and part of the linker.

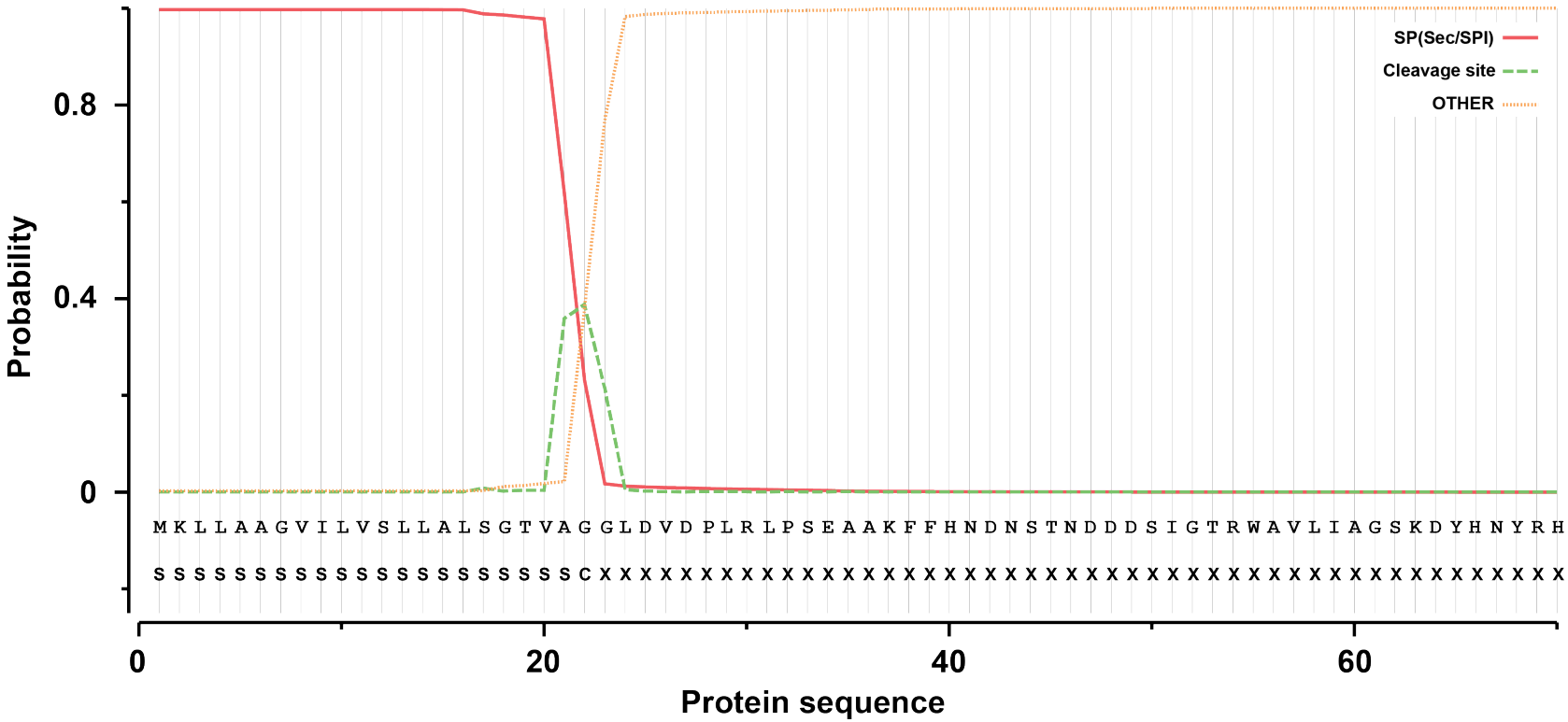


Figure 54. Signal peptide prediction using the output of VuPAL1 as an example. The amino acid sequences are located on top of the symbols indicating the identity of each residue. The presence of signal peptides were predicted by SignalP 5.0 [178]. S indicates a residue that is part of the signal peptide, C indicates the cleavage site, and X indicates other amino acids. The red line indicates that the signal peptide of VuPAL1 is predicted to be transported and removed by the secretory pathway and Signal Peptidase I, respectively. The green dashed line indicates the probability and location of the cleavage site of the signal peptide. The yellow dashed line indicates the presence of residues that are not predicted to be part of the signal peptide.

The sequences were cloned into a pET28(a)+ plasmid and expressed as zymogenic precursors in *Escherichia coli* T7 SHuffle (New England Biolabs, USA). To facilitate correct disulfide shuffling, *E. coli* also contains a pMJS9 plasmid that encodes a disulfide isomerase Erv1p gene and a protein disulfide isomerase (PDI) gene (a kind gift from Professor Lloyd Ruddock, University of Oulu, Finland). Initial attempts to express VoPAL1 and VvPAL1 resulted in low yield with no detectable soluble protein on the SDS-PAGE and no detectable activity by MALDI-TOF MS. The short Asp-containing region, Gly23-Pro28 of VoPAL1 and Gly22-Pro28 of VvPAL1, at the beginning of the N-terminal domain, which may facilitate the autoactivation of enzymes without acidic activation during the expression and purification process, was thus deleted. The yields of both VoPAL1 and VvPAL1 increased after truncation of the N-terminal short Asp-containing domain. The mutagenesis was performed using the Q5 mutagenesis kit (New England Biolabs, USA) through polymerase chain reaction (PCR).

The expressed His-Ub-zymogens of PALs and AEPs were purified by immobilized metal affinity chromatography (IMAC), anion exchange chromatography, and size exclusion chromatography using fast protein liquid chromatography (FPLC) (**Figure 55**). Anion exchange and size exclusion chromatography (SEC) were optional as long as the enzymes were shown to be purified on the SDS-PAGE after IMAC. The His-Ub-zymogens of BmAEP1, PePAL1, PiPAL1, SiPAL1, VoPAL1, VuPAL1, VvPAL1, and VyPAL4-5 were obtained and visualized by SDS-PAGE and Coomassie staining (**Figure 56**). Approximately 0.5 mg to 1 mg of His-Ub-zymogens was obtained. The sizes of the His-Ub-zymogens range from 50 kDa to 55 kDa. For PiPAL1 and BmAEP21,

the His-Ub-zymogens observed were not homogenous, indicating that the N-terminal domain or C-terminal cap domain may be partly cleaved and thus resulted in His-Ub-zymogens of the same enzyme of different sizes. Immature processing of the C-terminal cap domain was also observed in human legumain [122].

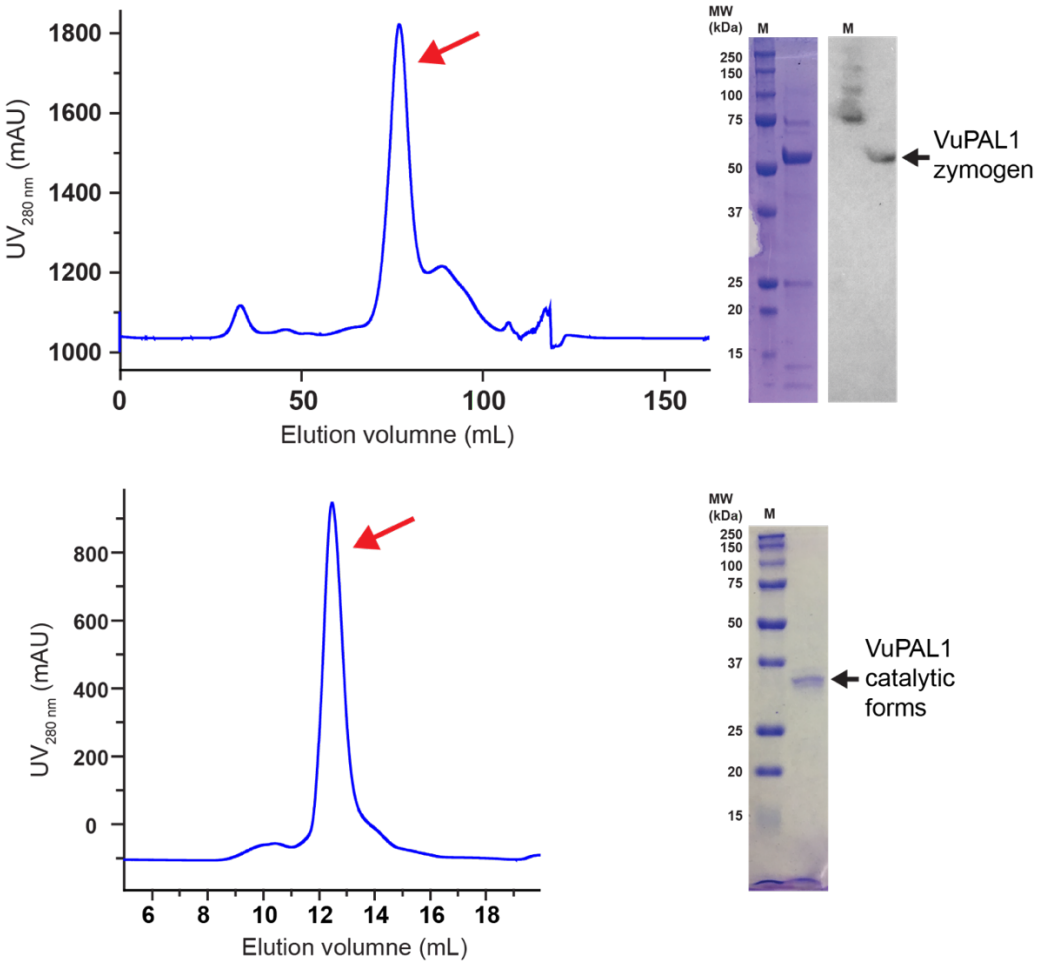


Figure 55. The FPLC profiles of purification of VuPAL1. The profiles of size exclusion chromatography (SEC) of the VuPAL1 His-Ub-zymogen and catalytic forms are on the top and the bottom, respectively. The proteins were visualized using SDS-PAGE and Coomassie staining and Western blot (shown on the right). HiLoad Superdex 75 prep grade column (GE healthcare, USA) was used for purification of the His-Ub-zymogen. HiLoad Superdex 200 pg preparative size exclusion chromatography columns (GE healthcare, USA) was used for the purification of the catalytic forms. Both columns were coupled to an AKTA system (GE Healthcare, USA), respectively. The red arrows indicate the target peaks containing VuPAL1. M stands for protein marker.

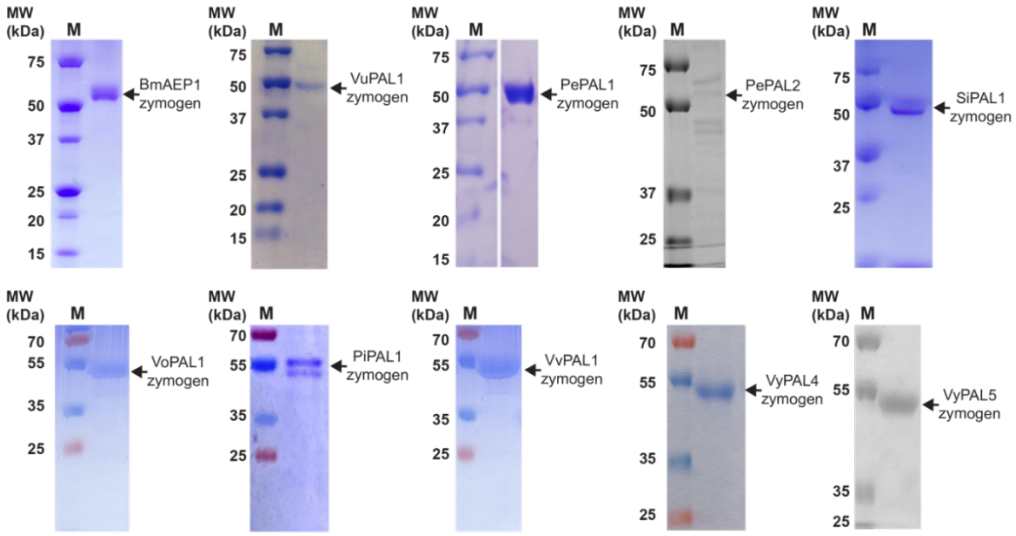


Figure 56. Visualization of the His-Ub-zymogens of putative PALs and AEPs. They included BmAEP1, PePAL1-2, PiPAL1, SiPAL1, VoPAL1, VuPAL1, VvPAL1, and VyPAL4-5. The proteins were visualized by SDS-PAGE and Coomassie staining.

5.2.3 Determination of C-terminal Autolytic Cleavage Site of PALs

To mimic the pH change during the maturation of PALs and AEPs *in vivo*, the purified His-Ub-zymogens were incubated at pH 4.5 for 10 min to 2 h at 37 °C, with the addition of 0.5 mM N-lauroylsarcosine, 1 mM EDTA, and 5 mM β-ME. Bands of about 31 kDa to 37 kDa were visualized using SDS-PAGE and Coomassie staining. A decrease in the size of the bands indicates that the catalytic core domain was separated from the N-terminal domain and C-terminal cap domain, exposing the catalytic sites and the substrate-binding pockets. The yields of the activated enzymes were within the range of 0.1 mg to 0.5 mg. Some of the purified and activated enzymes showed multiple bands, such as VuPAL1, PePAL1, VyPAL5, and BmAEP1 (**Figure 55 and Figure 57**), suggesting that there were multiple cleavage sites at the terminus through enzymatic activation, which is commonly reported in AEPs [122, 124, 134].

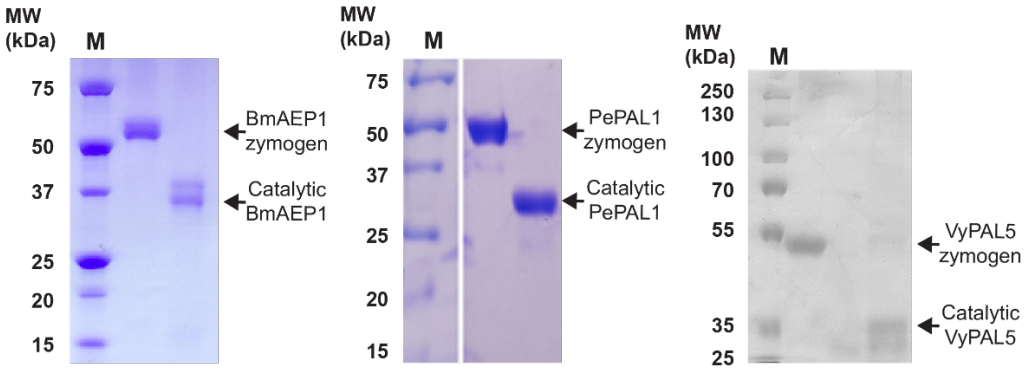


Figure 57. BmAEP1, PePAL1, and VyPAL5 showed several catalytic forms. The proteins were visualized by the SDS-PAGE and Coomassie staining.

To determine the processing sites, the gels of activated enzymes were subjected to tryptic digestion and *de novo* sequencing using LC-MS/MS. Peptide fragments with canonical cleavage sites of trypsin were excluded (**Figure 58**), and the putative C-terminal cleavage sites are summarized in **Figure 59**. At the N-terminus, the common cleavage sites were observed in the Asn/Asp-rich region, Asp39-Asn41 of butelase-1, Asn41-Asp49 of VuPAL1, Asn46-Asn48 of VyPAL5, Asn34Asm39 of PePAL1, and Asp31 and Asp36 of BmAEP1, before the catalytic core domain. At the C-terminus, multiple cleavage sites were observed at or near the linker region before the α6-helix (Asn322-Asp325 according to butelase-1 numbering), such as Asn322 and Asn327 of PePAL1, suggesting that removal of cap domain abolished the inhibitory effect caused by the presence of Gln at the tip of α6-helix. Notably, the residues after detected cleavage site Asn333 of VuPAL1 are Ser334 and Leu335, and the Asn-Ser-Leu tripeptide motif at the P1-P1'-P2' position is also the common recognition motif of the cyclotides found in *Viola uliginosa* [38].

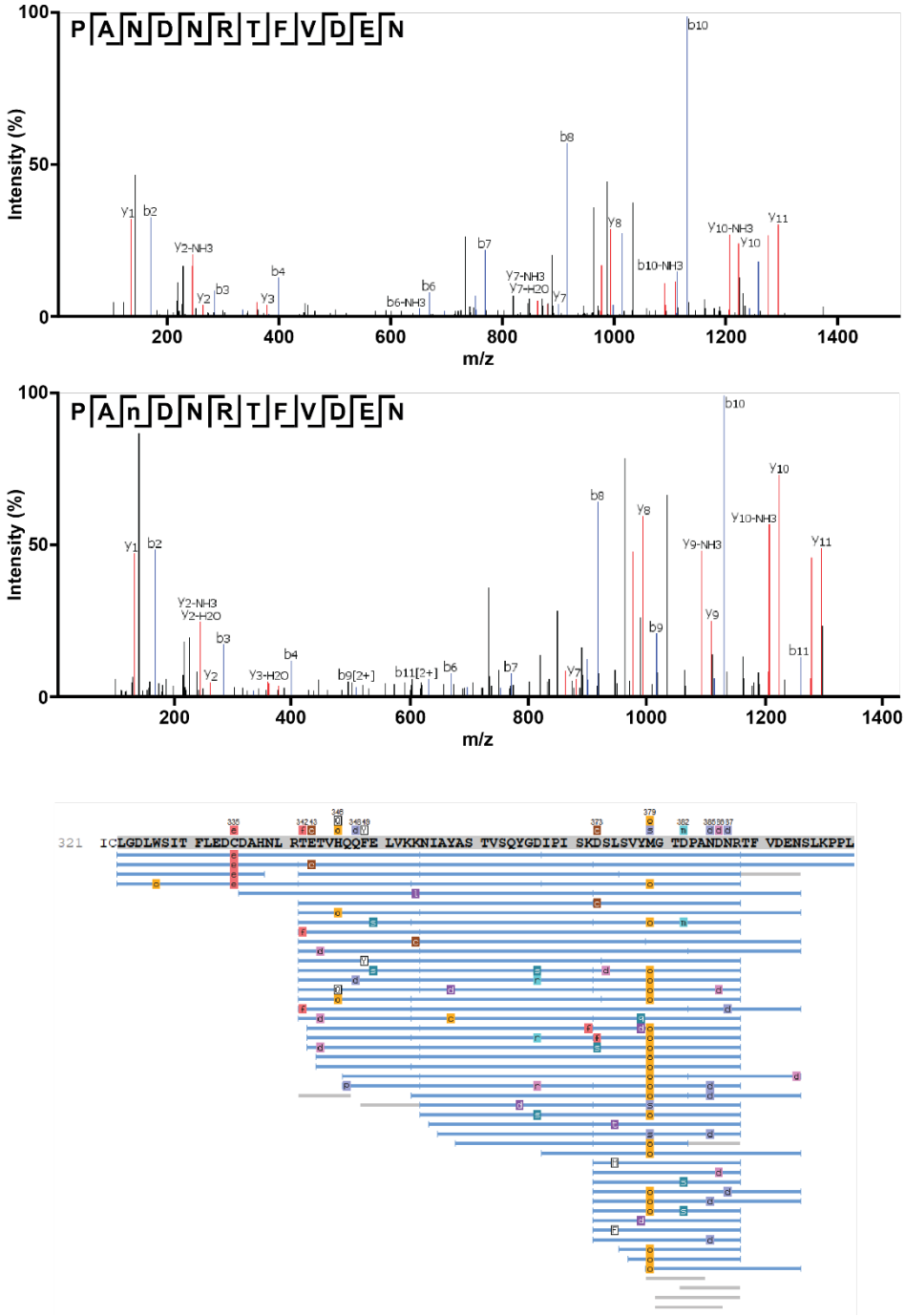


Figure 58. Peptide fragments of VuPAL1 detected through in-gel tryptic digestion and *de novo* sequencing by LC/MS-MS. The peptide fragments obtained through LC/MS-MS were visualized through PEAKS Studio 7.5 (Bioinformatics Solutions, Waterloo, ON, USA).

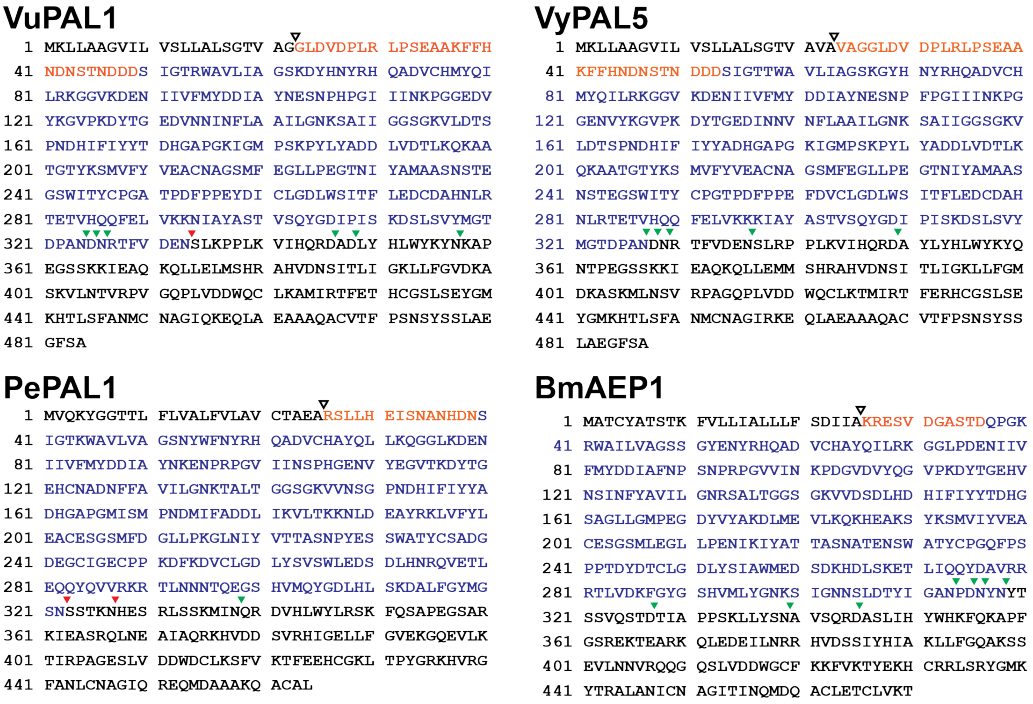


Figure 59. Cleavage sites of the selected PALs and AEPs. The signal peptide cleavage sites were predicted using SignalP 5.0 and are indicated by the white triangle. Residues belonging to the signal peptide, N-terminal domain, core catalytic domain, and C-terminal cap domain are colored in black, orange, blue, and black, respectively. The putative C-terminal cleavage sites were indicated by the green triangles and experimentally validated C-terminal cleavage sites (indicated by red triangles) were confirmed by Orbitrap MS/MS.

5.2.4 Screening of Ligase Activity of Selected PALs and AEPs

We first determined the enzymatic activities of selected examples of different classifications. They included the predicted butelase-1-like PAL, VuPAL1 (LAD+/+), the predicted partial ligases, PiPAL1, PePAL1-2, SiPAL1, and predicted protease BmAEP1. VcAEP, a previously reported protease that contains a Val and a dipeptide Tyr-Pro at the LAD1 and the LAD2, respectively [10], was also screened too as an example of partial ligase. Model peptide substrate GN14-X_{0-4} (GISTKSIPPISYRN-X_{0-4}) was modified based on the sunflower trypsin inhibitor (SFTI) precursor with the leaving groups (X_{0-4}) same as the native precursors found in the species or families of the PALs or AEPs on CyBase. For example, a C-terminal tripeptide motif Asn-Ser-Leu (P1-P1'-P2' position) was commonly found in the precursors of cyclic peptides in *Viola uliginosa* [38], model peptide GN14-NSL was thus used for the VuPAL1-mediated cyclization assay. Cyclization of the substrate GN14-X_{0-4} yielded the cyclic product cGN14 with the molecular weight of 1515, while hydrolysis of the substrate resulted in the production of the linear product GN14 with the molecular weight of 1530 (**Figure 60**). The cyclization assays for all seven enzymes were conducted at pH 6.5 at 37 °C and qualified by MALDI-TOF MS. All seven tested PALs and AEPs were shown to be able to cyclize and/or hydrolyze the model peptide substrate GN14-X_{0-4} (**Figure 61**).

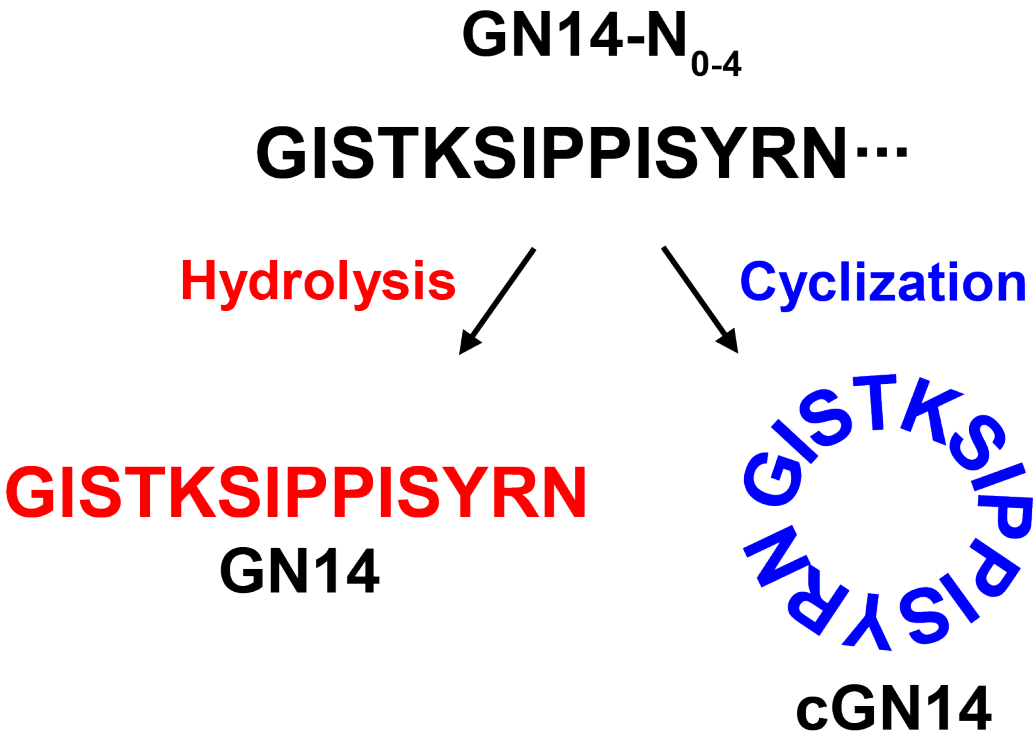


Figure 60. Schematic representation of the cyclization assay using the model peptide substrate GN14-X_{0-4}. The GN14 represents GISTKSIPPISYRN with the Asn at the P1 position, the X_{0-4} represents the C-terminal leaving groups. Cyclization of the GN14-X_{0-4} produces the cyclic peptides cGN14 (colored in blue), while hydrolysis of the GN14-X_{0-4} results in the linear products GN14 (colored in red).

VuPAL1, a predicted butelase-1-like PAL, exhibited predominant ligase activity and yielded no detectable linear product at pH 6.5 at 37 °C within 30 min. The predicted partial ligases, on the other hand, showed different levels of ligase and hydrolysis activity. Among the partial ligases, PiPAL1 (LAD-/+) showed the highest ligase activity with almost negligible linear products detected by MALDI-TOF MS. Another predicted partial ligase classified as 'LAD-/+', PePAL1, was shown to slightly prefer to ligate than to cleave, which produced slightly more cyclic than linear products. The other partial ligase of the group of 'LAD-/+' was SiPAL1, which preferred to cleave than to cyclize, producing mostly linear products. The cyclic-to-linear product ratio of PePAL2 (LAD+/-), an isoform of PePAL1, is similar to VcAEP (LAD+/-). They both produced more linear products than cyclic products, and their product distributions were similar to that of SiPAL1. Lastly, BmAEP1, a predicted to be a butelase-2-like protease and classified as 'LAD-/-,' showed predominant protease activity and the presence of cyclic product was not detectable by MALDI-TOF MS at pH 6.5 at 37 °C (**Figure 61**).

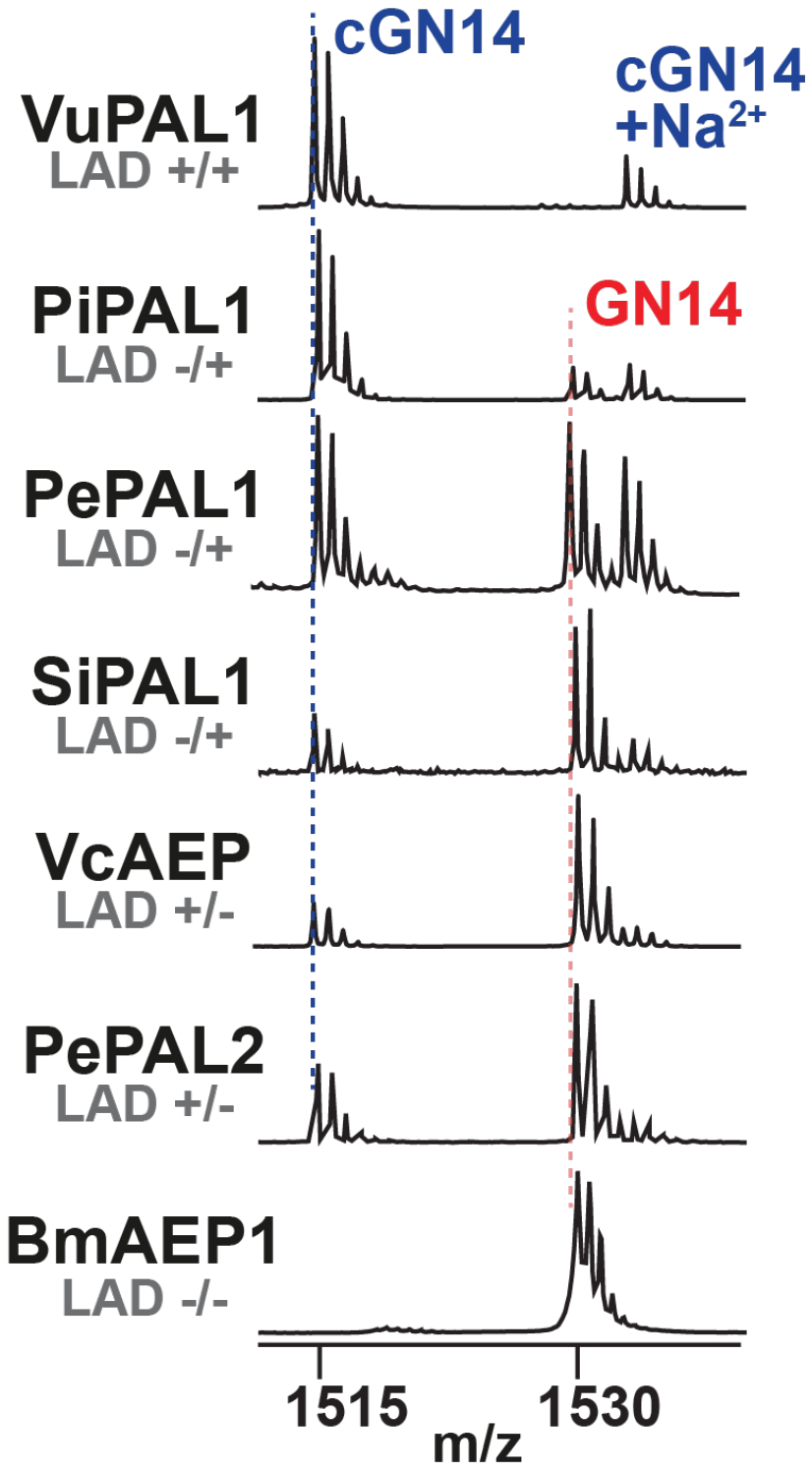


Figure 61. MALDI-TOF MS/MS spectra of enzymatic activities of selected PALs and AEPs. The 1515-Da peptides and the 1532 are the cyclized products (indicated by blue dashed line), and the 1530-Da peptides are the hydrolyzed products (indicated by the red dashed line).

PiPAL1 (LAD-/+), as an example of partial ligase, was next selected and screened for its activity from pH 4.5 to 8.0 using the model substrate GN12-GL (GLYRRGRLYRRNGL) (**Figure 62**). The reaction was performed at 37 °C for 30 min with the enzyme-to-substrate ratio of 1:1000, quantified and qualified by Reverse-Phase High-Performance Liquid Chromatography (RP-HPLC), and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), respectively (**Figure 63**). Butelase-2 was also screened at the same condition using the same substrate, GN12-GL, as a negative for comparison (**Figure 64**). The disparity of enzymatic activity of PiPAL1 and butelase-2, a prototypic protease, was clearly demonstrated. PiPAL1 behaved predominantly as a ligase and showed no detectable linear product (GN12) from pH 6.5 to 8.0. The cyclization and hydrolysis activity of PiPAL1 peaked at pH 6.0 and pH 5.0, respectively. In contrast, the butelase-2-mediated GN12-GL processing resulted in linear GN12 as a major product. The hydrolysis activity of butelase-2 dropped largely from pH 7.0 to 8.0 compared to pH 4.5 to 6.5.

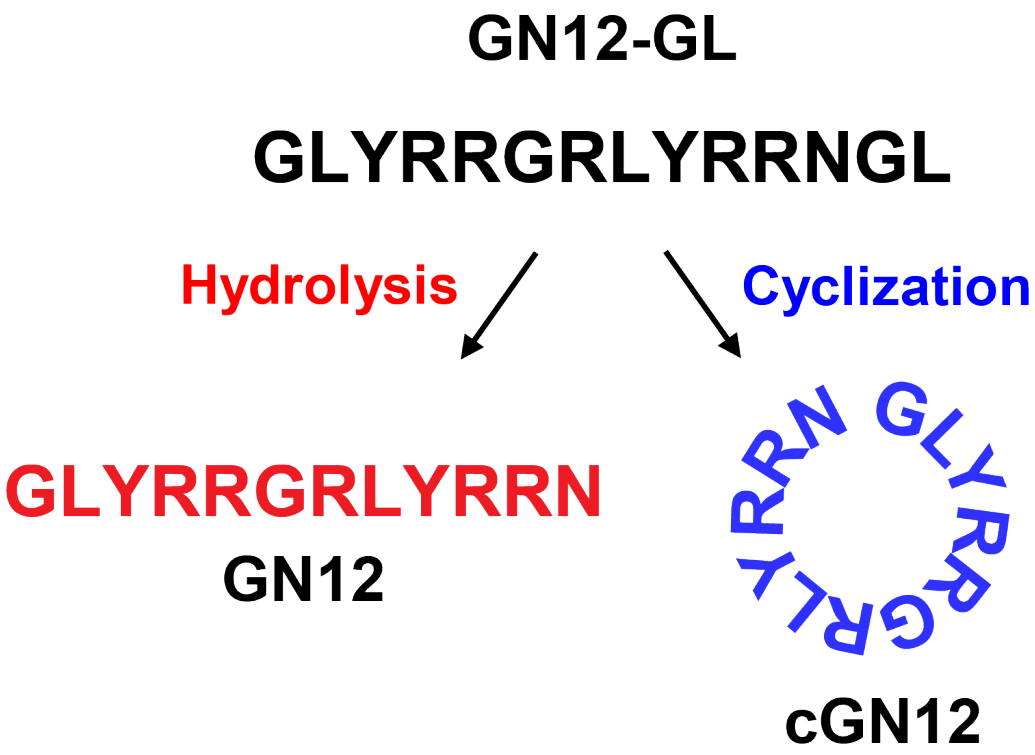


Figure 62. Schematic representation of the cyclization assay using the model peptide substrate GN12-GL. The full sequence of GN12-GL is GLYRRGRLYRRNGL. Cyclization of the GN12-GL produces the cyclic peptides cGN12 (colored in blue), while hydrolysis of the GN12-GL results in the linear products GN12 (colored in red).

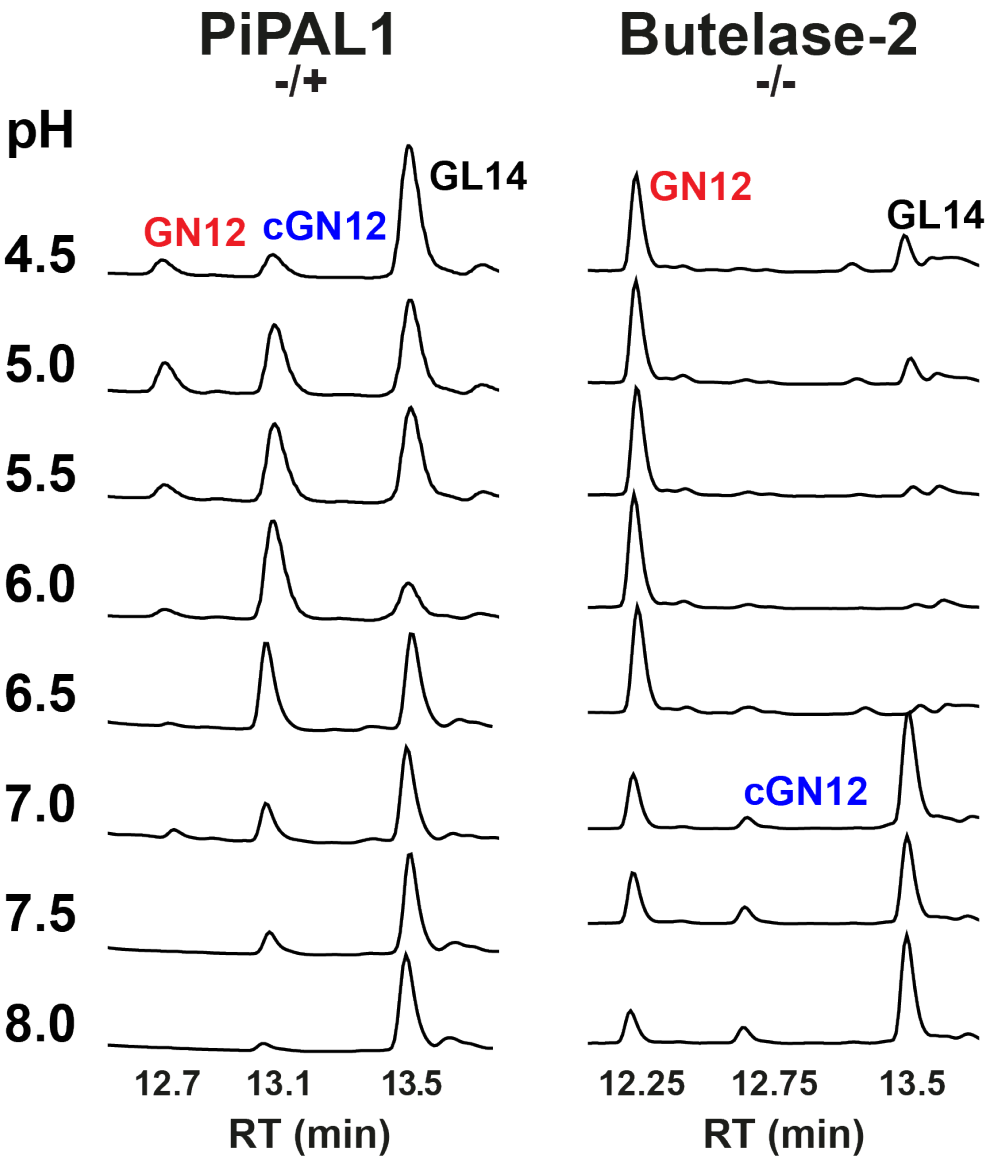


Figure 63. The RP-HPLC profiles of cyclization and hydrolysis of the peptide substrate GN12-GL by PiPAL1 and butelase-2. RT stands for retention time.

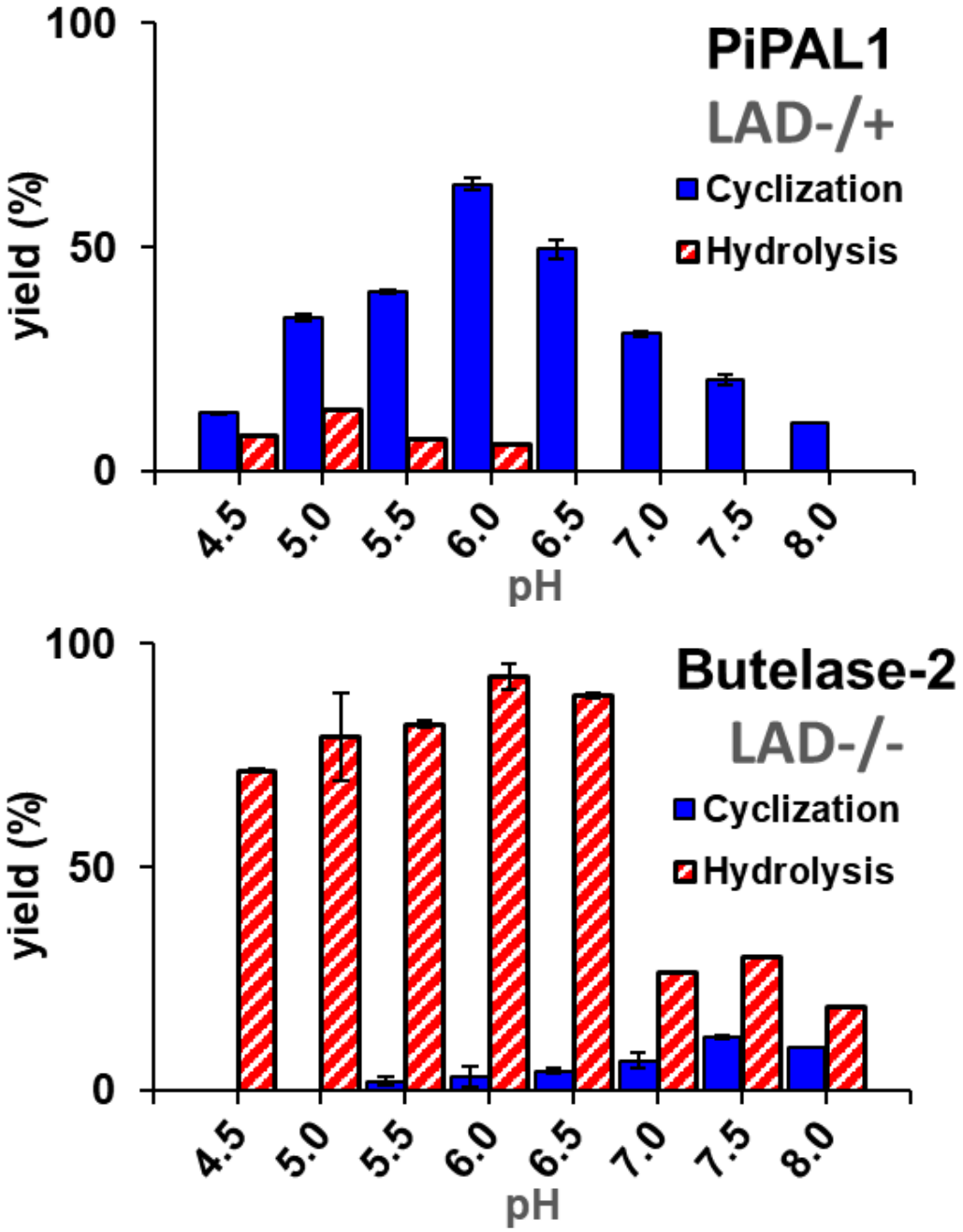


Figure 64. Characterization of PiPAL1 and butelase-2 using the model substrate GN12-GL. The reactions were performed at 37 °C from pH 4.5 to 8.0.

5.2.5 Predicting Putative PALs from the Dataset of 1570 Sequences

Having established that the peptide substrate GN12-GL was able to be processed by both partial ligase and protease, the activity of five predicted PALs, VoPAL1 (LAD+/+), VuPAL1 (LAD+/+), VyPAL4 (LAD+/+), VyPAL5 (LAD+/+), and VvPAL1 (LAD+/+), was next screened using the GN12-GL at 37 °C from pH 4.0 to 8.0. The reactions were quantified and qualified by RP-HPLC and MADLTI-TOF MS (**Figure 65**).

All five predicted putative butelase-1-like PALs showed predominant ligase activity with no detectable hydrolysis product from pH 4.0 to 8.0 using the model peptide substrate GN12-GL (**Figure 66**). The optimal cyclization efficiency of VuPAL1 was found to be from pH 5.5 to pH 6.5 with more than 85% of the substrates converted to cyclic products cGN12 within 1 h with the enzyme-to-substrate ratio of 1:500. The activity of VuPAL1 dropped largely when pH was equal to or higher than 7.5, or equal to and lower than 4.5. The lowest activity of VuPAL1 was found at pH 4.0, and the cyclic products cGN12 converted at pH 4.0 can only be detected by MADLTI-TOF and not RP-HPLC (**Figure 66**).

For VyPAL4, the cyclization yield of cyclized products cGN12 exhibited a bell-shaped curve pattern and reached a maximum at pH 6.0 37 °C, with 71.3% starting materials cyclized, using the model peptide GN12-GL. The catalytic efficiency of VyPAL4 decreased gradually from pH 6.0, and the lowest activity was observed at ph 4.0 with only about 5% GN12-GL converted to cyclic products (**Figure 66**).

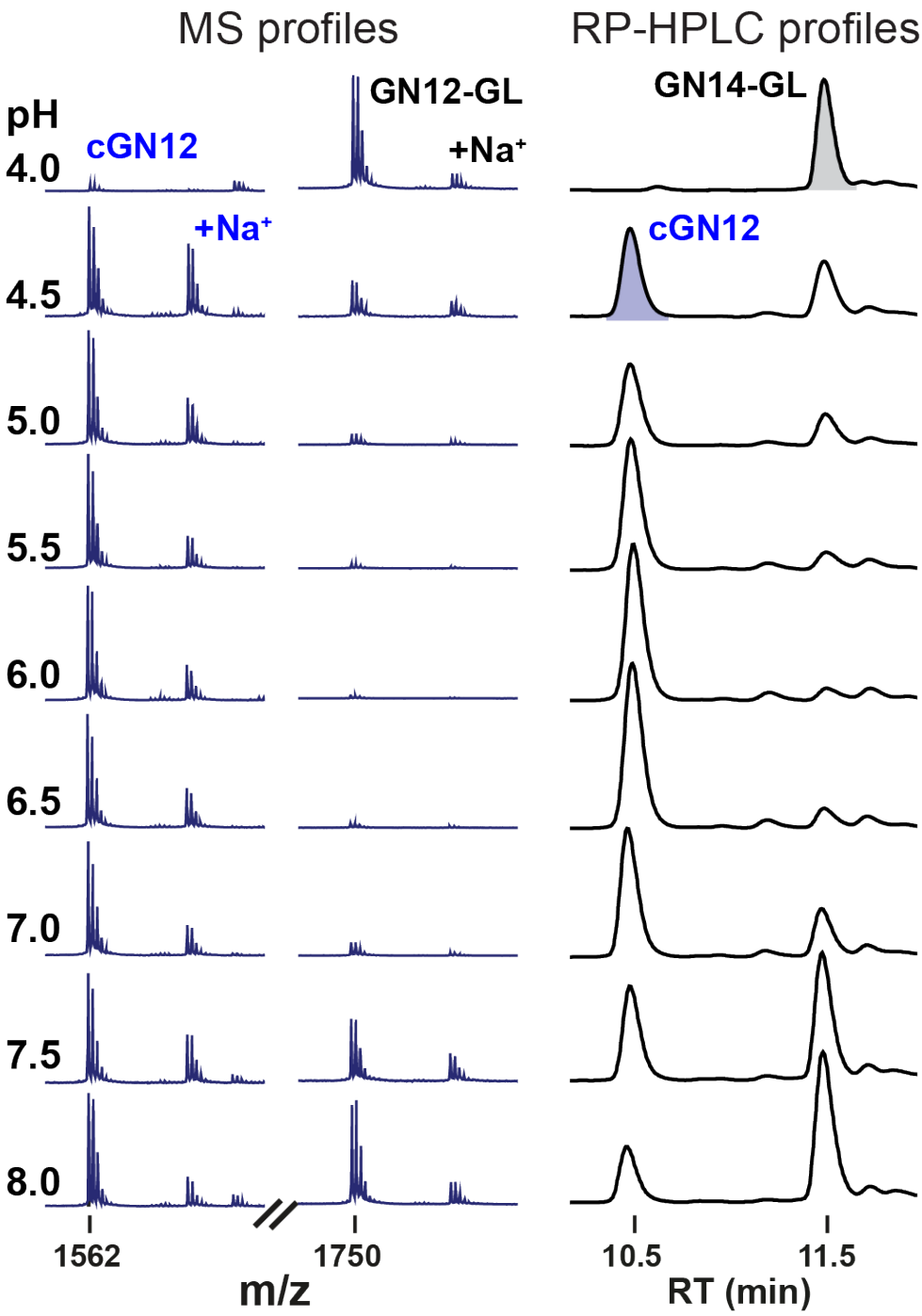


Figure 65. The MALDI-TOF MS spectra and RP-HPLC profiles of VuPAL1-mediated cyclization. The reactions were performed using the peptide substrate GN12-GL at 37 °C from pH 4.0 to 8.0. RT stands for retention time.

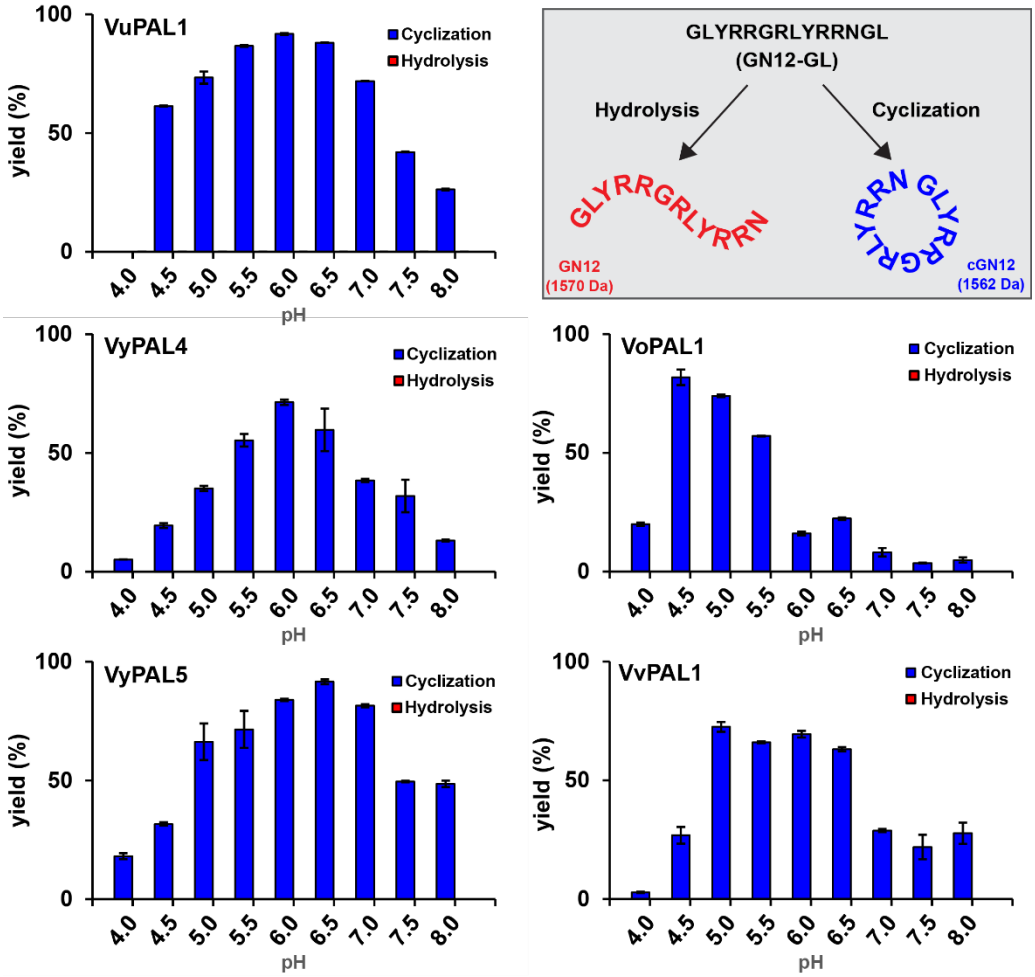


Figure 66. Characterization of selected PALs from pH 4.0 to 8.0. The reactions of VuPAL1, VoPAL1, VyPAL4-5, and VvPAL1 were performed using the peptide substrate GN12-GL at 37 °C from pH 4.5 to 8.0. The reaction scheme was shown on the top right corner of the figure.

The cyclization efficiency of VyPAL5 peaked from pH 6.0 to pH 7.0, and dropped from 81.4% to 49.5% when pH was increased from 7.0 to pH 7.5. Same as VuPAL1 and VyPAL4, VyPAL5 also had its lowest activity observed at pH 4.0. The yield of cyclic product cGN12 was found to be lowest at pH 4.0 with only 18.1% of the substrate GN12-GL converted to cGN12 (**Figure 66**).

The enzymatic activity profile of VvPAL1 was similar to VuPAL1, VvPAL1-mediated cyclization efficiency was highest from pH 5.0 to 6.5, and the yields of cyclic products cGN12 decreased significantly when the reactions were not performed within the range of pH 5.0 to 6.5, with the lowest yield of cyclic products (less than 5%) observed at pH 4.0 (**Figure 66**).

VoPAL1, different from the other four PALs, had its cyclization efficiency peaked at pH 4.5, with more than 81.7% of starting materials cyclized. VoPAL1-mediated cyclization was less efficient at basic pH. The yields of cyclic products cGN12 were lower than 10% from pH 7.0 to 8.0 (**Figure 66**). This pattern, that PALs cyclize with higher efficiency at acidic pH, was also observed in OaAEP3-5 [9].

5.2.6 Substrate Specificity Screening of VuPAL1

VuPAL1 was selected for substrate-specificity screening. The minimum requirement of the substrate recognition of VuPAL1 was first examined. An Asx residue at the P1 position is the essential requirement of substrate recognition of PALs and AEPs, whereas the following dipeptide leaving group varies. For example, butelase-1 prefers an Asn-His-Val tripeptide motif at the P1-P1'-P2' position [6], while OaAEP1b favors an Asn-Gly-Leu tripeptide motif at the C-terminus [7]. It had been established that VuPAL1 is capable of rendering ligation reaction from pH 4.0 to 8.0 with peptide substrates containing a tripeptide motif Asn-Ser-Leu at the C-terminus, the effect of the length of the C-terminal leaving group on the ligation efficiency and product distribution of VuPAL1-mediated cyclization was thus tested.

Several cyclic peptides were found in the plant *Viola uliginosa*. They include kalata S, cycloviolacin 13 (CyO 13), cycloviolacin 2 (CyO 2), and cycloviolacin 8 (CyO 8) [38, 202]. One feature these cyclic peptides all have in common is the C-terminal Asn-Sere-Leu tripeptide motif at the P1-P1'-P2' position (**Table 11**). The peptide substrates modified from the model peptide GN14-X_{0-4} with degenerated tails derived from the native cyclic peptide precursors in *Viola uliginosa* were synthesized and tested (**Table 12**, no. 1-7).

VuPAL1 was unable to cyclize or cleave the peptide substrates carrying an Asn-Ser dipeptide motif at the P1-P1' position (**Table 12**, no. 4) and a P1-Asn without leaving group (**Table 12**, no. 5).

Table 11. Selected native cyclic peptide precursors identified in *Viola uliginosa*. The sequences were retrieved from CyBase [38].

Cyclic peptide	Sequence
kalata S	GLPVC-GETCVGGTC---NTPGCSCSW-PVCTR**N**SLAM*
CyO 13	G-IPC-GESCVWIPC-ISAAIGCSCKS-KVCYR**N**SLDN*
CyO 2	G-IPC-GESCVWIPC-ISSAIGCSCKS-KVCYR**N**SLDN*
CyO 8	GTLPC-GESCVWI-C-ISSVVGCSCKS-KVCYK**N**SLA*

The P1-Asn is colored in red and bold. The leaving groups are underlined. The asterisk indicates end of the sequence.

The results showed that the length of the residue downstream of the Asn at the P1 position did not influence the efficiency and substrate distribution of VuPAL1-mediated cyclization (**Figure 67**), suggesting that a C-terminal tripeptide is sufficient for efficient backbone cyclization by VuPAL1.

To test if P1-Asp can be accepted by VuPAL1 at the P1 position, a 16-residue peptide substrate (GISTKSIPPISYRDSL) with the tripeptide motif Asp-Ser-Leu at the P1-P1'-P2' position, instead of Asn-Ser-Leu, was synthesized (**Table 11**, no. 8). No detectable cyclic or linear product was observed (**Figure 67**), suggesting that similar to VyPAL2 [10], VuPAL1 is unable to efficiently catalyze the cyclization of peptide substrates with an Asp at the P1 position.

To screen the substrate preference of VuPAL1 at the P1' and P2' position of the peptide substrates, two peptide libraries with C-terminal Asn-Xaa-Leu and Asn-Gly-Xaa (Xaa represents all 20 native amino acids) at the P1-P1'-P2' position were tested (**Table 11**, no 9-10). Similar to VyPAL2, VuPAL1 accepted almost all natural amino acids at the P1' position except Pro (**Figure 68**). At the P2' position, VuPAL1 preferred hydrophobic and aromatic residues, such as Leu, Ile, and Phe. These residues are also favored by VyPAL2 at the P2' position [10] (**Figure 68**).

Table 12. Peptide substrates used in this thesis for substrate preference screening.

No.	Peptide Substrate
1	GISTKSIPPISYR**N**SLAM*
2	GISTKSIPPISYR**N**SLA*
3	GISTKSIPPISYR**N**SL*
4	GISTKSIPPISYR**N**S*
5	GISTKSIPPISYR**N***
6	GISTKSIPPISYR**N**SLDN*
7	GISTKSIPPISYR**N**SLD*
8	GISTKSIPPISYR**D**SL*
9	GLYRRGRLYRR**N**XL* (X stands for all 20 native amino acids)
10	GLYRRGRLYRR**N**GX* (X stands for all 20 native amino acids)

The P1-Asx is colored in red and bold. The leaving groups are underlined. The asterisk indicates end of the sequence.

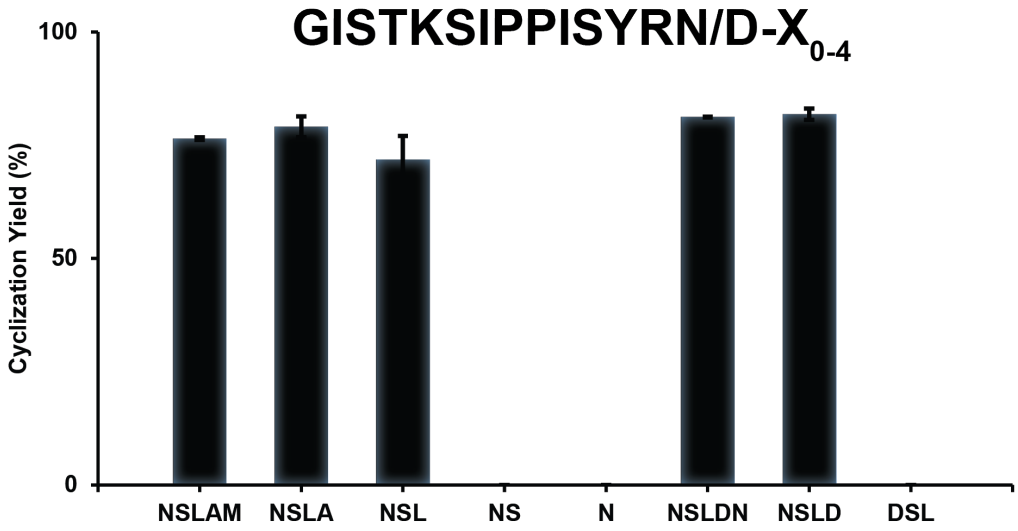


Figure 67. Substrate specificity of VuPAL1 against peptide substrates carrying degenerated recognition motifs. The C-terminal degenerated motifs were derived from the plant *Viola uliginosa*. All reactions were performed in triplicate at pH 6.0 at 37 °C for 1 h. The enzyme-to-substrate ratio was 1:500. The yields were quantified and qualified by RP-HPLC and MALDI-TOF MS, respectively. See **Table 12** for the complete amino acid sequences.

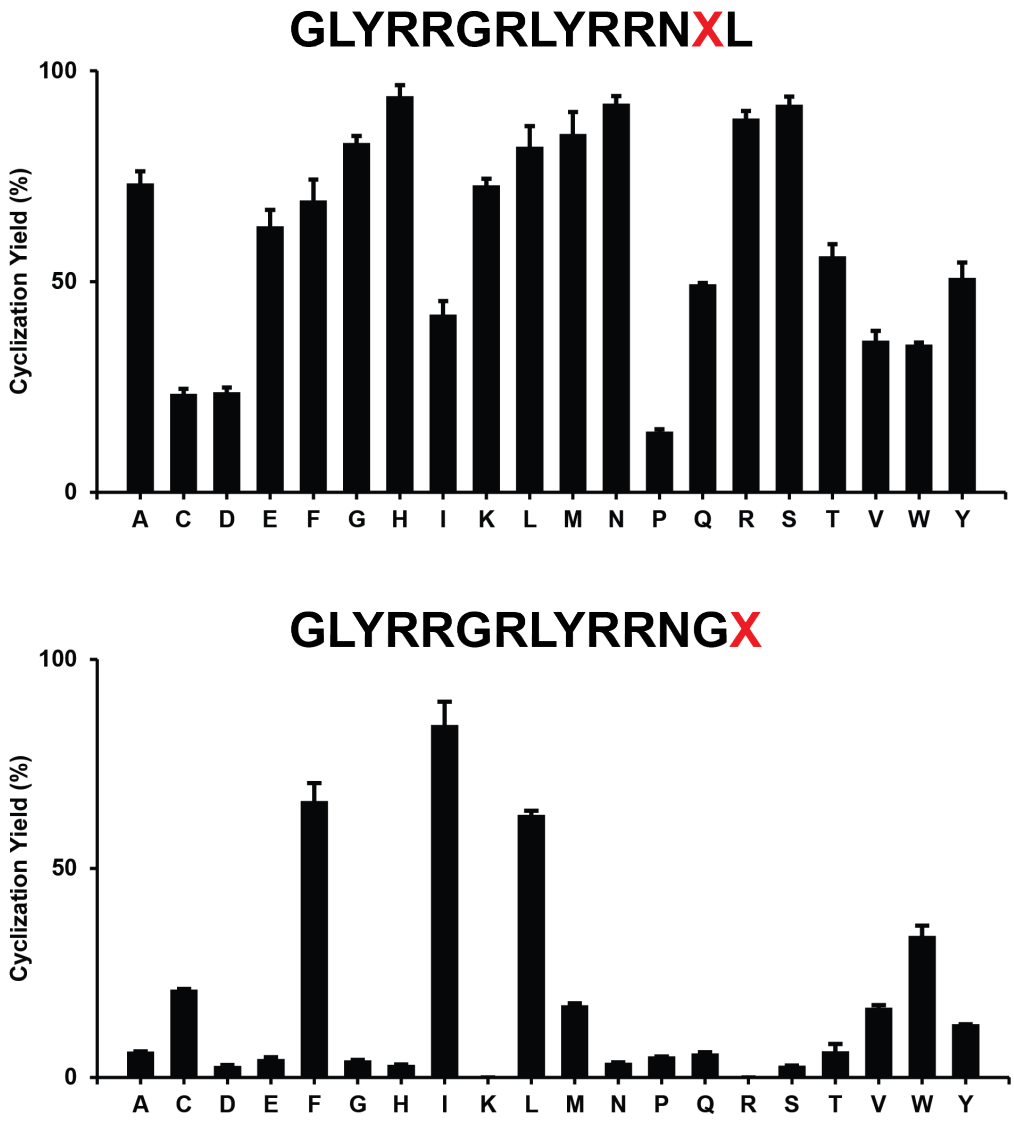


Figure 68. Substrate specificity of VuPAL1 at the P1' and P2' positions. The peptide substrates were based on the model peptide substrate GN12-GL with 20 different native amino acids at the P1' position and the P2' position (the positions of variation were colored in red). All reactions were performed in triplicate at pH 6.0 at 37 °C for 1 h. The enzyme-to-substrate ratio was 1:500. The yields were quantified and qualified by RP-HPLC and MALDI-TOF MS, respectively. See **Table 12** for complete amino acid sequences.

Substrate preference screening of VuPAL1 utilizing 51 peptide substrates demonstrated that a tripeptide recognition signal at the P1-P1'-P2' position is sufficient for VuPAL to render efficient cyclization. No detectable hydrolysis product was observed in all the reactions. The P1 position can only accept Asn residue, and the presence of P1-Asp resulted in significantly lower catalytic efficiency, yielding no detectable linear or cyclic product. In contrast to the strict requirement at the P1 position, the P1' position displayed a broad substrate tolerance, accepting all 20 natural amino acids. At the P2' position, VuPAL1 prefers hydrophobic residues, but except for Arg, the rest 19 amino acids can still be accepted and cyclized.

5.2.7 Kinetics Study of VuPAL1

FRET-based cyclization assay was performed to determine the catalytic efficiency of VuPAL1. The FRET-based peptide substrate GISTKSIPPIE(EDANS)YRNSLK(DABCYL) contained a C-terminal recognition motif Asn-Ser-Leu-Lys (P1-P2'-P2'-P3'), the quencher DABCYL, and the N-terminal fluorophore EDANS. The reactions were performed at 37 °C using Cytation 5 Cell Imaging Multi-Mode Reader (BioTek, USA). Through VuPAL-mediated cyclization, the quencher DABCYL was cleaved, and the fluorescence was released, accompanying the shift in molecular weight from 2401.8 Da to 1805 Da (GISTKSIPPIE-(EDANS)TYN) (**Figure 69**). Michaelis-Menten kinetics using the FRET-based substrates showed that VuPAL1 had a turnover rate (k_{cat}) of 0.01557 s^{-1}, the K_m value of 1.501 μM, and the catalytic efficiency (k_{cat}/K_m) of 10373.08 M^{-1} s^{-1} (**Figure 70**).

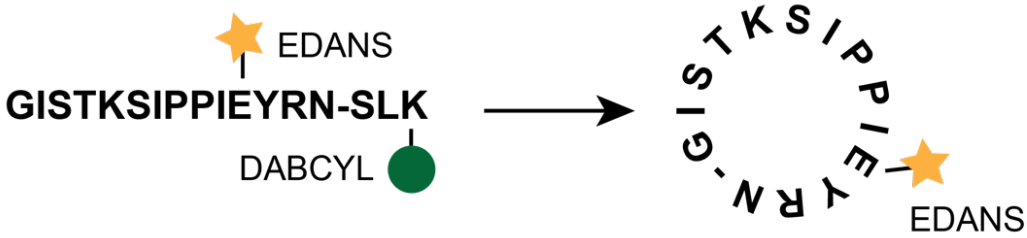


Figure 69. Schematic representation of the cyclization of the FRET-based peptide substrate. The full sequence of the FRET-based peptide substrate was GISTKSIPPIE(EDANS)YRNSLK(DABCYL). Cyclization of the starting materials (2401.8 Da) results in the removal of the leaving group (Ser-Leu-Lys) and the quencher DABCYL (green circle), allowing the fluorescence (yellow star) to be released. The cyclized product (GISTKSIPPIE-(EDANS)TYN) has a molecular weight of 1805 Da.

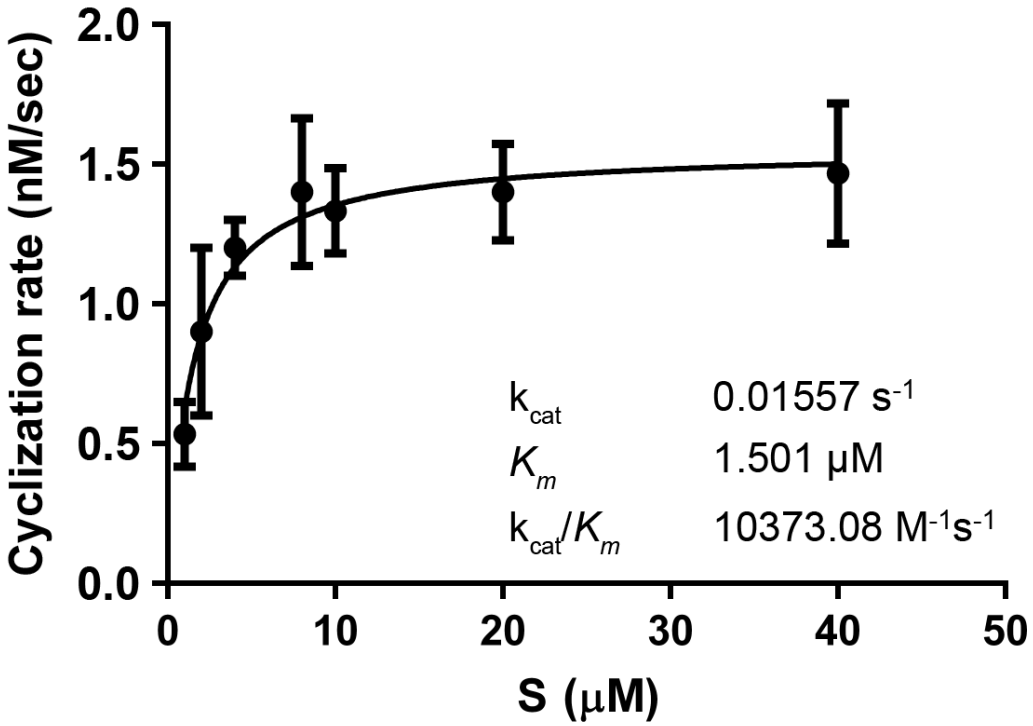


Figure 70. Cyclization efficiency of VuPAL1 against FRET-based peptide substrate. The peptide substrate, GISTKSIPPIE(EDANS)YRNSLK(DABCYL), has a molecular weight of 2401.8 Da. All reactions were performed in triplicate at pH 6.0 at 37 °C.

5.2.8 Characterization of the VuPAL1-I244A Mutant

Previously, it was reported that mutating the Cys247 of the substrate-binding pocket S2 to an Ala enhanced the catalytic efficiency of OaAEP1b 160-fold. Residue at this position was termed 'gatekeeper' and was also identified as one of the residues of the LAD1. To enhance the catalytic efficiency of VuPAL1, the residue of the same position, Ile244 (Val237 of butelase-1, Cys247 of OaAEP1b), was mutated to an Ala, obtaining the mutant VuPAL1-I244A. MALDI-TOF MS profiles of VuPAL1 and the mutant VuPAL1-I244A showed that the mutant VuPAL1-I244A catalyzed the cyclization of peptide substrate GN14-SLDI (GISTKSIPPISYRNSLDI) faster than the wild-type VuPAL1 at pH 6.5 at 42 °C (**Figure 71**). The Michaelis-Menten kinetics of VuPAL1-I244A using the FRET-based substrates was determined under the same condition as VuPAL1. It was revealed that the turnover rate (k_{cat}) of VuPAL1-I244A was 0.0172 s^{-1}, the *K*m value was 1.215 μM, and the catalytic efficiency (k_{cat}/K_m) was 14156.38 M^{-1} s^{-1} (**Figure 72**). Same peptide substrate was used to determine the catalytic efficiency of butelase-1 as a comparison. It was shown that efficiency of butelase-1 was two times higher than that of VuPAL1-I244A mutant. The turnover rate (k_{cat}) of butelase-1 was 0.05350 s^{-1}, the *K*m value was 1.818 μM, and the catalytic efficiency (k_{cat}/K_m) was 29427.94 M^{-1} s^{-1} (**Figure 73**).

Figure 71. MALDI-TOF MS spectra of VuPAL1 and VuPAL1-I244A against peptide substrate GN14-SLDI. The full sequence of the peptide substrate was GISTKSIPPISYRNSLDI. The reaction was performed at pH 6.5 at 42 °C for 10 min with the enzyme-to-substrate ratio of 1:500.

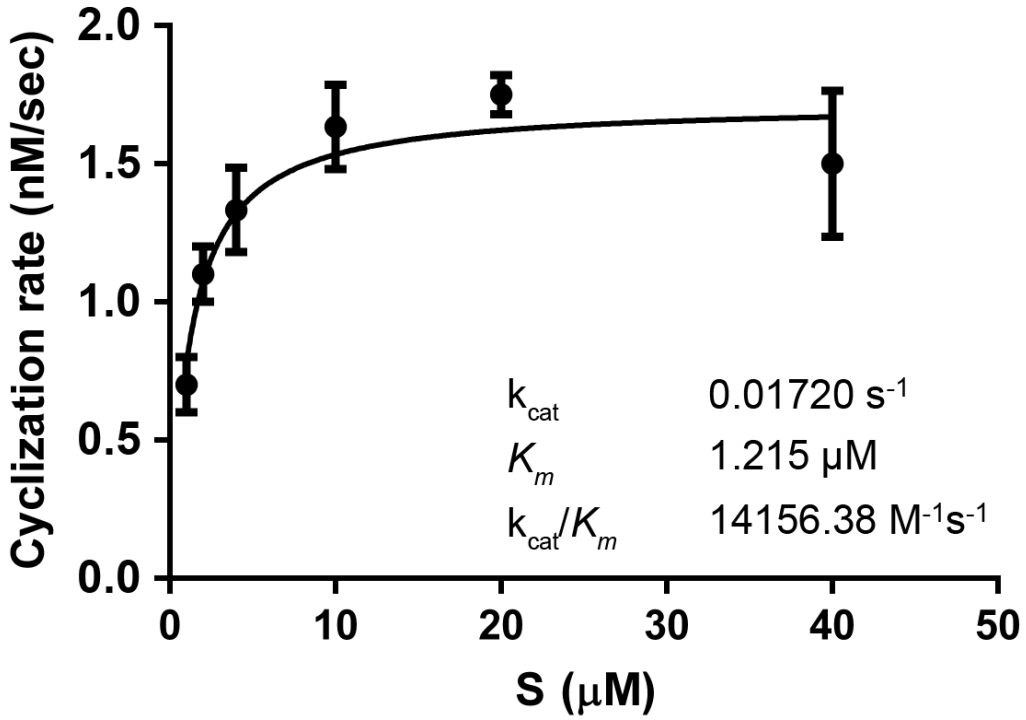


Figure 72. Cyclization efficiency of VuPAL1-I244A against FRET-based peptide substrate. The peptide substrate, GISTKSIPPIE(EDANS)YRNSLK(DABCYL), has a molecular weight of 2401.8 Da. All reactions were performed in triplicate at pH 6.0 at 37 °C.

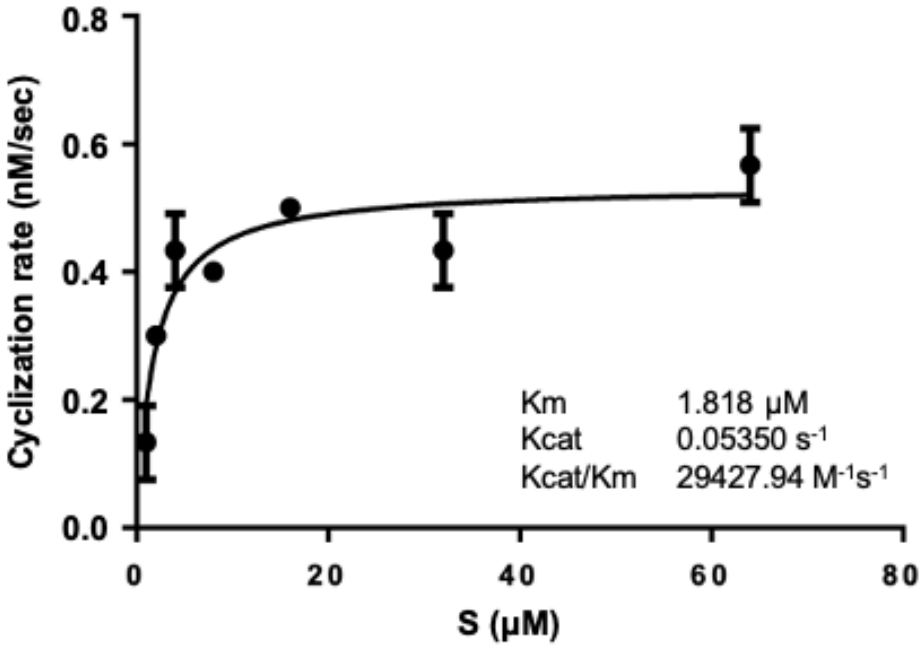


Figure 73. Cyclization efficiency of butelase-1 against FRET-based peptide substrate. The peptide substrate, GISTKSIPPIE(EDANS)YRNSLK(DABCYL), has a molecular weight of 2401.8 Da. All reactions were performed in triplicate at pH 6.0 at 37 °C.

To determine the effect of the mutation on the substrate distribution, cyclization assays using the peptide substrate GN14-SL (GISTKSIPPISYRNSL) carrying the native C-terminal recognition motif were performed. All reactions by both enzymes were performed in triplicate at 37 °C for 30 min with the enzyme-to-substrate ratio of 1:500. OaAEP2, a previously reported prototypic protease, was used as a negative control and screened using the peptide substrate GN14-GL (GISTKSIPPISYRNGL). The yields were quantified and qualified by RP-HPLC and MALDI-TOF MS, respectively (**Figure 74 & Figure 75**). The activity of wild-type VuPAL1 dropped largely when pH was equal to or higher than 7.5, or equal to and lower than 4.5. The lowest activity of VuPAL1 was found at pH 4.0 with less than 5% yield of cyclic products cGN12. Under the same condition, the optimum reaction pH of the mutant VuPAL1-I244A seemed to shift toward basic pH slightly. The cyclization efficiency of VuPAL1-I244A peaked at pH 6.5, and more than 99% of the starting materials were converted to cyclic product cGN12. At pH 7.5 and pH 8.0, the yields of the cyclic product were 62.1% and 57.6% for VuPAL1-I244A, which were higher than that of VuPAL1 (11.0% at pH 7.5 and 11.2% at pH 8.0). No linear product was detected in all assays of VuPAL1-mediated cyclization (**Figure 76**). Linear products GN12 were detected in reactions catalyzed by the mutant VuPAL1-I244A by MALDI-TOF MS, however, the linear products were unable to be detected and quantified by RP-HPLC (**Figure 76**). The results suggested that making the residue at the LAD1 slightly smaller and making the LAD1 leakier did not shift the enzymatic activity of VuPAL1 significantly.

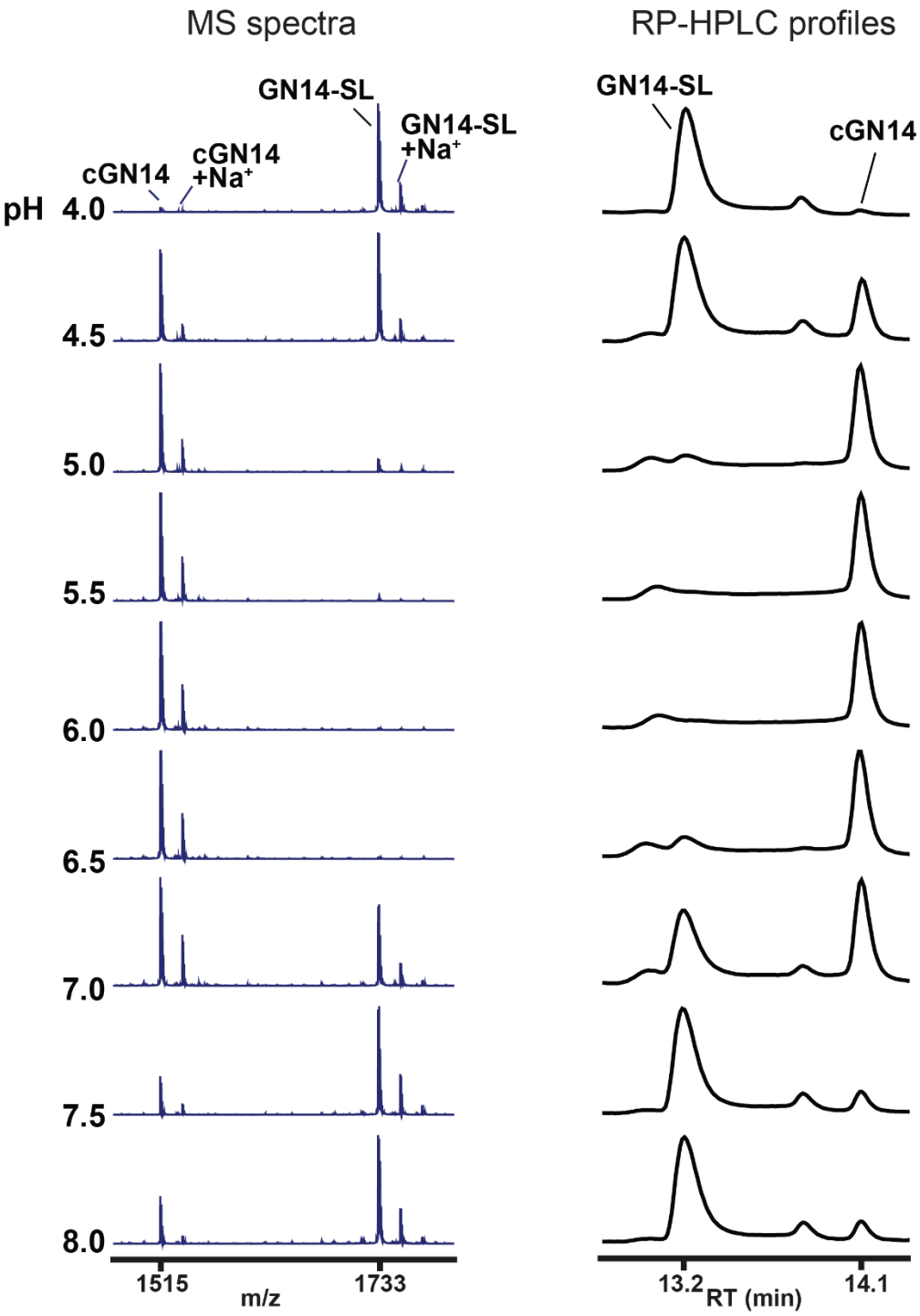


Figure 74. The MALDI-TOF MS spectra and RP-HPLC profiles of VuPAL1-mediated cyclization. The peptide substrate used was GN14-SL (GISTKSIPPISYRNSL). The reactions were performed at 37 °C from pH 4.0 to 8.0. RT stands for retention time.

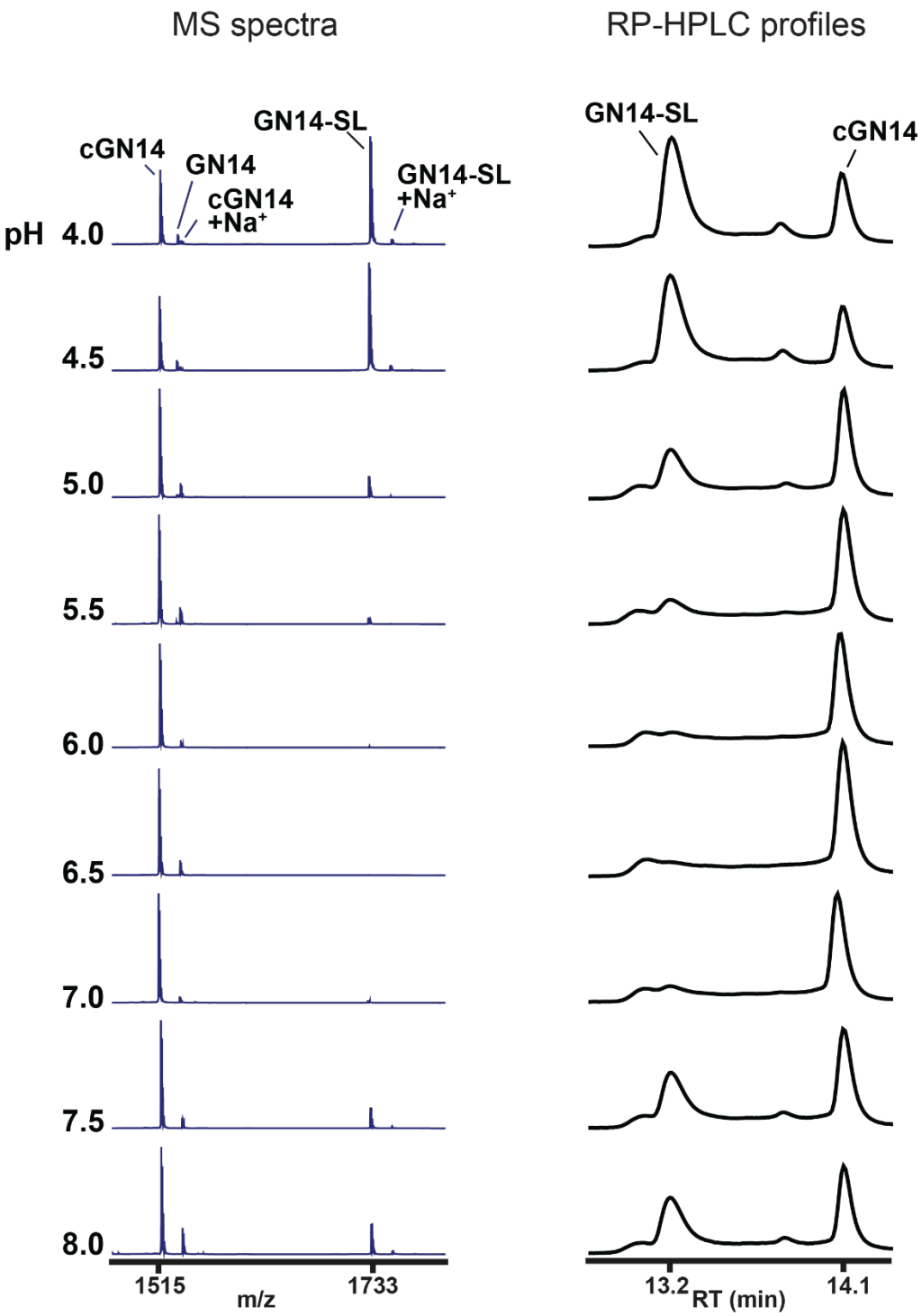


Figure 75. The MALDI-TOF MS spectra and RP-HPLC profiles of cyclization catalyzed by the mutant VuPAL1-I244A. The peptide substrate used was GN14-SL (GISTKSIPPISYRNSL). The reactions were performed at 37 °C from pH 4.0 to 8.0. RT stands for retention time.

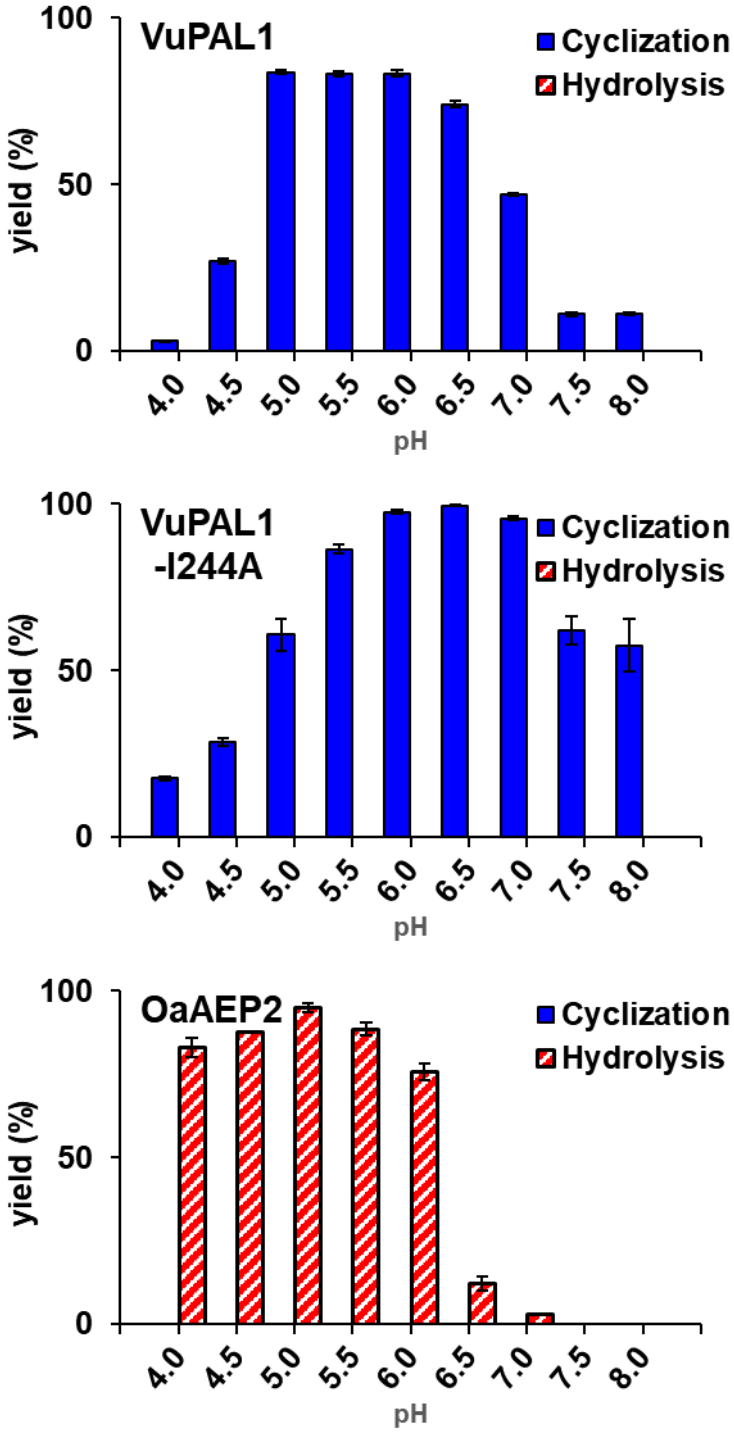


Figure 76. pH-dependent cyclization efficiency of VuPAL1, VuPAL1-I244A, and OaAEP2. OaAEP2 was used as a negative control. The peptide substrate used was GN14-SL (GISTKSIPPISYRNSL) and GN14-GL (GISTKSIPPISYRNGL) for VuPAL1 and VuPAL-I244A, and OaAEP2, respectively. All reactions were performed in triplicate at 37 °C for 30 min with the enzyme-to-substrate ratio of 1:500. The yields were quantified and qualified by RP-HPLC and MALDI-TOF MS, respectively.

5.2.9 Modulation of Enzymatic Activity of BmAEP1 by Mutating the LAD2

To provide additional validation for the pivotal roles of LAD2 in modulating ligase activity, a mutation at the LAD2 was introduced to the selected butelase-2-like protease BmAEP1. The Ser-Ala motif at the LAD2 of BmAEP1 was replaced with an Ala-Ala dipeptide motif, which is commonly observed in butelase-1-like PALs, by mutagenesis, generating the mutant BmAEP1-S161A. BmAEP1 and its mutant BmAEP1-S161A were mixed with the model peptide substrate GN12-GL at 37 °C for 15 min and 24 h, respectively. Linear products GN12 were observed in BmAEP1-mediated processing of GN12-GL from pH 4.0 to 8.0. In contrast, only cyclic products were detected in the reactions by BmAEP1-S161A by MALDI-TOF MS (**Figure 77**). The results showed that The Ser-to-Ala mutation abolished the hydrolysis activity of BmAEP1 largely. However, as the yields of cyclic products of BmAEP1-S161A-mediated cyclization were low, the mutation also seemed to reduce the catalytic efficiency of BmAEP1 largely.

Notably, although there were more cyclized products than linear products detected from pH 6.5 to pH 8.0, the cyclized products were subjected to cleavage by BmAEP1. BmAEP1 was incubated with peptide substrate GN12-GL at pH 6.5, and the reactions were stopped at specific time intervals, 5 min, 15 min, 30 min, and 24 h, then immediately monitored by MALDI-TOF MS to examine the enzymatic activity of BmAEP1 over time. It was shown that the starting materials were depleted after 30 min of incubation at pH 6.5 at 37 °C and the cyclized products could be further cleaved, resulting in an increase and decrease of the amount of hydrolyzed products and cyclized products detected, respectively, after 24 h of incubation (**Figure 78**).

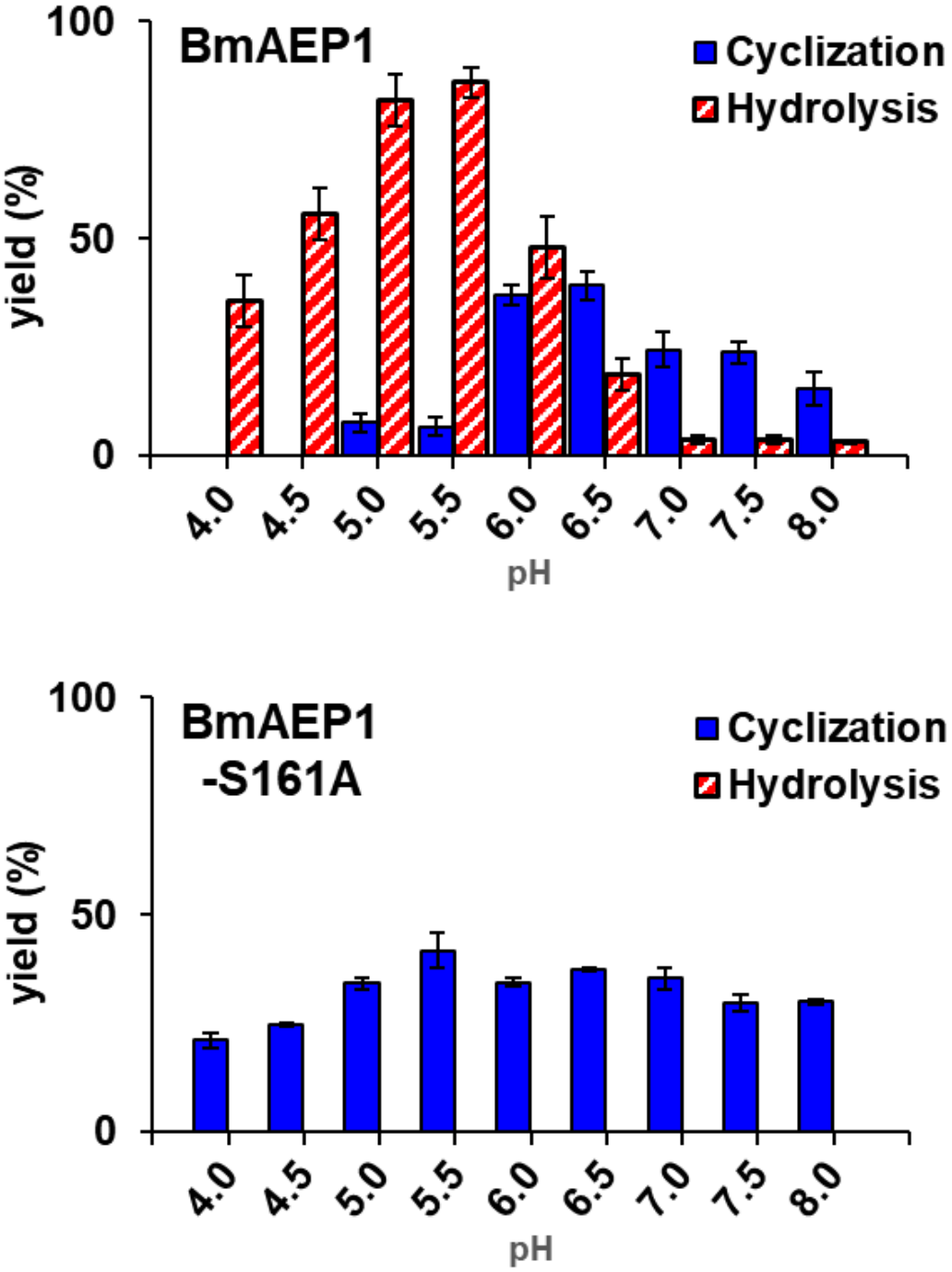


Figure 77. pH-dependent cyclization and hydrolysis efficiency of BmAEP1 and BmAEP1-S161A. All reactions were performed in triplicate at 37 °C with the enzyme-to-substrate ratio of 1:500. The reactions were performed for 15 min and 24 h for BmAEP1 and BmAEP1-S161A, respectively. The yields were quantified and qualified by RP-HPLC and MALDI-TOF MS, respectively.

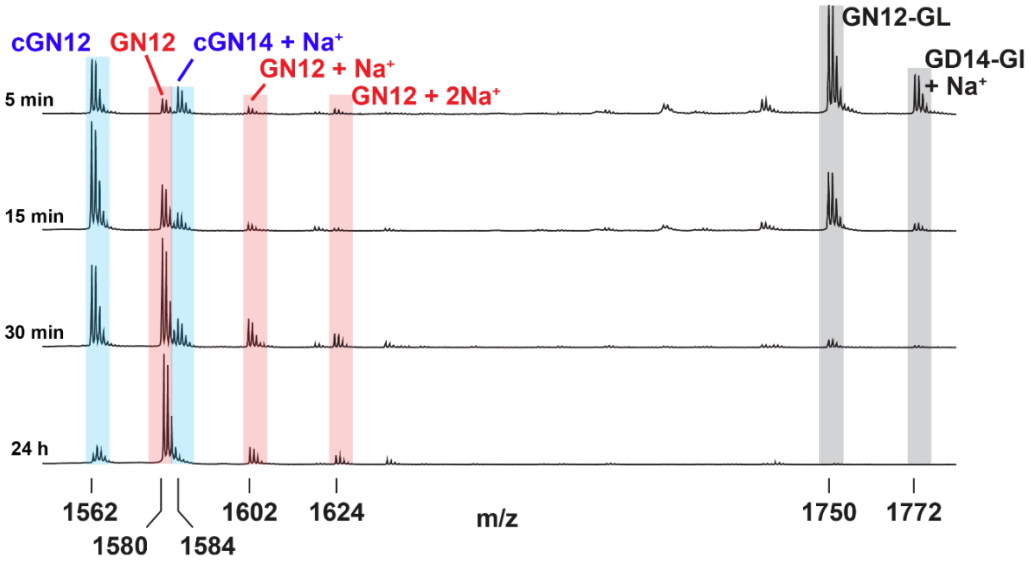


Figure 78. The MALDI-TOF MS spectra of BmAEP1-mediated cyclization and hydrolysis. The reactions were performed using the peptide substrate GN12-GL at specific time intervals (5 min, 15 min, 30 min, and 24 h) at pH 6.5 at 37 °C with the enzyme-to-substrate ratio of 1:500. The cyclized products were shaded blue, the hydrolyzed products were shaded red, and the starting materials were shaded grey.

5.2.10 Substrate-Dependent Peptide Bond Formation of BmAEP1

To further characterize BmAEP1, several peptide substrates modified from GN14-X_{0-4} with different C-terminal tails (P1-P2' position) were tested at pH 6.5 at 37 °C for 30 min. It was found that BmAEP1 preferred to cyclize substrates with an Asp, rather than an Asn, at the P1 position of the peptide substrates. Using the peptide substrate GN14-GI, BmAEP1-mediated cyclization and hydrolysis resulted in almost undetectable amounts of cyclic products cGN14. In contrast, using the 16-mer peptide substrate GD14-GI, which contained an Asp at the P1 position, led to the conversion of starting materials to more cyclic products than linear products (**Figure 79**). Similarly, BmAEP1 favored cyclization when the peptide substrate GD14-SL was used, and BmAEP1 cleaved most of the starting materials, GN14-SL (**Figure 80**).

These resulted suggested that using peptide substrates with a P1-Asp may facilitate the cyclization reaction of BmAEP1. As there is no cyclic peptide reported in the plant *Momordica charantia* [38], linear peptide substrates for BmAEP1, GN10-AL (GLRRGYSGSNAL) and GD10-AL (GLRRGYSGSDAL), were modified from the native cyclic peptide precursor McoTI-II of *Momordica cochinchinensis*, which is a member of the cucumber family. GD10-AL contained the native C-terminal tail, Ser-Asp-Ala-Leu at P2-P1-P1'-P2' position, of McoTI-II.

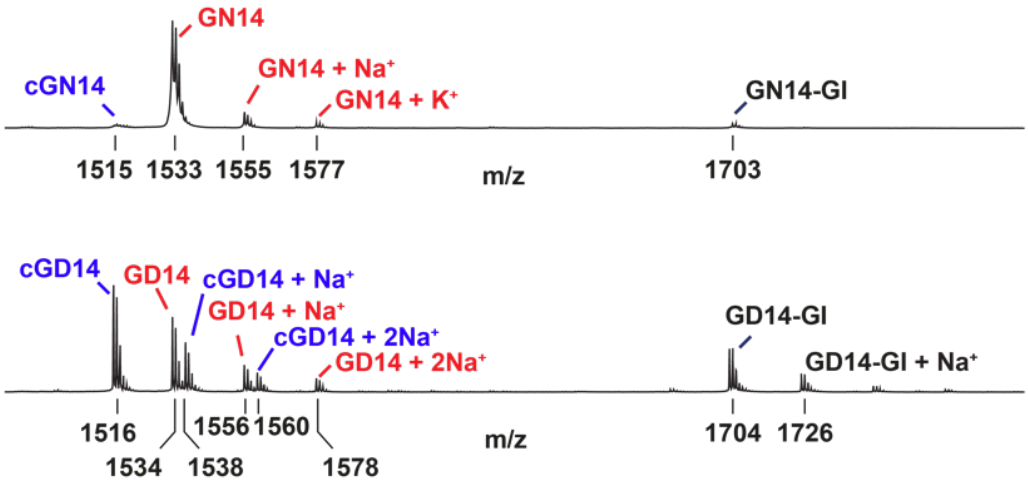


Figure 79. MALDI-TOF MS spectra of BmAEP1-mediated cyclization and hydrolysis of peptide substrate GN14-GI and GD14-GI. The full sequences of the peptide substrates GN14-GI and GD14-GI were GISTKSIPPISYRNGI and GISTKSIPPISYRDGI, respectively. The reactions were performed at pH 6.5 at 37 °C with the enzyme-to-substrate ratio of 1:500 for 30 min. The cyclized products (cGN14 and cGD14) were labeled in blue, the hydrolyzed products (GN14 and GD14) were labeled in red, and the starting materials were labeled in black.

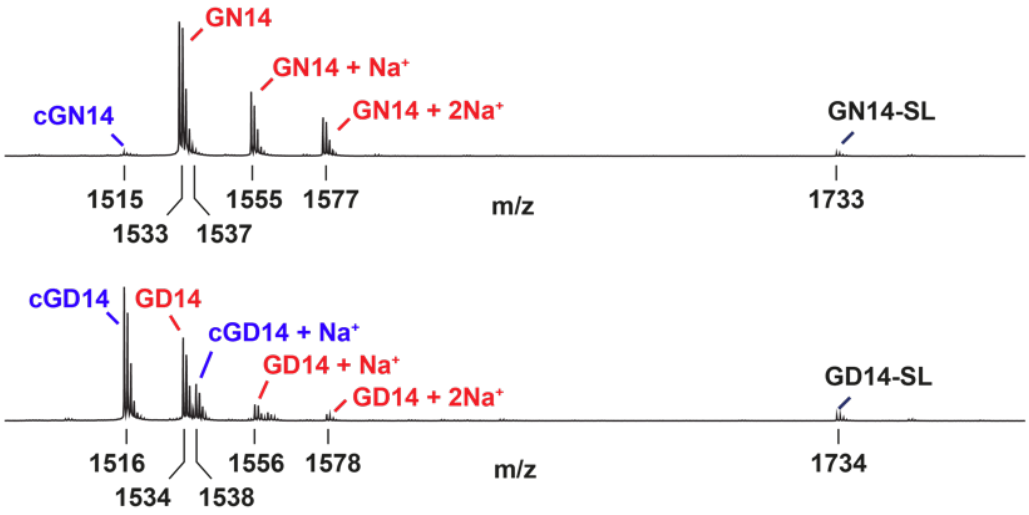


Figure 80. MALDI-TOF MS spectra of BmAEP1-mediated cyclization and hydrolysis of peptide substrate GN14-SL and GD14-SL. The full sequences of the peptide substrates GN14-SL and GD14-SL were GISTKSIPPISYRNSL and GISTKSIPPISYRDSL, respectively. The reactions were performed at pH 6.5 at 37 °C with the enzyme-to-substrate ratio of 1:500 for 30 min. The cyclized products (cGN14 and cGD14) were labeled in blue, the hydrolyzed products (GN14 and GD14) were labeled in red, and the starting materials were labeled in black.

The cyclase activity of BmAEP1 was increased largely at acidic pH (4.5-6.0) using GD10-AL. At pH 6.0, using GN10-AL as starting materials resulted in more than 50% of linear products, while using GD10-AL yielded no detectable linear product at pH 6.0 (**Figure 81**). From pH 7.0 to 8.0, GN10-AL was predominantly cyclized. While at the same pH range, the catalytic efficiency of BmAEP1 against GD10-AL was very low. There was no cyclic product cGD10 detected in the reactions at pH 7.5 and pH 8.0. The results showed that BmAEP1, a predicted protease, could also cyclize using specific substrates at certain pH.

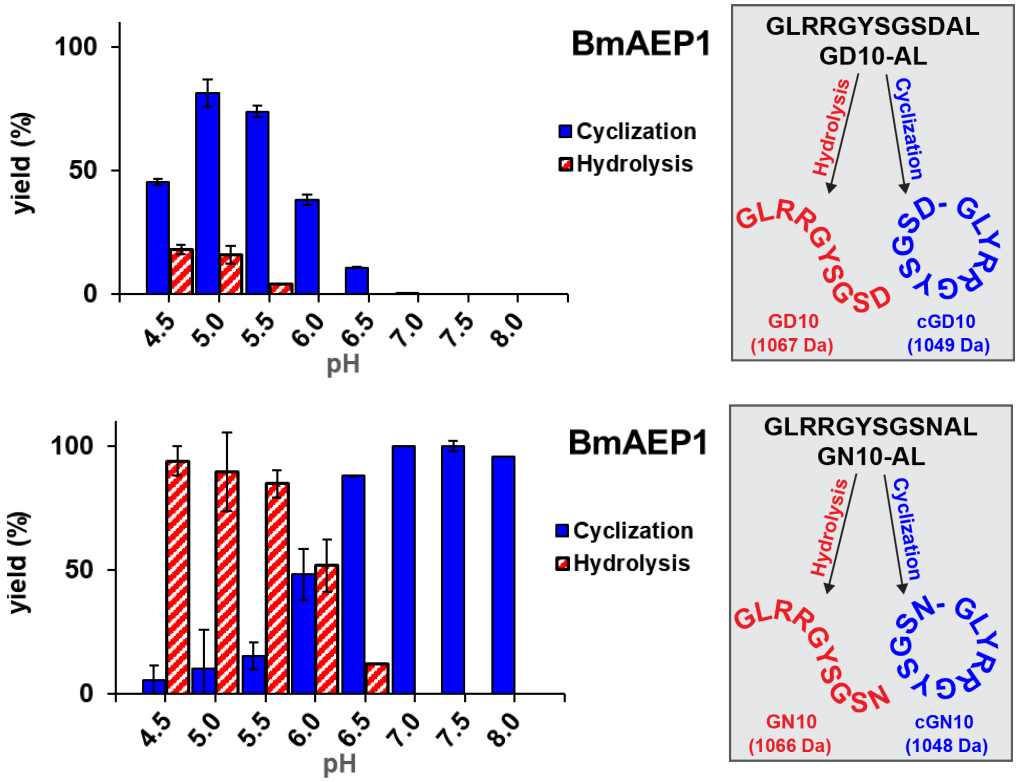


Figure 81. pH-dependent and substrate-dependent cyclization and hydrolysis efficiency of BmAEP1. All reactions were performed in triplicate at 37 °C with the enzyme-to-substrate ratio of 1:500. The reactions were performed for 30 min. The results were quantified and qualified by MALDI-TOF MS. Schematic representations of the processing of starting materials, GD10-AL (GLYRRGYSGSDAL) and GN10-AL (GLRRGYSGSNAL), are on the right.

5.3 Discussion

AEPs are comprised of nearly 500 amino acids, and it was demonstrated that we were able to predict the presence of butelase-1-like PALs using LADs as a filtering criterion in various plant species. The predicted PALs, VoPAL1, VuPAL1, VvPAL1, VyPAL4, and VyPAL5, all functioned as an Asn-specific ligase from pH 4.0 to 8.0. While PiPAL1, PePAL1, PePAL2, SiPAL1, and VcAEP, which have only one LAD conserved with the motifs in known PALs, demonstrated how partly conserved motifs indicate an AEP exhibiting both ligase and protease activity in a pH-dependent manner.

It was previously reported that a partial ligase from the common sunflower (*Helianthus annuus*), HaAEP, was able to cyclize the sunflower trypsin inhibitor 1 (SFTI-1), a potent trypsin inhibitor, but with low efficiency [98]. Bernath-Levin *et al.* proposed that the low efficiency may be a result of insufficient selection pressure, as SFTI-1 is not abundant, not ubiquitous in the genus *Helianthus,* and not the only trypsin inhibitors in the common sunflower [98, 203]. Similarly, the presence of several butelase-1like PALs in the Violaceae family may be a result of evolutionary selection due to the abundance of the cyclic peptides in the Violaceae family (**Figure 13, Appendix A**) [38]. And the inability of VuPAL1 to efficiently cyclize peptide substrates containing an Asp at the P1 position may result from the insufficient selection pressure, as there is no P1-Asp-containing cyclic peptide discovered in the plant *Viola uliginosa* hitherto. Overall, the discovery of five more butelase-1-like PALs expands the catalog of available PALs for biochemical applications.

The pH-dependent cyclization efficiency profiles of VoPAL1 showed that its cyclase activity was higher at acidic pH and peaked at pH 4.5, which was

also reported in the OaAEP3-5 [9]. As VoPAL1 and OaAEP3-5 performed cyclization without producing detectable hydrolysis products at acidic pH and exhibited different substrate preferences, it is possible to apply them in tandem or a one-pot condition for modifications of peptides and proteins at acidic condition.

Modification of the catalytic residues or residues nearest the catalytic site has been exploited to impair the protease activity with ligase activity retained. For instance, mutation of catalytic Ser of subtilisin BPN' impaired the hydrolysis activity [69, 70]. Similarly, mutations of residues lie within substrate-binding pockets of trypsin (K60E, N143H, E151H, D189K) not only impaired the protease activity but also altered the substrate preference of trypsin, resulting in an improved variant of trypsin, trypsiligase [81]. Thus, our study on the LADs, which include residues inside or in the proximity of the substrate-binding pockets, could play a critical role in modulating the directionality of AEP enzymes to act as ligases. Mutagenesis studies at the LADs further reinforced the importance of LADs in modulating the ligase activity of PALs and AEPs. At LAD1, mutating the bulky and hydrophobic Ile to smaller Ala in VuPAL1 resulted in no increase in hydrolysis activity. Compared to previously reported OaAEP-C237A [126], which also had its LAD1 mutated, the ligase kinetics of VuPAL1-I244A is similar to that of wild-type VuPAL1. Mutating the residues at LAD1 to Ala did not lead to enhanced hydrolysis activity for both OaAEP1b and VuPAL1. Our results and the study by Yang *et al.* [126] suggest that the S2 pocket is capable of modulating the catalytic efficiency of PALs but may only play a minor role in modulating the enzymatic directionality of PALs and AEPs.

The importance of the primed side pockets has been highlighted in a recent study. Zauner et al. showed that the hydrophobicity at the S2' pocket plays an essential role in modulating the catalytic efficacy. The substitution of Tyr190 in AtLEGγ to less hydrophobic His, significantly increased proteolysis activity [128]. We speculate that the amino acids having smaller and less bulky side chains nearby the S1 site can lead to the leakage of a water molecule into the S1 catalytic pocket, resulting in the hydrolysis of the intermediate thioester bond. In contrast, displacing water molecule before the nucleophile attack by the incoming amino acid results in ligation reaction (**Figure 82**).

Zauner *et al.* showed that retaining the cleaved leaving group (P1'-P2' position) of the substrate at the primed pockets promotes cyclization. In contrast, the lack of a hydrophobic patch at the S2' pocket promotes proteolysis [128]. Therefore, it was postulated that the hydrophobic interaction between the P1' −P2' and the S1'-S2' pockets retain the substrate leaving group, displacing water molecule before the incoming P1" nucleophile attacks on the acyl-enzyme intermediate. However, in the case of VcAEP, a hydrophobic Y168 near the S1' pocket was found to disfavor ligation, and substitution to a less bulky Ala promoted ligase activity. The Y168A mutation of VcAEP at the LAD1 enhanced the ligase activity of the VcAEP drastically [10]. These results also showed that replacing the hydrophilic Ser with a small and hydrophobic residue, Ala, at the LAD1 increased ligase activity of BmAEP1 largely. It was speculated that the bulkiness of the Tyr aromatic sidechain hinders the entry of the incoming peptide nucleophile P1" and facilitates the departure of the leaving dipeptide (P1'-P2'). Subsequently, the recruited catalytic water attacks the acyl-enzyme intermediate at the S1 pocket and thus promotes hydrolysis. Conversely, the presence of a

small hydrophobic patch, which is conserved in known PALs, such as Gly-Ala 167-168 of butelase-1 at the S1' pocket, retains the cleaved leaving group. Subsequently, the access of catalytic water toward the thioester bond is blocked, allowing the nucleophilic P1" residue to binding to the S1' pocket, thereby displacing the leaving peptide group and forming a peptide bond between P1 and P1" residues of substrate. We propose that the LAD2 (S1' pocket) is more important than the LAD1 (the S2 pocket) in terms of governing the tendency of the enzyme to ligate or to cleave. The LAD2 is involved in the accessibility of catalytic water and amine nucleophile to the S-acyl intermediate. In contrast, the LAD1 appears only to affect the orientation of the substrate binding.

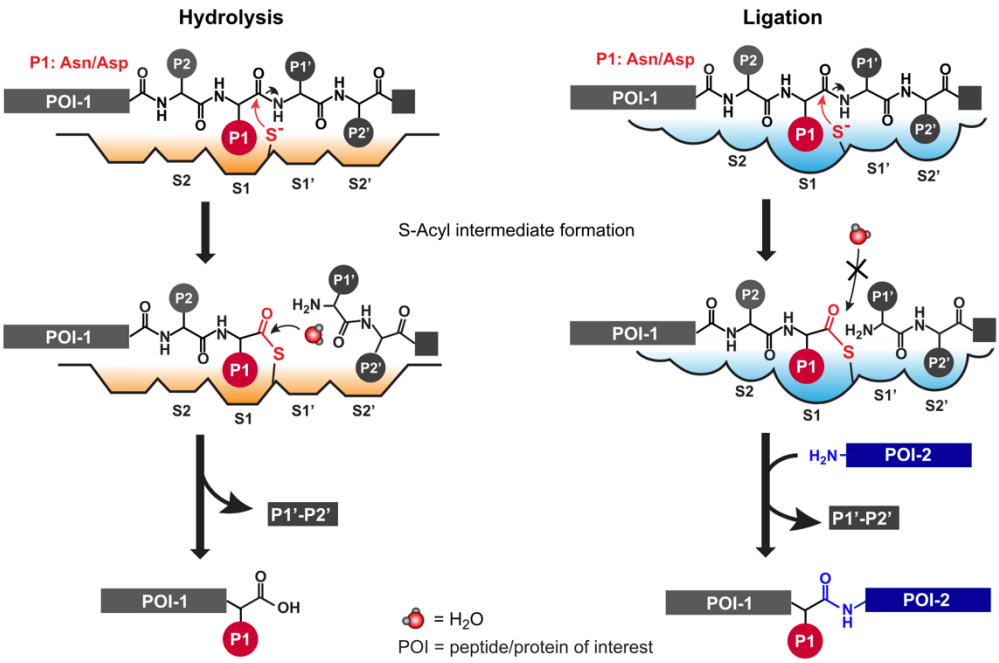


Figure 82. Schematic representation of putative mechanism of proteolysis and ligation mediated by AEPs and PALs. The leaving group replaced by a water molecule results in hydrolysis. In contrast, the leaving group replaced by incoming peptide or protein results in ligation. Figure taken from Hemu et al [131].

Datasets of this study include 23 sequences from the cucumber family, and most of them are predicted to be predominant protease (**Figure 83**). However, by using P1-Asp-containing peptide substrates modified from cyclic peptide precursor McoTI-II of *Momordica cochinchinensis*, BmAEP1 was able to catalyze cyclization reactions at acidic pH. Recent studies also revealed a predicted protease, McoAEP1 (LAD1: Gly, LAD2: Gly-Pro), to process peptide substrates with a P1-Asp at acidic pH [99]. It is also worthwhile to note that BmAEP1 is from the Cucurbitaceae family, which is known to produce cyclotides that often contain an Asp instead of an Asn at the P1 site [204]. Our results and the previous studies suggested that LAD-based prediction may be Asn-specific. An important caveat is that the ligase activity could also be influenced by the substrates used. Recent reports suggested that AEPs could act as ligases using different substrates. AtLEGγ was reported to process the seed storage protein precursors, which require the proteolytic activity of AEPs [32]. However, using the modified sunflower trypsin inhibitor 1 (SFTI-1), Brandstetter, and coworkers obtained about 90% cyclic products [128]. Similarly, CeAEP1 from jack bean hydrolyzed most SFTI-GLDN substrates [98], and the same enzyme is responsible for the post-translational modification of the circularly permutated concanavalin A [96].

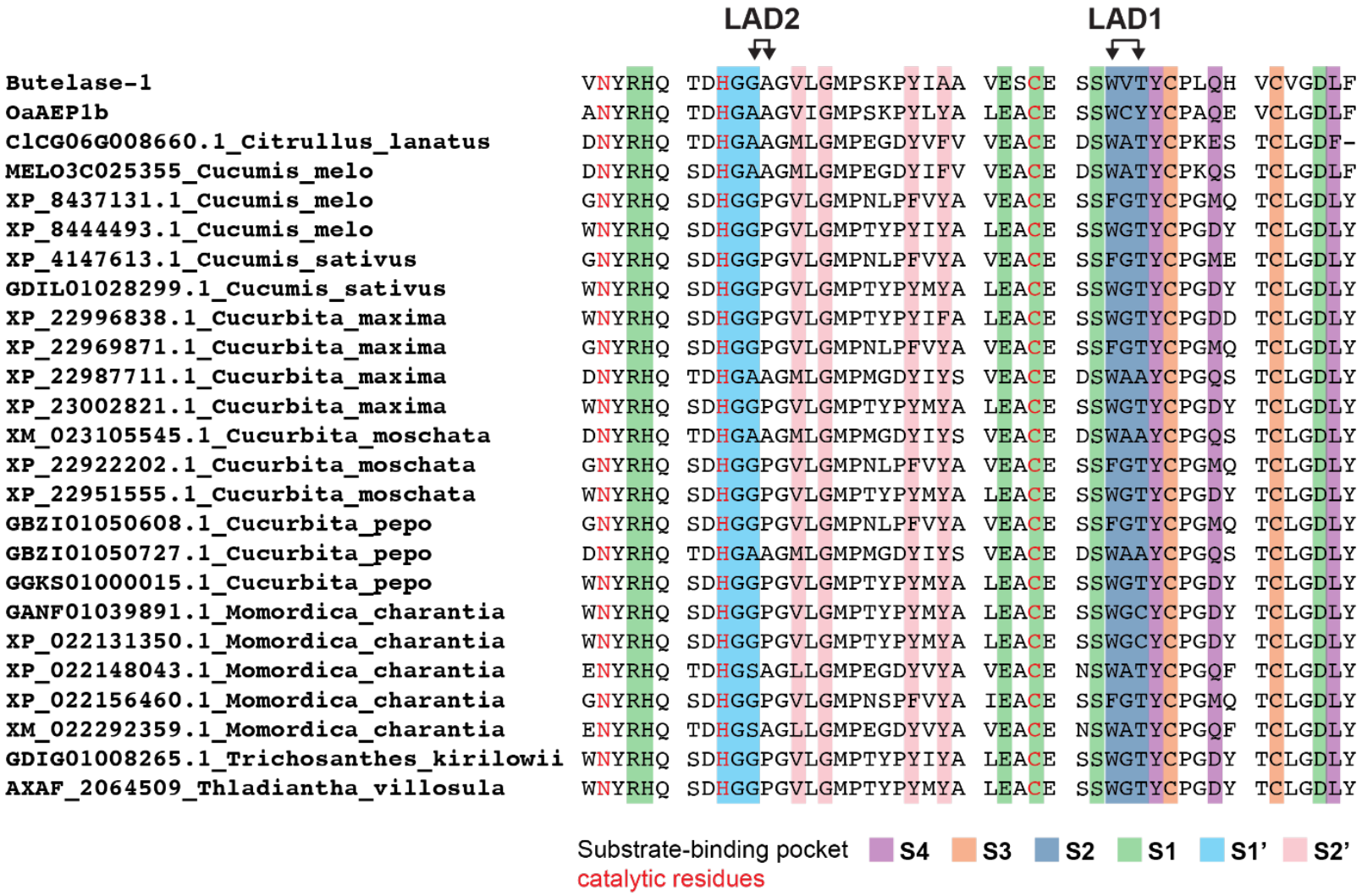

LAD2
LAD1
Butelase-1 VNYRHQ TDHGGAGVLGMPSKPYIAA VESCE SSWVTYCPLQH VCVGDLF
OaAEP1b ANYRHQ TDHGAAGVIGMPSKPYLYA LEACE SSWCYYCPAQE VCLGDLF
ClCG06G008660.1_Citrullus_lanatus DNYRHQ TDHGAAGMLGMPEGDYVFV VEACE DSWATYCPKES TCLGDF-
MELO3C025355_Cucumis_melo DNYRHQ SDHGAAGMLGMPEGDYIFV VEACE DSWATYCPKQS TCLGDLF
XP_8437131.1_Cucumis_melo GNYRHQ SDHGGPGVLGMPNLPFVYA VEACE SSFGTYCPGMQ TCLGDLY
XP_8444493.1_Cucumis_melo WNYRHQ SDHGGPGVLGMPTYPYIYA LEACE SSWGTYCPGDY TCLGDLY
XP_4147613.1_Cucumis_sativus GNYRHQ SDHGGPGVLGMPNLPFVYA VEACE SSFGTYCPGME TCLGDLY
GDIL01028299.1_Cucumis_sativus WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGTYCPGDY TCLGDLY
XP_22996838.1_Cucurbita_maxima WNYRHQ SDHGGPGVLGMPTYPYIFA LEACE SSWGTYCPGDD TCLGDLY
XP_22969871.1_Cucurbita_maxima GNYRHQ SDHGGPGVLGMPNLPFVYA VEACE SSFGTYCPGMQ TCLGDLY
XP_22987711.1_Cucurbita_maxima DNYRHQ TDHGAAGMLGMPMGDYIYS VEACE DSWAAYCPGQS TCLGDLY
XP_23002821.1_Cucurbita_maxima WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGTYCPGDY TCLGDLY
XM_023105545.1_Cucurbita_moschata DNYRHQ TDHGAAGMLGMPMGDYIYS VEACE DSWAAYCPGQS TCLGDLY
XP_22922202.1_Cucurbita_moschata GNYRHQ SDHGGPGVLGMPNLPFVYA VEACE SSFGTYCPGMQ TCLGDLY
XP_22951555.1_Cucurbita_moschata WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGTYCPGDY TCLGDLY
GBZI01050608.1_Cucurbita_pepo GNYRHQ SDHGGPGVLGMPNLPFVYA VEACE SSFGTYCPGMQ TCLGDLY
GBZI01050727.1_Cucurbita_pepo DNYRHQ TDHGAAGMLGMPMGDYIYS VEACE DSWAAYCPGQS TCLGDLY
GGKS01000015.1_Cucurbita_pepo WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGTYCPGDY TCLGDLY
GANF01039891.1_Momordica_charantia WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGCYCPGDY TCLGDLY
XP_022131350.1_Momordica_charantia WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGCYCPGDY TCLGDLY
XP_022148043.1_Momordica_charantia ENYRHQ TDHGSAGLLGMPEGDYVYA VEACE NSWATYCPGQF TCLGDLY
XP_022156460.1_Momordica_charantia GNYRHQ SDHGGPGVLGMPNSPFVYA IEACE SSFGTYCPGMQ TCLGDLY
XM_022292359.1_Momordica_charantia ENYRHQ TDHGSAGLLGMPEGDYVYA VEACE NSWATYCPGQF TCLGDLY
GDIG01008265.1_Trichosanthes_kirilowii WNYRHQ SDHGGPGVLGMPTYPYIYA LEACE SSWGTYCPGDY TCLGDLY
AXAF_2064509_Thladiantha_villosula WNYRHQ SDHGGPGVLGMPTYPYMYA LEACE SSWGTYCPGDY TCLGDLY
Substrate-binding pocket S4 S3 S2 S1 S1' S2'
catalytic residues

Figure 83. Multiple sequence alignment of butelase-1, OaAEP1b, and sequences of the cucumber family in this dataset. The sequences were aligned by Clustal Omega (available online at: https://www.ebi.ac.uk/Tools/msa/clustalo/) [138] using JalView [139, 140].

Apart from the substrate, pH also affects ligase activity. Brandstetter and coworkers demonstrated that a protease AEP, AtLEGγ, can display both protease and ligase activity, which is regulated by pH [127]. Moreover, incomplete activation of AEP could reverse the catalytic AEP to inactive proenzyme [124], and activation of recombinantly expressed AEPs frequently leads to active forms of various sizes that differ by more than 10 kDa [132]. These catalytic enzymes of various sizes could possess different levels of ligase activity. Notably, not all expression systems allow glycosylation, which leads to recombinant AEPs with potentially diminished stability and affected enzyme activity. It is thus very important to note that the enzymatic activity of PALs could be modulated by not only the evolutionary important sequence motifs, but also the substrates, the reaction pH, and the activation of the enzymes.

Chapter 6 Summary

By bioinformatic analyses and biochemical assays, it is concluded that the amino acid compositions at the substrate-binding pockets flanking the S1 pocket of the S-acyl intermediate, in particular, LAD1 at the non-primed side and the LAD2 at the primed side, are crucial in governing the activity preference between ligation and hydrolysis. As efficient PALs, hydrophobic amino acids such as Ile, Val, and Cys are preferred at the LAD1, and at the LAD2, dipeptide motifs of Ala-Pro, Ala-Ala, and Gly-Ala are preferred as they are conserved in PALs reported hitherto. The presence of aromatic residues such as Tyr or hydrophilic residue such as Ser is disfavored at the LAD2, as illustrated in VcAEP1 and BmAEP1. Using the LAD hypothesis, five butelase-1-like PALs were identified. Our findings elaborate the molecular basis of PALs and significantly expand their repertoire library, which is beneficial in the industrial-scale synthesis of macrocyclic peptides in novel pharmaceutics and drug discovery. Additionally, LAD1 and LAD2 represent potential candidates in rational design and engineering of PAL with tailored catalytic properties and broaden substrate specificity.

Publications

Chan, N-Y, Hemu, X, Wang, K, Chen, Y, Liew, HT., Hu, S, Zhang, X, Serra, A, Sze, SK, Lescar, J, Tam, JP., *Discovery and Characterization of Ligase Activity Determinants in Plant Asparaginyl Endopeptidases,* (In preparation).

Liu, C-F, Zhang, D, Wang, Z, Hu, S, Chan, N-Y, Liew, HT, Lescar, J, Tam, JP, *AEP-mediated protein C-terminal hydrazinolysis for the synthesis of bioconjugates*, (In preparation).

Liew, HT, To, J, Zhang, X, Hemu, X, Chan, N-Y, Serra, A, Sze, SK, Liu, C-F, Tam, JP, *The legumain McPAL1 from Momordica cochinchinensis is a highly stable Asx-specific splicing enzyme*. Journal of Biological Chemistry, 2021. **297**(6): p. 101325.

Xia, Y, To, J, Chan, N-Y, Hu, S, Liew, HT, Balamkundu, S, Zhang, X, Lescar, J, Bhattacharjya, S, Tam, JP, Liu, C-F, *Nγ-Hydroxyasparagine: A Multifunctional Unnatural Amino Acid That is a Good P1 Substrate of Asparaginyl Peptide Ligases.* Angewandte Chemie International Edition, 2021. **60**: p. 22207.

Tam, JP, Chan, N-Y, Liew, HT, Tan, SJ, and Chen, Y, *Peptide asparaginyl ligases—renegade peptide bond makers*. Science China Chemistry, 2020. **63**: p. 296-307.

Appendix A

Appendix A. Cyclic peptide reported in the Violaceae family. Sequences retrieved from Cybase [38].

Name	Sequence	Species
cycloviolacin H1	GIPCGESCVYIPCLTSAIGCSCKSKVCYRN	*Viola hederacea*
cycloviolacin H2	SAIACGESCVYIPCFIPGCSCRNRVCYLN	*Viola hederacea*
cycloviolacin H3	GLPVCGETCFGGTCNTPGCICDPWPVCTRN	*Viola hederacea*
cycloviolacin H4	GIPCAESCVWIPCTVTALLGCSCSNNVCYN	*Viola hederacea*
CyO1	GIPCAESCVYIPCTVTALLGCSCSNRVCYN	*Viola odorata*, *Viola uliginosa*
CyO10	GIPCGESCVYIPCLTSAVGCSCKSKVCYRN	*Viola odorata*
CyO11	GTLPCGESCVWIPCISAVVGCSCKSKVCYKN	*Viola odorata*,*Viola tricolor*, *Viola arvensis*, *Viola baoshanensis*, *Viola yedoensis*, *Viola tianshanica* *Viola abyssinica*, *Viola philippica*
CyO12	GLPICGETCVGGTCNTPGCSCSWPVCTRN	*Viola arcuata*
CyO13	GIPCGESCVWIPCISAAIGCSCKSKVCYRN	*Viola uliginosa*
CyO14	GSIPACGESCFKGKCYTPGCSCSKYPLCAKN	*Viola odorata*
CyO15	GLVPCGETCFTGKCYTPGCSCSYPICKKN	*Viola odorata*
CyO16	GLPCGETCFTGKCYTPGCSCSYPICKKIN	*Viola odorata*
CyO17	GIPCGESCVWIPCISAAIGCSCKNKVCYRN	*Viola odorata*
CyO18	GIPCGESCVYIPCTVTALAGCKCKSKVCYN	*Viola odorata*
CyO19	GTLPCGESCVWIPCISSVVGCSCKSKVCYKD	*Viola odorata*
CyO2	GIPCGESCVWIPCISSAIGCSCKSKVCYRN	*Viola odorata*, *Hybanthus enneaspermus*, *Viola baoshanensis*, *Viola biflora*, *Viola philippica*,

		Viola uliginosa, Viola arcuate, Viola austrosinensis
CyO20	GIPCGESCVWIPCLTSAIGCSCKSKVCYRD	*Viola odorata*
CyO21	GLPVCGETCVTGSCYTPGCTCSWPVCTRN	*Viola odorata*
CyO22	GLPICGETCVGGTCNTPGCTCSWPVCTRN	*Viola tricolor, Viola arvensis, Viola baoshanensis. Viola yedoensis, Viola tianshanica, Viola abyssinica, Viola philippica, Viola arcuata*
CyO23	GLPTCGETCFGGTCNTPGCTCDSSWPICTHN	*Viola odorata*
CyO24	GLPTCGETCFGGTCNTPGCTCDPWPVCTHN	*Viola odorata*
CyO25	DIFCGETCAFIPCITHVPGTCSCKSKVCYFN	*Viola odorata*
CyO26	GSIPACGESCFRGKCYTPGCSCSKYPLCAKD	*Viola odorata*
CyO27	GSIPACGESCFKGWCYTPGCSCSKYPLCAKD	*Viola odorata*
CyO28	GLPVCGETCVGGTCNTPGCSCSWPVCFRD	*Viola odorata,Viola tricolor*
CyO29	GIPCGESCVWIPCISGAIGCSCKSKVCYKN	*Viola odorata*
CyO3	GIPCGESCVWIPCLTSAIGCSCKSKVCYRN	*Viola uliginosa*
CyO30	GIPCGESCVWIPCISSAIGCSCKNKVCFKN	*Viola odorata*
CyO31	GLPVCGETCVGGTCNTPGCSCSIPVCTRN	*Viola odorata*
CyO32	GAPVCGETCFGGTCNTPGCTCDPWPVCTND	*Viola odorata*
CyO33	GLPVCGETCVGGTCNTPYCTCSWPVCTRD	*Viola odorata*
CyO34	GLPVCGETCVGGTCNTEYCTCSWPVCTRD	*Viola odorata*
CyO35	GLPVCGETCVGGTCNTPYCFCSWPVCTRD	*Viola odorata*
CyO36	GLPTCGETCFGGTCNTPGCTCDPFPVCTHD	*Viola odorata*
CyO4	GIPCGESCVWIPCISSAIGCSCKNKVCYRN	*Viola odorata, Viola tricolor, Viola baoshanensis, Pombalia calceolaria*
CyO5	GTPCGESCVWIPCISSAVGCSCKNKVCYKN	*Viola odorata*
CyO6	GTLPCGESCVWIPCISAAVGCSCKSKVCYKN	*Viola odorata*

CyO7	SIPCGESCVWIPCTITALAGCKCKSKVCYN	*Viola odorata*
CyO8	GTLPCGESCVWIPCISSVVGCSCKSKVCYKN	*Viola odorata, Viola baoshanensis, Viola adunca, Viola uliginosa*
CyO9	GIPCGESCVWIPCLTSAVGCSCKSKVCYRN	*Viola odorata, Viola biflora*
cycloviolacin T1	GIPVCGETCVGGTCNTPGCSCSWPVCTRN	*Viola tianshanica*
cycloviolacin Y1	GGTIFDCGETCFLGTCYTPGCSCGNYGFCYGTN	*Viola yedoensis*
cycloviolacin Y2	GGTIFDCGESCFLGTCYTAGCSCGNWGLCYGTN	*Viola yedoensis*
cycloviolacin Y3	GGTIFDCGETCFLGTCYTAGCSCGNWGLCYGTN	*Viola yedoensis*
cycloviolacin Y4	GVPCGESCVFIPCITGVIGCSCSSNVCYLN	*Viola yedoensis*
cycloviolacin Y5	GIPCAESCVWIPCTVTALVGCSCSDKVCYN	*Viola yedoensis*
cycloviolin A	GVIPCGESCVFIPCISAAIGCSCKNKVCYRN	*Leonia cymosa*
cycloviolin B	GTACGESCYVLPCFTVGCTCTSSQCFKN	*Leonia cymosa*
cycloviolin C	GIPCGESCVFIPCLTTVAGCSCKNKVCYRN	*Leonia cymosa*
cycloviolin D	GFPCGESCVFIPCISAAIGCSCKNKVCYRN	*Leonia cymosa, Viola uliginosa*
cyI1	GTFPCGESCVYIPCISSVVGCSCKSKVCYKN	*Viola inconspicua*
cyI2	GTFPCGESCVWIPCISSVVGCSCKSKVCYKN	*Viola inconspicua*
cyI3	GNPGACGETCIWGKCYSASIGCSCSKYKVCTLN	*Viola inconspicua*
cyI4	GNPGACGETCVWGKCYSASIGCSCNKYKVCTLN	*Viola inconspicua*
cyI5	GNPGACGETCIWGKCYSASIGCSCSRYKVCTLN	*Viola inconspicua*
cyI6	GNPGACGETCIWGKCYSAKIGCSCSKYKICTLN	*Viola inconspicua*
Globa A	GIPCGESCVFIPCITAAIGCSCKTKVCYRN	*Gloeospermum blakeanum*
Globa B	GVIPCGESCVFIPCISAVLGCSCKSKVCYRN	*Gloeospermum blakeanum*
Globa C	APCGESCVYIPCLLTAPIGCSCSNIVCYRN	*Gloeospermum blakeanum*
Globa D	GIPCGETCVFMPCISGPMGCSCKHMVCYRN	*Gloeospermum blakeanum*
Globa E	GSAFGCGETCVKGKCNTPGCVCSWPVCKKN	*Gloeospermum blakeanum*
Globa F	GSFPCGESCVFIPCISAIAGCSCKNKVCYKN	*Gloeospermum blakeanum*
Glopa A	GGSIPCIETCVWTGCFLVPGCSCKSDKKCYLN	*Gloeospermum blakeanum*

Glopa B	GGSVPCIETCVWTGCFLVPGCSCKSDKKCYLN	*Gloeospermum blakeanum*
Glopa C	GDIPLCGETCFEGGNCRIPGCTCVWPFCSKN	*Gloeospermum blakeanum*
Glopa D	GVPCGESCVWVPCTVTALMGCSCVREVCRKD	*Gloeospermum blakeanum*
Glopa E	GIPCAESCVWIPCTVTKMLGCSCKDKVCYN	*Gloeospermum blakeanum*
Glopa F	GRLPCGESCVFLPCLSVSLGCSCKNKVCYRN	*Gloeospermum pauciflorum pauciflorum*
Glopa G	GRLPCGESCVFLPCLSAVLGCSCKNKVCYRN	*Gloeospermum pauciflorum pauciflorum*
Hobo A	GLPTCGETCTLGTCNTPGCTCSWPLCTKN	*Melicytus obovatus*
Hyde B	GVLPCGESCVFDRTCHLAGCGCGSTVPLCVRN	*Hybanthus denticulatus*
Hyfl A	SISCGESCVYIPCTVTALVGCTCKDKVCYLN	*Hybanthus floribundus E*
Hyfl B	GSPIQCAETCFIGKCYTEELGCTCTAFLCMKN	*Hybanthus floribundus E*
Hyfl C	GSPRQCAETCFIGKCYTEELGCTCTAFLCMKN	*Hybanthus floribundus E*
Hyfl D	GSVPCGESCVYIPCFTGIAGCSCKSKVCYYN	*Hybanthus floribundus E*
Hyfl E	GEIPCGESCVYLPCFLPNCYCRNHVCYLN	*Hybanthus floribundus E*
Hyfl F	SISCGETCTTFNCWIPNCKCNHHDKVCYWN	*Hybanthus floribundus E*
Hyfl I	GIPCGESCVFIPCISGVIGCSCKSKVCYRN	*Hybanthus floribundus E*
Hyfl J	GIACGESCAYFGCWIPGCSCRNKVCYFN	*Hybanthus floribundus E*
Hyfl K	GTPCGESCVYIPCFTAVVGCTCKDKVCYLN	*Hybanthus floribundus E*
Hyfl L	GTPCAESCVYLPCFTGVIGCTCKDKVCYLN	*Hybanthus floribundus E*
Hyfl M	GNIPCGESCIFFPCFNPGCSCKDNLCYYN	*Hybanthus floribundus E*
Hyla-br1	GVIPCGESCVFIPCISSFLGCSCKNKVCYRN	*Pombalia lanata*
Hypa A	GIPCAESCVYIPCTITALLGCSCKNKVCYN	*Hybanthus parviflorus*
Ltri A	GVACGESCVYLPCFTVGCTCTSSQCFKN	*Leonia triandra*
Mang A	GFPTCGETCTLGTCNTPGCTCSWPICTRD	*Melicytus angustifolius*
Mden A	GIPTCGETCTLGTCNTPGCTCSWPICTKN	*Melicytus dentatus*
Mden B	GLPICGETCFTGKCYTPGCTCSYPICKKN	*Melicytus dentatus*
Mden C	GKPICGETCFKGKCYTPGCTCSYPVCKKN	*Melicytus dentatus*
Mden E	GIPCGESCVYIPCITAAIGCSCKSKVCYRN	*Melicytus dentatus*

Mden F	GLPICGETCFFGKCNTPKCTCINPICYKN	*Melicytus dentatus*
Mden G	GIPCAESCVYIPCITAALGCSCKNKVCYRN	*Melicytus dentatus*
Mden H	GIPICGETCFFGKCNTPKCTCNKPLCYKN	*Melicytus dentatus*
Mden I	GIPCGESCVYIPCITTAIGCSCKNKVCYRN	*Melicytus dentatus*
Mden J	GSIPCGESCVYIPCISSIVGCACKSKVCYKN	*Melicytus dentatus*
Mden K	GSIPCGESCVWIPCISSVVGCACKNKVCYKN	*Melicytus dentatus*
Mden L	GSIPCGESCVYIPISAVLGCSCKNKVCYRN	*Melicytus dentatus*
Mden M	GTIPCGESCVYIPCITSALGCSCKKKVCYKN	*Melicytus dentatus*
Mden N	GTIPCGESCVYIPCLTSALGCSCKNKVCYRN	*Melicytus dentatus*
mech 1	GVIPCGESCVFIPCINKKKCSCKNKVCYRD	*Melicytus chathamicus*
mech 2	GLPTCGETCTLGKCNTPKCTCNWPICYKD	*Melicytus chathamicus*
mech 3	GLPTCGETCTLGKCNTPKCTCNWPICYKN	*Melicytus chathamicus*
mech 4	GSIPCGESCVYIPCISSLLGCSCKSKVCYKD	*Melicytus chathamicus*
mech 5	GVIPCGESCVFIPCISSVVGCTCKNKVCYRD	*Melicytus chathamicus*
mech 6	GVIPCGESCVFIPCISSVVGCTCKNKVCYRN	*Melicytus chathamicus*
mech 7	GIPICGETCTIGTCNTPGCTCSWPVCTRD	*Melicytus chathamicus*
mela 1	GKYTCGETCFKGKCYTPGCTCSYPICKKD	*Melicytus latifolius*
mela 2	GKPTCGETCFKGKCYTPGCTCSYPLCKKD	*Melicytus latifolius*
mela 3	GKPICGETCFKGKCYTPGCTCSYPICKKD	*Melicytus latifolius*
mela 4	GKPICGETCFKGKCYTPGCTCSYPICKKN	*Melicytus latifolius*
mela 5	GSAIACGESCFKFKCYTPGCSCSYPICKKD	*Melicytus latifolius*
mela 6	GIPTCGETCFKGKCYTPGCSCSYPICKKD	*Melicytus latifolius*
mela 7	GLPTCGETCFKGKCYTPGCSCSYPICKKN	*Melicytus latifolius*
Mema A	GLPCAESCVWLPCTVTALLGCSCKDKVCYRN	*Melicytus macrophyllus*
Mema B	GTVPCGESCVWLPCLTGLVGCSCKNNVCYTN	*Melicytus macrophyllus*
Mobo A	GFPTCGETCTLGTCNTPGCTCSWPICTRN	*Melicytus obovatus*
Mobo B	GKPICGETCAKGKCYTPKCTCNWPICYKN	*Melicytus obovatus*

Mra1	GIPCAESCVYIPCLTSIGCSCKSKVCYRN	*Melicytus ramiflorus*
Mra13	GIPCGESCVYLPCFTTIIGCKCQGKVCYH	*Melicytus ramiflorus*
Mra14a	GSIPCGESCVFIPCISSVVGCSCKNKVCYKN	*Melicytus ramiflorus*
Mra14b	GTIPCGESCVFIPCLTSAIGCSCKSKVCYKN	*Melicytus ramiflorus*
Mra17a	GSIPCGESCVYIPCISSLLGCSCESKVCYKN	*Melicytus ramiflorus*
Mra2	GIPCAESCVYIPCLTSAGCSCKSKVCYRN	*Melicytus ramiflorus*
Mra22	GVPCGESCVWIPCLTSIVGCSCKNNVCTLNS	*Melicytus ramiflorus*
Mra23	GVIPCGESCVFIPCISSVLGCSCKNKVCYRN	*Melicytus ramiflorus*
Mra24	GHPTCGETCLLGTCYTPGCTCKRPVCYKN	*Melicytus ramiflorus*
Mra25	GSAILCGESCTLGECYTPGCTCSWPICTKN	*Melicytus ramiflorus*
Mra26	GHPICGETCVGNKCYTPGCTCTWPVCYRN	*Melicytus ramiflorus*
Mra29	GSIPCGESCVFIPCISSIVGCSCKSKVCYKN	*Melicytus ramiflorus*
Mra3	GSIPCGESCVYIPCITSIVGCSCKSKVCYKN	*Melicytus ramiflorus*
Mra30	GIPCGESCVFIPCLTSAIGCSCKSKVCYRN	*Viola tricolor, Hybanthus enneaspermus, Viola baoshanensis, Melicytus ramiflorus, Viola philippica, Viola uliginosa, Viola arcuata*
Mra30a	GSIPCGEGCVFIPCISSIVGCSCKSKVCYKN	*Melicytus ramiflorus*
Mra4	GSIPCGESCVYIPCISSLLGCSCKSKVCYKN	*Melicytus ramiflorus*
Mra5	GIPCAESCVYIPCLTSAIGCSCKSKVCYRN	*Melicytus ramiflorus*
NorA	GVIPCGESCVFIPCISTLIGCSCKNKVCYRN	*Noisettia orchidiflora*
Oak6 cyclotide 1	GLPVCGETCFGGTCNTPGCACDPWPVCTRN	*Oldenlandia affinis, Viola tricolor*
Orto A	GLPCGESCVYLPCLLTAPLGCSCKNKVCYRN	*Orthion oblanceolatum*
Poca A	GLPCAESCVFIPCTITAILGCSCRDRVCYD	*Pombalia calceolaria*
Poca B	GIPCAESCVFIPCVTAILGCSCKDRVCYN	*Pombalia calceolaria*
Rigra A	GVPCGESCVWLPCTVTALLGCKCETRGCTLN	*Rinorea gracilipes*
Rili A	GLPCAESCVWLPCTVTALLGCTCVDRVCFLD	*Rinorea lindeniana*
Rili B	GLPVCGETCAGGTCNTPGCSCTWPLCTRN	*Rinorea lindeniana*

Rivi1	GLPICGETCVFGKCNTPGCSCRRPICYKN	*Rinorea virgata*
Rivi2	GSYLCGETCVQGKCYTPGCTCSWPICKKN	*Rinorea virgata*
Rivi3	GLPICGETCLLGKCYTPGCSCRRPVCYKN	*Rinorea virgata*
Rivi4	GKPICGETCLLGKCYTPGCTCGKRVLCYKN	*Rinorea virgata*
Rivi5	GLPICGETCTLGTCNTPGCTCSWPVCFRN	*Rinorea virgata*
tricyclon A	GGTIFDCGESCFLGTCYTKGCSCGEWKLCYGTN	*Viola arvensis, Viola tricolor*
tricyclon B	GGTIFDCGESCFLGTCYTKGCSCGEWKLCYGEN	*Viola tricolor*
vaby A	GLPVCGETCAGGTCNTPGCSCSWPICTRN	*Viola abyssinica*
vaby B	GLPVCGETCAGGTCNTPGCSCTWPICTRN	*Viola abyssinica*
vaby C	GLPVCGETCAGGRCNTPGCSCSWPVCTRN	*Viola abyssinica, Viola tricolor*
vaby D	GLPVCGETCFGGTCNTPGCTCDPWPVCTRN	*Viola abyssinica*
vaby E	GLPVCGETCFGGTCNTPGCSCDPWPVCTRN	*Viola abyssinica*
varv peptide B	GLPVCGETCFGGTCNTPGCSCDPWPMCSRN	*Viola arvensis, Viola tricolor*
varv peptide C	GVPICGETCVGGTCNTPGCSCSWPVCTRN	*Viola arvensis, Viola tricolor*
varv peptide D	GLPICGETCVGGSCNTPGCSCSWPVCTRN	*Viola arvensis, Viola tricolor*
varv peptide F	GVPICGETCTLGTCYTAGCSCSWPVCTRN	*Viola arvensis, Viola tricolor*
varv peptide G	GVPVCGETCFGGTCNTPGCSCDPWPVCSRN	*Viola arvensis, Viola tricolor*
varv peptide H	GLPVCGETCFGGTCNTPGCSCETWPVCSRN	*Viola arvensis, Viola tricolor*
Vdif A	GIPCGESCVFIPCISSVVGCSCKSKVCYRN	*Viola diffusa*
vhl-1	SISCGESCAMISFCFTEVIGCSCKNKVCYLN	*Viola hederacea*
vhl-2	GLPVCGETCFTGTCYTNGCTCDPWPVCTRN	*Viola hederacea*
vhr1	GIPCAESCVWIPCTVTALLGCSCSNKVCYN	*Viola hederacea*
Viba 1	GIPCGEGCVYLPCFTAPLGCSCSSKVCYRN	*Viola baoshanensis*
Viba 10	GIPCAESCVYLPCVTIVIGCSCKDKVCYN	*Viola baoshanensis*
Viba 11	GIPCGESCVWIPCISGAIGCSCKSKVCYRN	*Viola baoshanensis, Viola philippica, Viola tricolor*
Viba 12	GIPCAESCVWIPCTVTALLGCSCKDKVCYN	*Viola baoshanensis*

Viba 13	TIPCAESCVWIPCTVTALLGCSCKDKVCYN	*Viola baoshanensis*
Viba 14	GRLCGERCVIERTRAWCRTVGCICSLHTLECVRN	*Viola baoshanensis*
Viba 15	GLPVCGETCVGGTCNTPGCACSWPVCTRN	*Viola baoshanensis, Viola philippica, Viola tricolor*
Viba 17	GLPVCGETCVGGTCNTPGCGCSWPVCTRN	*Viola baoshanensis, Viola philippica*
Viba 2	GIPCGESCVYLPCFTAPLGCSCSSKVCYRN	*Viola baoshanensis*
Viba 3	GIPCGESCVWIPCLTAAIGCSCSSKVCYRN	*Viola baoshanensis*
viba 30 linear	GPPVCGETCVGGTCNTPGCSCSWPVCTRN	*Viola tricolor*
viba 32	GLPVCGEACVGGTCNTPGCSCSWPVCTRN	*Viola tricolor*
Viba 4	GVPCGESCVWIPCLTSAIGCSCKSSVCYRN	*Viola baoshanensis*
Viba 5	GIPCGESCVWIPCLTATIGCSCKSKVCYRN	*Viola baoshanensis*
Viba 6	GIPCGESCVLIPCISSVIGCSCKSKVCYRN	*Viola baoshanensis*
Viba 7	GVIPCGESCVFIPCISSVIGCSCKSKVCYRN	*Viola baoshanensis*
Viba 8	GAGCIETCYTFPCISEMINCSCKNSRCQKN	*Viola baoshanensis*
Viba 9	GIPCGESCVWIPCISSAIGCSCKNKVCYRK	*Viola tricolor, Viola baoshanensis*
Viba18	GIHCAETCLWGTCRTAYIGCSCENKICYKN	*Viola baoshanensis*
Viba19	GLPVCGETCFGGTCNTPGCSCEWPVCTRN	*Viola baoshanensis*
Viba20	PISCGETCKFSRCFTSIFGCKCINKVCHT	*Viola baoshanensis*
Viba21	GLFCGEYCVQFPCRSPGCICLRGLCVRN	*Viola baoshanensis*
Viba22	GIPCGESCVFIPCISSVIGCSCKSKVCYRN	*Viola baoshanensis*
Viba23	GIPCGESCVWIPCFSAAIGCSCSSKVCYRN	*Viola baoshanensis*
Viba24	GKIPCGESCVWIPCITTVVGCSCSNKVCYKN	*Viola baoshanensis*
Viba25	GRVPCGESCVYIPCFTSIAGCSCSDKVCWHN	*Viola baoshanensis*
Viba26	GFPCGESCVYVPCLTAAIGCSCKNKVCYKN	*Viola baoshanensis*
Viba27	GVQCGPTCRFGPVNCGINPRCNCRNTRCVWD	*Viola baoshanensis*
Viba28	GIHCAETCIWGTCRTAIIGCSCENRICYKN	*Viola baoshanensis*
Viba29	GLPVCGETCVGGTCSTPGCGCSWPVCTRN	*Viola baoshanensis*

Viba30	GPPVCGETCVGGTCNTPGCSCSWPVCTRN	*Viola baoshanensis*
Viba31	GLPVCGETCVGGACNTPGCACSWPVCTRN	*Viola baoshanensis*
Viba32	GLPVCGEACVGGTCNTPGCSCSWPVCTRN	*Viola baoshanensis*
Viba33	GLPVCGETCFGGTCNTPGCPCEWPVCTRN	*Viola baoshanensis*
Viba34	GSIPSCGESCFKGKCYTPGCSCSKYPLCAKN	*Viola baoshanensis*
Viba35	GQDFCGGQICFASDCIVVGCECDAWSRCVPS	*Viola baoshanensis*
Viba36	GSYYSCGETCRKTKCYTPDCICAWPGLCGKN	*Viola baoshanensis*
Viba37	GIPCAESCVYLPCVTIVIGCSCKDEVCYNS	*Viola baoshanensis*
Viba38	GSIPSCGESCFKGKCYTPGCSCSKYPLCAKK	*Viola baoshanensis*
Viba39	GSIPSCGESCFKGKCYTPGCSCSKYPLCAYE	*Viola baoshanensis*
Viba40	GIPCGESCVWIPCITAAIGCSCSSKVCYRN	*Viola baoshanensis*
Viba41	GSIPSCGESCFKGKCYTPVCSCSKYPLCAKN	*Viola baoshanensis*
Viba42	GASCVETCNYFPCISEMINCYCQSKRCVKN	*Viola baoshanensis*
Viba43	GSLPCGESCVFIPCISSVIGCACKSKVCYKN	*Viola baoshanensis*
Viba44	GIHCAETCFWGTCRTAYIGCSCENRICYKN	*Viola baoshanensis*
vibi A	GLPVCGETCFGGTCNTPGCSCSYPICTRN	*Viola biflora*
vibi B	GLPVCGETCFGGTCNTPGCTCSYPICTRN	*Viola biflora, Palicourea tetragona*
vibi C	GLPVCGETCAFGSCYTPGCSCSWPVCTRN	*Viola biflora, Viola tricolor*
vibi D	GLPVCGETCFGGRCNTPGCTCSYPICTRN	*Viola biflora*
vibi E	GIPCAESCVWIPCTVTALIGCGCSNKVCYN	*Viola biflora*
vibi F	GTIPCGESCVFIPCLTSALGCSCKSKVCYKN	*Viola biflora*
vibi G	GTFPCGESCVFIPCLTSAIGCSCKSKVCYKN	*Psychotria leptothyrsa, Viola biflora, Viola tricolor*
vibi H	GLLPCAESCVYIPCLTTVIGCSCKSKVCYKN	*Viola biflora, Psychotria leptothyrsa*
vibi I	GIPCGESCVWIPCLTSTVGCSCKSKVCYRN	*Viola biflora*
vibi J	GTFPCGESCVWIPCISKVIGCACKSKVCYKN	*Viola biflora*
vibi K	GIPCGESCVWIPCLTSAVGCPCKSKVCYRN	*Viola biflora*

vico A	GSIPCAESCVYIPCFTGIAGCSCKNKVCYYN	*Viola cotyledon*
vico B	GSIPCAESCVYIPCITGIAGCSCKNKVCYYN	*Viola cotyledon*
Vide A	GLPCGESCVFLPCLTSALGCSCKSKVCYRN	*Viola decumbens*
vigno 1	GLPLCGETCAGGTCNTPGCSCSWPVCVRN	*Viola ignobilis*
vigno 10	GTIPCGESCVWIPCISSVVGCSCKSKVCYKD	*Viola ignobilis, Viola tricolor*
vigno 2	GSSPLCGETCAGGTCNTPGCSCSWPVCVRD	*Viola ignobilis*
vigno 3	GLPLCGETCVGGTCNTPGCSCSWPVCTRN	*Viola ignobilis, Viola tricolor*
vigno 4	GLPLCGETCVGGTCNTPACSCSWPVCTRN	*Viola ignobilis, Viola tricolor*
vigno 5	GLPLCGETCVGGTCNTPGCSCGWPVCVRN	*Viola ignobilis, Viola tricolor*
vigno 6	GIPCGESCVWIPCISSAIGCSCKGSKVCYRN	*Viola ignobilis, Viola tricolor*
vigno 7	GTLPCGESCVWIPCISSVVGCSCKNKVCYKN	*Viola ignobilis, Viola tricolor*
vigno 8	GIPCGESCVWIPCITSAVGCSCKSKVCYRN	*Viola ignobilis*
vigno 9	GIPCGESCVWIPCISSALGCSCKSKVCYRN	*Viola ignobilis, Viola tricolor*
vila A	GIPCGESCVWIPCISSAIGCSCKDKVCYRD	*Viola labridorica*
vila B	GIPCGESCVWIPCISSAIGCSCRSKVCYRD	*Viola labridorica*
vila C	GTLPCGESCVWIPCISSVVGCSCKDKVCYKD	*Viola labridorica*
Vinc A	GIPVCGETCTLGTCYTAGCSCSWPVCTRN	*Viola inconspicua*
Vinc B	GSIPACGESCFKGKCYTPGCTCSKYPLCAKN	*Viola inconspicua*
Vini A	GSVPCGESCVWLPCLSGLAGCSCKNKVCYYD	*Viola nivalis*
violacin A	SAISCGETCFKFKCYTPRCSCSYPVCK	*Viola odorata, Psychotria leptothyrsa*
violapeptide 1	GLPVCGETCVGGTCNTPGCSCSRPVCTXN	*Viola arvensis*
viphi A	GSIPCGESCVFIPCISSVIGCACKSKVCYKN	*viola philippica*
viphi B	GLPVCGETCTIGTCYTAGCTCSWPICTRN	*viola philippica*
viphi C	GVPCGESCVYIPCITSVIGCSCSSKVCYIN	*viola philippica*
viphi D	GIPCGESCVFIPCISSVIGCSCSSKVCYRN	*viola philippica*
viphi E	GSIPCGESCVFIPCISAVIGCSCSNKVCYKN	*viola philippica*
viphi F	GSIPCGESCVFIPCISAIIGCSCSSKVCYKN	*viola philippica*

viphi G	GSIPCGESCVFIPCISAIIGCSCSNKVCYKN	*viola philippica*
viphi H	GIPCAESCVWIPCTVTAIVGCSCSWGVCYN	*viola philippica*
Visu 1	GLPVCGETCVGGTCNTPGCTCTWPVCTRN	*Viola sumatrana*
Visu 2	GIPCAESCVYIPCTITALLGCSCKSKVCYN	*Viola sumatrana*
vitri A	GIPCGESCVWIPCITSAIGCSCKSKVCYRN	*Psychotria leptothyrsa, Viola biflora, Viola tricolor*
vitri B	GYPICGESCVGGTCNTPGCSCSWPVCTTN	*viola tricolor*
vitri C	GLPICGETCVGGTCNTPGCFCTWPVCTRN	*viola tricolor*
vitri D	GLPVCGETCFTGSCYTPGCSCNWPVCNRN	*Viola tricolor*
vitri E	GLPVCGETCVGGTCNTPGCSCSWPVCFRN	*Viola odorata, Viola tricolor*
vitri F	GTLPCGESCVWIPCISSVVGCACKSKVCYKD	*Viola tricolor*
vitri peptide 1	GLIPCGESCVWIPCISSVIGCSCKSKVCYKN	*Viola tricolor*
vitri peptide 14	GSSCGETCEVFSCFITRCACIDGLCYRN	*Viola tricolor*
vitri peptide 18a	GVPICGETCFQGTCNTPGCTCKWPICERN	*Viola tricolor*
vitri peptide 18b	GSVFNCGETCVFGTCFTSGCSCVYRVCSKD	*Viola tricolor*
vitri peptide 2	GSIPCGESCVWIPCISGIAGCSCSNKVCYLN	*Viola tricolor*
vitri peptide 20	GDLVPCGESCVYIPCLTTVLGCSCSENVCYRN	*Viola tricolor*
vitri peptide 21	GGPLDCQETCTLSDRCYTKGCTCNWPICYKN	*Viola tricolor*
vitri peptide 22a	GAPVCGETCFTGLCYSSGCSCIYPVCNRN	*Viola tricolor*
vitri peptide 23	GLPTCGETCTLGTCYTPGCTCSWPLCTKN	*Viola tricolor*
vitri peptide 24/28	GEPVCGDSCVFFGCDDEGCTCGPWSLCYRN	*Viola tricolor*
vitri peptide 24a	GGTIFNCGESCFQGTCYTKGCACGDWKLCYGEN	*Viola tricolor*
vitri peptide 27a	GAFTPCGETCLTGECHTEGCSCVGQTFCVKK	*Viola tricolor*
vitri peptide 29	GVPSSDCLETCFGGKCNAHRCTCSQWPLCAKN	*Viola tricolor*
vitri peptide 3	GSWPCGESCVYIPCITSIAGCECSKNVCYKN	*Viola tricolor*
vitri peptide 30	GFACGETCIFTSCFITGCTCNSSLCFRN	*Viola tricolor*
vitri peptide 36/37	GGTIFSCGESCFQGTCYTKGCACGDWKLCYGEN	*Viola tricolor*

vitri peptide 38	GDTCYETCFTGFCFIGGCKCDFPVCVKN	*Viola tricolor*
vitri peptide 39	GAPICGESCFTGTCYTVQCSCSWPVCTRN	*Viola tricolor*
vitri peptide 4	GTPCGESCIYVPCISAVFGCWCQSKVCYKD	*Viola tricolor*
vitri peptide 50	GDIPCGESCVYIPCITGVLGCSCSHNVCYYN	*Viola tricolor*
vitri peptide 8	PTPCGETCIWISCVTAAIGCYCHESICYR	*Viola tricolor*
vitri peptide 94b	GVAVCGETCTLGTCYTPGCSCDWPICKRN	*Viola tricolor*
vitri peptide 9a/53	GTIFDCGETCLLGKCYTPGCSCGSWALCYGQN	*Viola tricolor*
ViulA	GIPCGESCVWIPCISSLIGCSCRGKVCYH	*Viola uliginosa*
ViulB	GVPCGESCVWIPCLTGAIGCSCSNKVCYLN	*Viola uliginosa*
ViulC	GTFCGETCVMFPCFSSARGCGCHNLGCELN	*Viola uliginosa*
ViulD	GIPCGESCVWIPCLTSAIGCSCKSKVCYKN	*Viola uliginosa*
ViulE	GGHCGESCMLLPCFTARIGCSCSRSICYKN	*Viola uliginosa*
ViulF	GRFCGEICSRGFCSNPRCTCNASRQCVRN	*Viola uliginosa*
ViulG	GRAVCGETCFAGICYTPVCVCGKWDLCRMN	*Viola uliginosa*
ViulH	SELPCGESCVFIPCITSIAGCSCSHKVCYLN	*Viola uliginosa*
ViulI	GPTPCGETCIWISCVTAVMGCSCKNSICYMN	*Viola uliginosa*
ViulJ	GEVTCNGETCFTGKCNAKGCNCKNWPLCTRN	*Viola uliginosa*
ViulK	SIFCSETCRTFPCFTKAVGCSCVSKRCYKN	*Viola uliginosa*
ViulL	GIPCAETCLWRPCRTAIMGCSCEYNFCYKN	*Viola uliginosa*
vocC	GLPVCGETCVGGTCNTPGCSCSWPVCIRN	*Viola odorata*
Vodo I1	GVFCGEACAQASCSIAGCECIAGLCYKN	*Viola odorata*
Vodo I2	GVFCGELCIKASCSIPGCECIAGLCYKN	*Viola odorata*
Vodo I3	GVFCGEPCIKASCSIPGCECIAGLCYKN	*Viola odorata*
vodo M	GAPICGESCFTGKCYTVQCSCSWPVCTRN	*Viola odorata, Viola tricolor*
vodo N	GLPVCGETCTLGKCYTAGCSCSWPVCYRN	*Viola odorata, Viola tricolor*
Vpf-1	GIPCGESCVFIPCLTAAIGCSCRSKVCYRN	*Viola pinetorum*
Vpl-1	GSQSCGESCVLIPCISGVIGCSCSSMICYFN	*Viola pinetorum*

Vpub A	GVIPCGESCVFIPCISAVIGCSCKSKVCYRN	*Viola pubescens*
Vpub B	GIIPCGESCVFIPCITSVVGCSCKSKVCYKN	*Viola pubescens*
Vpub C	GIIPCGESCVFIPCITSIVGCSCKSKVCYKN	*Viola pubescens*

Appendix B

Appendix B. Sequence homology chart of PALs and AEPs. Sequence similarity (%) of the enzymes was calculated by EMBL-EBI EMBOSS Water Pairwise Sequence Alignment (available online at https://www.ebi.ac.uk/Tools/psa/emboss_water/) with default setting. See section 3.1.3 for the accession codes used.

Enzyme	AtLEGβ	AtLEGγ	Butelase-1	Butelase-2	Jack bean AEP	Huamn legumain	Mouse legumain	OaAEP1b	VcAEP	VyPAL2
AtLEGβ	100.0									
AtLEGγ	73.5	100.0								
Butelase-1	70.2	80.7	100.0							
Butelase-2	79.1	70.9	72.0	100.0						
Jack bean AEP	80.5	72.4	71.3	90.7	100.0					
Huamn legumain	55.2	54.5	51.3	56.0	55.1	100.0				
Mouse legumain	56.1	55.8	51.4	56.1	57.6	91.3	100.0			
OaAEP1b	68.5	78.8	80.0	70.1	69.6	54.0	53.8	100.0		
VcAEP	65.6	75.3	74.9	67.8	65.7	50.8	53.1	74.8	100.0	
VyPAL2	66.9	78.1	77.9	68.0	67.5	50.2	51.1	79.4	78.5	100.0

References

1. Rawlings, ND, Barrett, AJ, Thomas, PD, Huang, X, Bateman, A, and Finn, RD, *The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database.* Nucleic Acids Research, 2018. **46**(D1): p. D624-D632.
2. Gerovac, M and Tampé, R, *Control of mRNA Translation by Versatile ATP-Driven Machines.* Trends in Biochemical Sciences, 2019. **44**(2): p. 167-180.
3. Arnison, PG, Bibb, MJ, Bierbaum, G, Bowers, AA, Bugni, TS, Bulaj, G, Camarero, JA, Campopiano, DJ, Challis, GL, Clardy, J, Cotter, PD, Craik, DJ, Dawson, M, Dittmann, E, Donadio, S, Dorrestein, PC, Entian, K-D, Fischbach, MA, Garavelli, JS, Göransson, U, Gruber, CW, Haft, DH, Hemscheidt, TK, Hertweck, C, Hill, C, Horswill, AR, Jaspars, M, Kelly, WL, Klinman, JP, Kuipers, OP, Link, AJ, Liu, W, Marahiel, MA, Mitchell, DA, Moll, GN, Moore, BS, Müller, R, Nair, SK, Nes, IF, Norris, GE, Olivera, BM, Onaka, H, Patchett, ML, Piel, J, Reaney, MJT, Rebuffat, S, Ross, RP, Sahl, H-G, Schmidt, EW, Selsted, ME, Severinov, K, Shen, B, Sivonen, K, Smith, L, Stein, T, Süssmuth, RD, Tagg, JR, Tang, G-L, Truman, AW, Vederas, JC, Walsh, CT, Walton, JD, Wenzel, SC, Willey, JM, and van der Donk, WA, *Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature.* Natural Product Reports, 2013. **30**(1): p. 108-160.
4. Montalbán-López, M, Scott, TA, Ramesh, S, Rahman, IR, van Heel, AJ, Viel, JH, Bandarian, V, Dittmann, E, Genilloud, O, and Goto, Y, *New developments in RiPP discovery, enzymology and engineering.* Natural Product Reports, 2020.
5. Barber, CJ, Pujara, PT, Reed, DW, Chiwocha, S, Zhang, H, and Covello, PS, *The two-step biosynthesis of cyclic peptides from linear precursors in a member of the plant family Caryophyllaceae involves cyclization by a serine protease-like enzyme.* Journal of Biological Chemistry, 2013. **288**(18): p. 12500-12510.
6. Nguyen, GKT, Wang, S, Qiu, Y, Hemu, X, Lian, Y, and Tam, JP, *Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis.* Nature Chemical Biology, 2014. **10**: p. 732.
7. Harris, KS, Durek, T, Kaas, Q, Poth, AG, Gilding, EK, Conlan, BF, Saska, I, Daly, NL, Van Der Weerden, NL, and Craik, DJ, *Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase.* Nature Communications, 2015. **6**: p. 10199.
8. Jackson, MA, Gilding, EK, Shafee, T, Harris, KS, Kaas, Q, Poon, S, Yap, K, Jia, H, Guarino, R, Chan, LY, Durek, T, Anderson, MA, and Craik, DJ, *Molecular basis for the production of cyclic peptides by plant asparaginyl endopeptidases.* Nature Communications, 2018. **9**(1): p. 2411.
9. Harris, KS, Guarino, RF, Dissanayake, RS, Quimbar, P, McCorkelle, OC, Poon, S, Kaas, Q, Durek, T, Gilding, EK, Jackson, MA, Craik, DJ, van der Weerden, NL, Anders, RF, and Anderson, MA, *A suite of kinetically superior AEP ligases can cyclise an intrinsically disordered protein.* Scientific Reports, 2019. **9**(1): p. 10820-10820.
10. Hemu, X, El Sahili, A, Hu, S, Wong, K, Chen, Y, Wong, YH, Zhang, X, Serra, A, Goh, BC, Darwis, DA, Chen, MW, Sze, SK, Liu, C-F, Lescar, J, and Tam, JP, *Structural determinants for peptide-bond formation by asparaginyl ligases.* Proceedings of the National Academy of Sciences, 2019. **116**(24): p. 11737.

11. Luo, H, Hong, SY, Sgambelluri, RM, Angelos, E, Li, X, and Walton, JD, *Peptide macrocyclization catalyzed by a prolyl oligopeptidase involved in α-amanitin biosynthesis.* Chemistry and Biology, 2014. **21**(12): p. 1610-1617.
12. Mazmanian, SK, Liu, G, Ton-That, H, and Schneewind, O, *Staphylococcus aureus sortase, an enzyme that anchors surface proteins to the cell wall.* Science, 1999. **285**(5428): p. 760-763.
13. Lee, J, McIntosh, J, Hathaway, BJ, and Schmidt, EW, *Using Marine Natural Products to Discover a Protease that Catalyzes Peptide Macrocyclization of Diverse Substrates.* Journal of the American Chemical Society, 2009. **131**(6): p. 2122-2124.
14. Hirata, R, Ohsumk, Y, Nakano, A, Kawasaki, H, Suzuki, K, and Anraku, Y, *Molecular structure of a gene, VMA1, encoding the catalytic subunit of H (+)-translocating adenosine triphosphatase from vacuolar membranes of Saccharomyces cerevisiae.* Journal of Biological Chemistry, 1990. **265**(12): p. 6726-6733.
15. Kane, PM, Yamashiro, CT, Wolczyk, DF, Neff, N, Goebl, M, and Stevens, TH, *Protein splicing converts the yeast TFP1 gene product to the 69-kD subunit of the vacuolar H(+)-adenosine triphosphatase.* Science, 1990. **250**(4981): p. 651.
16. Kembhavi, AA, Buttle, DJ, Knight, CG, and Barrett, AJ, *The Two Cysteine Endopeptidases of Legume Seeds: Purification and Characterization by Use of Specific Fluorometric Assays.* Archives of Biochemistry and Biophysics, 1993. **303**(2): p. 208-213.
17. Hara-Nishimura, I and Nishimura, M, *Proglobulin Processing Enzyme in Vacuoles Isolated from Developing Pumpkin Cotyledons.* Plant Physiology, 1987. **85**(2): p. 440.
18. Hara-Nishimura, I, Inoue, K, and Nishimura, M, *A unique vacuolar processing enzyme responsible for conversion of several proprotein precursors into the mature forms.* FEBS Letters, 1991. **294**(1-2): p. 89-93.
19. Webb, EC, *Enzyme nomenclature : 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. 1992, London [etc.]: Academic Press.
20. Kembhavi, AA, Buttle, DJ, Knight, CG, and Barrett, AJ, *The two cysteine endopeptidases of legume seeds: purification and characterization by use of specific fluorometric assays.* Arch Biochem Biophys, 1993. **303**(2): p. 208-13.
21. Tam, JP, Chan, N-Y, Liew, HT, Tan, SJ, and Chen, Y, *Peptide asparaginyl ligases—renegade peptide bond makers.* Science China Chemistry, 2020. **63**: p. 296-307.
22. Hara-Nishimura, I, Takeuchi, Y, and Nishimura, M, *Molecular characterization of a vacuolar processing enzyme related to a putative cysteine proteinase of Schistosoma mansoni.* The Plant Cell, 1993. **5**(11): p. 1651-1659.
23. Chen, J-M, Rawlings, ND, Stevens, RA, and Barrett, AJ, *Identification of the active site of legumain links it to caspases, clostripain and gingipains in a new clan of cysteine endopeptidases.* FEBS Letters, 1998. **441**(3): p. 361-365.
24. Shutov, A, Do, NL, and Vaintraub, I, *Purification and partial characterization of protease B from germinating vetch seeds.* Biokhimiia (Moscow, Russia), 1982. **47**(5): p. 814-821.
25. Csoma, C and Polgár, L, *Proteinase from germinating bean cotyledons. Evidence for involvement of a thiol group in catalysis.* Biochemical Journal, 1984. **222**(3): p. 769-776.
26. Hara-Nishimura, I, Nishimura, M, and Akazawa, T, *Biosynthesis and intracellular transport of 11S globulin in developing pumpkin cotyledons.* Plant Physiology, 1985. **77**(3): p. 747-752.

27. Klinkert, M-Q, Felleisen, R, Link, G, Ruppel, A, and Beck, E, *Primary structures of Sm31/32 diagnostic proteins of Schistosoma mansoni and their identification as proteases.* Molecular and Biochemical Parasitology, 1989. **33**(2): p. 113-122.
28. Scott, MP, Jung, R, Muntz, K, and Nielsen, NC, *A protease responsible for post-translational cleavage of a conserved Asn-Gly linkage in glycinin, the major seed storage protein of soybean.* Proceedings of the National Academy of Sciences, 1992. **89**(2): p. 658-662.
29. Hiraiwa, N, Takeuchi, Y, Nishimura, M, and Hara-Nishimura, I, *A vacuolar processing enzyme in maturing and germinating seeds: its distribution and associated changes during development.* Plant and Cell Physiology, 1993. **34**(8): p. 1197-1204.
30. Shimada, T, Hiraiwa, N, Nishimura, M, and Hara-Nishimura, I, *Vacuolar processing enzyme of soybean that converts proproteins to the corresponding mature forms.* Plant and Cell Physiology, 1994. **35**(4): p. 713-718.
31. Hara-Nishimura, I, Shimada, T, Hiraiwa, N, and Nishimura, M, *Vacuolar processing enzyme responsible for maturation of seed proteins.* Journal of Plant Physiology, 1995. **145**(5-6): p. 632-640.
32. Shimada, T, Yamada, K, Kataoka, M, Nakaune, S, Koumoto, Y, Kuroyanagi, M, Tabata, S, Kato, T, Shinozaki, K, and Seki, M, *Vacuolar processing enzymes are essential for proper processing of seed storage proteins in Arabidopsis thaliana.* Journal of Biological Chemistry, 2003. **278**(34): p. 32292-32299.
33. Gruis, DF, Selinger, DA, Curran, JM, and Jung, R, *Redundant proteolytic mechanisms process seed storage proteins in the absence of seed-type members of the vacuolar processing enzyme family of cysteine proteases.* The Plant Cell, 2002. **14**(11): p. 2863-2882.
34. Gruis, D, Schulze, J, and Jung, R, *Storage protein accumulation in the absence of the vacuolar processing enzyme family of cysteine proteases.* The Plant Cell, 2004. **16**(1): p. 270-290.
35. Ishii, S, *Asparaginylendopeptidase: an enzyme probably responsible to post-translational proteolysis and transpeptidation of proconcanavalin A.* Seikagaku, 1993. **65**(3): p. 185-9.
36. Schechter, I and Berger, A, *On the size of the active site in proteases. I. Papain.* Biochemical and Biophysical Research Communications, 1967. **27**(2): p. 157-162.
37. Kaufmann, H and Tobschirbel, A, *An oligopeptide from flaxseed.* Chemische Berichte, 1959. **92**: p. 2805-2809.
38. Wang, CK, Kaas, Q, Chiche, L, and Craik, DJ, *CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering.* Nucleic acids research, 2007. **36**(suppl_1): p. D206-D210.
39. Tan, N-H and Zhou, J, *Plant Cyclopeptides.* Chemical Reviews, 2006. **106**(3): p. 840-895.
40. Morita, H, Yun, YS, Takeya, K, and Itokawa, H, *Segetalin A, a new cyclic hexapeptide from vaccaria segetalis.* Tetrahedron Letters, 1994. **35**(51): p. 9593-9596.
41. Condie, JA, Nowak, G, Reed, DW, Balsevich, JJ, Reaney, MJT, Arnison, PG, and Covello, PS, *The biosynthesis of Caryophyllaceae-like cyclic peptides in Saponaria vaccaria L. from DNA-encoded precursors.* The Plant Journal, 2011. **67**(4): p. 682-690.
42. Ludewig, H, Czekster, CM, Oueis, E, Munday, ES, Arshad, M, Synowsky, SA, Bent, AF, and Naismith, JH, *Characterization of the Fast and Promiscuous Macrocyclase from Plant PCY1 Enables the Use of Simple Substrates.* ACS Chemical Biology, 2018. **13**(3): p. 801-811.

43. Chekan, JR, Estrada, P, Covello, PS, and Nair, SK, *Characterization of the macrocyclase involved in the biosynthesis of RiPP cyclic peptides in plants.* Proceedings of the National Academy of Sciences, 2017. **114**(25): p. 6551-6556.
44. Lane, AL and Moore, BS, *A sea of biosynthesis: marine natural products meet the molecular age.* Natural Product Reports, 2011. **28**(2): p. 411-428.
45. Schmidt, EW, Nelson, JT, Rasko, DA, Sudek, S, Eisen, JA, Haygood, MG, and Ravel, J, *Patellamide A and C biosynthesis by a microcin-like pathway in Prochloron didemni, the cyanobacterial symbiont of Lissoclinum patella.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(20): p. 7315-7320.
46. Koehnke, J, Bent, A, Houssen, WE, Zollman, D, Morawitz, F, Shirran, S, Vendome, J, Nneoyiegbe, AF, Trembleau, L, Botting, CH, Smith, MCM, Jaspars, M, and Naismith, JH, *The mechanism of patellamide macrocyclization revealed by the characterization of the PatG macrocyclase domain.* Nature Structural & Molecular Biology, 2012. **19**: p. 767.
47. Montalbán-López, M, Scott, TA, Ramesh, S, Rahman, IR, van Heel, AJ, Viel, JH, Bandarian, V, Dittmann, E, Genilloud, O, Goto, Y, Grande Burgos, MJ, Hill, C, Kim, S, Koehnke, J, Latham, JA, Link, AJ, Martínez, B, Nair, SK, Nicolet, Y, Rebuffat, S, Sahl, H-G, Sareen, D, Schmidt, EW, Schmitt, L, Severinov, K, Süssmuth, RD, Truman, AW, Wang, H, Weng, J-K, van Wezel, GP, Zhang, Q, Zhong, J, Piel, J, Mitchell, DA, Kuipers, OP, and van der Donk, WA, *New developments in RiPP discovery, enzymology and engineering.* Natural Product Reports, 2021. **38**(1): p. 130-239.
48. Luo, H, Hallen-Adams, HE, Scott-Craig, JS, and Walton, JD, *Ribosomal biosynthesis of α-amanitin in Galerina marginata.* Fungal Genetics and Biology, 2012. **49**(2): p. 123-129.
49. Bushnell, DA, Cramer, P, and Kornberg, RD, *Structural basis of transcription: α-amanitin–RNA polymerase II cocrystal at 2.8 Å resolution.* Proceedings of the National Academy of Sciences, 2002. **99**(3): p. 1218-1222.
50. Sgambelluri, RM, Smith, MO, and Walton, JD, *Versatility of Prolyl Oligopeptidase B in Peptide Macrocyclization.* ACS Synthetic Biology, 2018. **7**(1): p. 145-152.
51. Marraffini, LA, DeDent, AC, and Schneewind, O, *Sortases and the Art of Anchoring Proteins to the Envelopes of Gram-Positive Bacteria.* Microbiology and Molecular Biology Reviews, 2006. **70**(1): p. 192-221.
52. Schneewind, O, Mihaylova-Petkov, D, and Model, P, *Cell wall sorting signals in surface proteins of gram-positive bacteria.* The EMBO Journal, 1993. **12**(12): p. 4803-4811.
53. Navarre, WW and Schneewind, O, *Proteolytic cleavage and cell wall anchoring at the LPXTG motif of surface proteins in Gram-positive bacteria.* Molecular Microbiology, 1994. **14**(1): p. 115-121.
54. Schneewind, O, Fowler, A, and Faull, KF, *Structure of the cell wall anchor of surface proteins in Staphylococcus aureus.* Science, 1995. **268**(5207): p. 103.
55. Ton-That, H, Liu, G, Mazmanian, SK, Faull, KF, and Schneewind, O, *Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif.* Proceedings of the National Academy of Sciences, 1999. **96**(22): p. 12424-12429.
56. Pishesha, N, Ingram, JR, and Ploegh, HL, *Sortase A: A Model for Transpeptidation and Its Biological Applications.* Annual Review of Cell and Developmental Biology, 2018. **34**(1): p. 163-188.
57. Guimaraes, CP, Witte, MD, Theile, CS, Bozkurt, G, Kundrat, L, Blom, AEM, and Ploegh, HL, *Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions.* Nature Protocols, 2013. **8**: p. 1787.

58. Shah, NH and Muir, TW, *Inteins: nature's gift to protein chemists.* Chemical Science, 2014. **5**(2): p. 446-461.
59. Muralidharan, V and Muir, TW, *Protein ligation: an enabling technology for the biophysical analysis of proteins.* Nature methods, 2006. **3**(6): p. 429.
60. Kang, HJ, Coulibaly, F, Clow, F, Proft, T, and Baker, EN, *Stabilizing Isopeptide Bonds Revealed in Gram-Positive Bacterial Pilus Structure.* Science, 2007. **318**(5856): p. 1625-1628.
61. Zakeri, B and Howarth, M, *Spontaneous Intermolecular Amide Bond Formation between Side Chains for Irreversible Peptide Targeting.* Journal of the American Chemical Society, 2010. **132**(13): p. 4526-4527.
62. Zakeri, B, Fierer, JO, Celik, E, Chittock, EC, Schwarz-Linek, U, Moy, VT, and Howarth, M, *Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin.* Proceedings of the National Academy of Sciences, 2012. **109**(12): p. E690-E697.
63. Fierer, JO, Veggiani, G, and Howarth, M, *SpyLigase peptide–peptide ligation polymerizes affibodies to enhance magnetic cancer cell capture.* Proceedings of the National Academy of Sciences, 2014. **111**(13): p. E1176-E1181.
64. Siegmund, V, Piater, B, Zakeri, B, Eichhorn, T, Fischer, F, Deutsch, C, Becker, S, Toleikis, L, Hock, B, Betz, UAK, and Kolmar, H, *Spontaneous Isopeptide Bond Formation as a Powerful Tool for Engineering Site-Specific Antibody-Drug Conjugates.* Scientific Reports, 2016. **6**: p. 39291.
65. Buldun, CM, Jean, JX, Bedford, MR, and Howarth, M, *SnoopLigase Catalyzes Peptide-Peptide Locking and Enables Solid-Phase Conjugate Isolation.* Journal of the American Chemical Society, 2018. **140**(8): p. 3008-3018.
66. Buldun, CM, Khairil Anuar, INA, and Howarth, M, *SnoopLigase-Mediated Peptide–PeptidePeptidesConjugation and Purification*, in *Polypeptide Materials: Methods and Protocols*, M.G. Ryadnov, Editor. 2021, Springer US: New York, NY. p. 13-31.
67. Banerjee, A and Howarth, M, *Nanoteamwork: covalent protein assembly beyond duets towards protein ensembles and orchestras.* Current Opinion in Biotechnology, 2018. **51**: p. 16-23.
68. Weeks, AM and Wells, JA, *Subtiligase-Catalyzed Peptide Ligation.* Chemical Reviews, 2020. **120**(6): p. 3127-3160.
69. Neet, KE and Koshland, DE, Jr., *The conversion of serine at the active site of subtilisin to cysteine: a "chemical mutation".* Proceedings of the National Academy of Sciences of the United States of America, 1966. **56**(5): p. 1606-1611.
70. Polgar, L and Bender, ML, *The reactivity of thiol-subtilisin, an enzyme containing a synthetic functional group.* Biochemistry, 1967. **6**(2): p. 610-620.
71. Wu, ZP and Hilvert, D, *Conversion of a Protease into an Acyl Transferase: Selenolsubtilisin.* Journal of the American Chemical Society, 1989. **111**(12): p. 4513-4514.
72. Abrahmsen, L, Tom, J, Burnier, J, Butcher, KA, Kossiakoff, A, and Wells, JA, *Engineering subtilisin and its substrates for efficient ligation of peptide bonds in aqueous solution.* Biochemistry, 1991. **30**(17): p. 4151-4159.
73. Chang, TK, Jackson, DY, Burnier, JP, and Wells, JA, *Subtiligase: a tool for semisynthesis of proteins.* Proceedings of the National Academy of Sciences, 1994. **91**(26): p. 12544-12548.
74. Toplak, A, Nuijens, T, Quaedflieg, PJLM, Wu, B, and Janssen, DB, *Peptiligase, an Enzyme for Efficient Chemoenzymatic Peptide Synthesis and Cyclization in Water.* Advanced Synthesis & Catalysis, 2016. **358**(13): p. 2140-2147.
75. Tan, X, Yang, R, and Liu, C-F, *Facilitating Subtiligase-Catalyzed Peptide Ligation Reactions by Using Peptide Thioester Substrates.* Organic letters, 2018. **20**(21): p. 6691-6694.

76. Rühlmann, A, Kukla, D, Schwager, P, Bartels, K, and Huber, R, *Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor: Crystal structure determination and stereochemistry of the contact region.* Journal of Molecular Biology, 1973. **77**(3): p. 417-436.
77. Graf, L, Craik, CS, Patthy, A, Roczniak, S, Fletterick, RJ, and Rutter, WJ, *Selective alteration of substrate specificity by replacement of aspartic acid-189 with lysine in the binding pocket of trypsin.* Biochemistry, 1987. **26**(9): p. 2616-23.
78. Evnin, LB, Vásquez, JR, and Craik, CS, *Substrate specificity of trypsin investigated by using a genetic selection.* Proceedings of the National Academy of Sciences, 1990. **87**(17): p. 6659-6663.
79. Willett, WS, Gillmor, SA, Perona, JJ, Fletterick, RJ, and Craik, CS, *Engineered Metal Regulation of Trypsin Specificity.* Biochemistry, 1995. **34**(7): p. 2172-2180.
80. Kurth, T, Grahn, S, Thormann, M, Ullmann, D, Hofmann, HJ, Jakubke, HD, and Hedstrom, L, *Engineering the S1' subsite of trypsin: design of a protease which cleaves between dibasic residues.* Biochemistry, 1998. **37**(33): p. 11434-11440.
81. Liebscher, S, Schöpfel, M, Aumüller, T, Sharkhuukhen, A, Pech, A, Höss, E, Parthier, C, Jahreis, G, Stubbs, MT, and Bordusa, F, *N-terminal protein modification by substrate-activated reverse proteolysis.* Angewandte Chemie International Edition, 2014. **53**(11): p. 3024-3028.
82. Kinoshita, T, Nishimura, M, and Hara-Nishimura, I, *Homologues of a vacuolar processing enzyme that are expressed in different organs in Arabidopsis thaliana.* Plant Molecular Biology, 1995. **29**(1): p. 81-89.
83. Kinoshita, T, Nishimura, M, and Hara-Nishimura, I, *The sequence and expression of the γ-VPE gene, one member of a family of three genes for vacuolar processing enzymes in Arabidopsis thaliana.* Plant and Cell Physiology, 1995. **36**(8): p. 1555-1562.
84. Hara-Nishimura, I, Kinoshita, T, Hiraiwa, N, and Nishimura, M, *Vacuolar processing enzymes in protein-storage vacuoles and lytic vacuoles.* Journal of plant physiology, 1998. **152**(6): p. 668-674.
85. Kinoshita, T, Yamada, K, Hiraiwa, N, Kondo, M, Nishimura, M, and Hara-Nishimura, I, *Vacuolar processing enzyme is up-regulated in the lytic vacuoles of vegetative tissues during senescence and under various stressed conditions.* The Plant Journal, 1999. **19**(1): p. 43-53.
86. Chen, J-M, Dando, PM, Rawlings, ND, Brown, MA, Young, NE, Stevens, RA, Hewitt, E, Watts, C, and Barrett, AJ, *Cloning, Isolation, and Characterization of Mammalian Legumain, an Asparaginyl Endopeptidase*.* Journal of Biological Chemistry, 1997. **272**(12): p. 8090-8098.
87. Lunde, NN, Bosnjak, T, Solberg, R, and Johansen, HT, *Mammalian legumain - A lysosomal cysteine protease with extracellular functions?* Biochimie, 2019. **166**: p. 77-83.
88. Caffrey, CR, McKerrow, JH, Salter, JP, and Sajid, M, *Blood 'n' guts: an update on schistosome digestive peptidases.* Trends in Parasitology, 2004. **20**(5): p. 241-248.
89. Ravipati, AS, Poth, AG, Troeira Henriques, Sn, Bhandari, M, Huang, Y-H, Nino, J, Colgrave, ML, and Craik, DJ, *Understanding the diversity and distribution of cyclotides from plants of varied genetic origin.* Journal of Natural Products, 2017. **80**(5): p. 1522-1530.
90. Craik, DJ, *Host-defense activities of cyclotides.* Toxins (Basel), 2012. **4**(2): p. 139-56.
91. Yamada, K, Basak, AK, Goto-Yamada, S, Tarnawska-Glatt, K, and Hara-Nishimura, I, *Vacuolar processing enzymes in the plant life cycle.* New Phytologist, 2020. **226**(1): p. 21-31.

92. Nonis, SG, Haywood, J, and Mylne, JS, *Plant asparaginyl endopeptidases and their structural determinants of function.* Biochemical Society Transactions, 2021. **49**(2): p. 965-976.
93. Nakaune, S, Yamada, K, Kondo, M, Kato, T, Tabata, S, Nishimura, M, and Hara-Nishimura, I, *A Vacuolar Processing Enzyme, δVPE, Is Involved in Seed Coat Formation at the Early Stage of Seed Development.* The Plant Cell, 2005. **17**(3): p. 876-887.
94. Hatsugai, N, Yamada, K, Goto-Yamada, S, and Hara-Nishimura, I, *Vacuolar processing enzyme in plant programmed cell death.* Frontiers in Plant Science, 2015. **6**: p. 234.
95. Hatsugai, N, Kuroyanagi, M, Yamada, K, Meshi, T, Tsuda, S, Kondo, M, Nishimura, M, and Hara-Nishimura, I, *A plant vacuolar protease, VPE, mediates virus-induced hypersensitive cell death.* Science, 2004. **305**(5685): p. 855-858.
96. Min, W and Jones, DH, *In vitro splicing of concanavalin A is catalyzed by asparaginyl endopeptidase.* Nature structural biology, 1994. **1**(8): p. 502.
97. Cunningham, BA, Hemperly, JJ, Hopp, TP, and Edelman, GM, *Favin versus concanavalin A: Circularly permuted amino acid sequences.* Proceedings of the National Academy of Sciences, 1979. **76**(7): p. 3218-3222.
98. Bernath-Levin, K, Nelson, C, Elliott, AG, Jayasena, AS, Millar, AH, Craik, DJ, and Mylne, JS, *Peptide macrocyclization by a bifunctional endoprotease.* Chemistry & Biology, 2015. **22**(5): p. 571-582.
99. Du, J, Yap, K, Chan, LY, Rehm, FBH, Looi, FY, Poth, AG, Gilding, EK, Kaas, Q, Durek, T, and Craik, DJ, *A bifunctional asparaginyl endopeptidase efficiently catalyzes both cleavage and cyclization of cyclic trypsin inhibitors.* Nature Communications, 2020. **11**(1): p. 1575.
100. Jennings, C, West, J, Waine, C, Craik, D, and Anderson, M, *Biosynthesis and insecticidal properties of plant cyclotides: the cyclic knotted proteins from Oldenlandia affinis.* Proceedings of the National Academy of Sciences, 2001. **98**(19): p. 10614-10619.
101. Lis, H and Sharon, N, *Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition.* Chemical Reviews, 1998. **98**(2): p. 637-674.
102. Carrington, D, Auffret, A, and Hanke, D, *Polypeptide ligation occurs during post-translational modification of concanavalin A.* Nature, 1985. **313**(5997): p. 64.
103. Bowles, DJ, Marcus, SE, Pappin, D, Findlay, J, Eliopoulos, E, Maycox, PR, and Burgess, J, *Posttranslational processing of concanavalin A precursors in jackbean cotyledons.* The Journal of Cell Biology, 1986. **102**(4): p. 1284-1297.
104. Abe, Y, Shirane, K, Yokosawa, H, Matsushita, H, Mitta, M, Kato, I, and Ishii, S, *Asparaginyl endopeptidase of jack bean seeds. Purification, characterization, and high utility in protein sequence analysis.* Journal of Biological Chemistry, 1993. **268**(5): p. 3525-3529.
105. Chen, J-M, Dando, PM, Rawlings, ND, Brown, MA, Young, NE, Stevens, RA, Hewitt, E, Watts, C, and Barrett, AJ, *Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase.* Journal of Biological Chemistry, 1997. **272**(12): p. 8090-8098.
106. Watts, C, *The endosome-lysosome pathway and information generation in the immune system.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2012. **1824**(1): p. 14-21.
107. Repnik, U, Stoka, V, Turk, V, and Turk, B, *Lysosomes and lysosomal cathepsins in cell death.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2012. **1824**(1): p. 22-33.
108. Manoury, B, Hewitt, EW, Morrice, N, Dando, PM, Barrett, AJ, and Watts, C, *An asparaginyl endopeptidase processes a microbial antigen for class II MHC presentation.* Nature, 1998. **396**(6712): p. 695-699.

109. Dall, E and Brandstetter, H, *Structure and function of legumain in health and disease.* Biochimie, 2016. **122**: p. 126-150.
110. Liu, C, Sun, CZ, Huang, HN, Janda, K, and Edgington, T, *Overexpression of legumain in tumors is significant for invasion/metastasis and a candidate enzymatic target for prodrug therapy.* Cancer Research, 2003. **63**(11): p. 2957-2964.
111. Ruppel, A, Shi, YE, Wei, DX, and Diesfeld, HJ, *Sera of Schistosoma japonicum-infected patients cross-react with diagnostic 31/32 kD proteins of S. mansoni.* Clinical & Experimental Immunology, 1987. **69**(2): p. 291-8.
112. Horn, M, Nussbaumerová, M, Šanda, M, Kovářová, Z, Srba, J, Franta, Z, Sojka, D, Bogyo, M, Caffrey, CR, Kopáček, P, and Mareš, M, *Hemoglobin Digestion in Blood-Feeding Ticks: Mapping a Multipeptidase Pathway by Functional Proteomics.* Chemistry & Biology, 2009. **16**(10): p. 1053-1063.
113. Sojka, D, Hajdušek, O, Dvořák, J, Sajid, M, Franta, Z, Schneider, EL, Craik, CS, Vancová, M, Burešová, V, Bogyo, M, Sexton, KB, McKerrow, JH, Caffrey, CR, and Kopáček, P, *IrAE – An asparaginyl endopeptidase (legumain) in the gut of the hard tick Ixodes ricinus.* International Journal for Parasitology, 2007. **37**(7): p. 713-724.
114. Alim, MA, Tsuji, N, Miyoshi, T, Islam, MK, Huang, X, Hatta, T, and Fujisaki, K, *HlLgm2, a member of asparaginyl endopeptidases/legumains in the midgut of the ixodid tick Haemaphysalis longicornis, is involved in blood-meal digestion.* Journal of Insect Physiology, 2008. **54**(3): p. 573-585.
115. Sajid, M, McKerrow, JH, Hansell, E, Mathieu, MA, Lucas, KD, Hsieh, I, Greenbaum, D, Bogyo, M, Salter, JP, and Lim, KC, *Functional expression and characterization of Schistosoma mansoni cathepsin B and its trans-activation by an endogenous asparaginyl endopeptidase.* Molecular and Biochemical Parasitology, 2003. **131**(1): p. 65-75.
116. Sojka, D, Franta, Z, Horn, M, Caffrey, CR, Mareš, M, and Kopáček, P, *New insights into the machinery of blood digestion by ticks.* Trends in Parasitology, 2013. **29**(6): p. 276-285.
117. Zhang, Y, *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**(1): p. 1-8.
118. Roy, A, Kucukural, A, and Zhang, Y, *I-TASSER: a unified platform for automated protein structure and function prediction.* Nature Protocols, 2010. **5**(4): p. 725-738.
119. Yang, J, Yan, R, Roy, A, Xu, D, Poisson, J, and Zhang, Y, *The I-TASSER Suite: protein structure and function prediction.* Nature Methods, 2015. **12**(1): p. 7-8.
120. Perutz, MF, Muirhead, H, Cox, JM, and Goaman, LCG, *Three-dimensional Fourier Synthesis of Horse Oxyhaemoglobin at 2.8 Å Resolution: The Atomic Model.* Nature, 1968. **219**(5150): p. 131-139.
121. Jílková, A, Řezáčová, P, Lepšík, M, Horn, M, Váchová, J, Fanfrlík, J, Brynda, J, McKerrow, JH, Caffrey, CR, and Mareš, M, *Structural basis for inhibition of cathepsin B drug target from the human blood fluke, Schistosoma mansoni.* Journal of Biological Chemistry, 2011. **286**(41): p. 35770-35781.
122. Dall, E and Brandstetter, H, *Activation of legumain involves proteolytic and conformational events, resulting in a context-and substrate-dependent activity profile.* Acta Crystallographica Section F: Structural Biology and Crystallization Communications, 2012. **68**(1): p. 24-31.
123. Dall, E and Brandstetter, H, *Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation.* Proceedings of the National Academy of Sciences, 2013. **110**(27): p. 10940-10945.
124. Zhao, L, Hua, T, Crowley, C, Ru, H, Ni, X, Shaw, N, Jiao, L, Ding, W, Qu, L, Hung, L-W, Huang, W, Liu, L, Ye, K, Ouyang, S, Cheng, G, and Liu, Z-J,

Structural analysis of asparaginyl endopeptidase reveals the activation mechanism and a reversible intermediate maturation stage. Cell Research, 2014. **24**(3): p. 344-358.
125. Dall, E, Fegg, JC, Briza, P, and Brandstetter, H, *Structure and Mechanism of an Aspartimide-Dependent Peptide Ligase in Human Legumain.* Angewandte Chemie International Edition, 2015. **54**(10): p. 2917-2921.
126. Yang, RL, Wong, YH, Nguyen, GKT, Tam, JP, Lescar, J, and Wu, B, *Engineering a Catalytically Efficient Recombinant Protein Ligase.* Journal of the American Chemical Society, 2017. **139**(15): p. 5351-5358.
127. Zauner, FB, Dall, E, Regl, C, Grassi, L, Huber, CG, Cabrele, C, and Brandstetter, H, *Crystal structure of plant legumain reveals a unique two-chain state with pH-dependent activity regulation.* The Plant Cell, 2018. **30**(3): p. 686-699.
128. Zauner, FB, Elsässer, B, Dall, E, Cabrele, C, and Brandstetter, H, *Structural analyses of Arabidopsis thaliana legumain γ reveal differential recognition and processing of proteolysis and ligation substrates.* Journal of Biological Chemistry, 2018. **293**(23): p. 8934-8946.
129. James, AM, Haywood, J, Leroux, J, Ignasiak, K, Elliott, AG, Schmidberger, JW, Fisher, MF, Nonis, SG, Fenske, R, and Bond, CS, *The macrocyclizing protease butelase 1 remains autocatalytic and reveals the structural basis for ligase activity.* The Plant Journal, 2019. **98**(6): p. 999.
130. Dall, E, Zauner, FB, Soh, WT, Demir, F, Dahms, SO, Cabrele, C, Huesgen, PF, and Brandstetter, H, *Structural and functional studies of Arabidopsis thaliana legumain beta reveal isoform specific mechanisms of activation and substrate recognition.* Journal of Biological Chemistry, 2020. **295**(37): p. 13047-13064.
131. Hemu, X, El Sahili, A, Hu, S, Zhang, X, Serra, A, Goh, BC, Darwis, DA, Chen, MW, Sze, SK, Liu, C-f, Lescar, J, and Tam, JP, *Turning an Asparaginyl Endopeptidase into a Peptide Ligase.* ACS Catalysis, 2020: p. 8825-8834.
132. James, AM, Haywood, J, Leroux, J, Ignasiak, K, Elliott, AG, Schmidberger, JW, Fisher, MF, Nonis, SG, Fenske, R, Bond, CS, and Mylne, JS, *The macrocyclizing protease butelase 1 remains autocatalytic and reveals the structural basis for ligase activity.* The Plant Journal, 2019. **98**(6): p. 988-999.
133. CHEN, J-M, FORTUNATO, M, and BARRETT, AJ, *Activation of human prolegumain by cleavage at a C-terminal asparagine residue.* Biochemical Journal, 2000. **352**(2): p. 327-334.
134. Li, DN, Matthews, SP, Antoniou, AN, Mazzeo, D, and Watts, C, *Multistep autoactivation of asparaginyl endopeptidase in vitro and in vivo.* Journal of Biological Chemistry, 2003. **278**(40): p. 38980-38990.
135. Hiraiwa, N, Nishimura, M, and Hara-Nishimura, I, *Vacuolar processing enzyme is self-catalytically activated by sequential removal of the C-terminal and N-terminal propeptides.* FEBS Letters, 1999. **447**(2-3): p. 213-216.
136. Wang, Z, Zhang, D, Hemu, X, Hu, S, To, J, Zhang, X, Lescar, J, Tam, JP, and Liu, C-F, *Engineering protein theranostics using bio-orthogonal asparaginyl peptide ligases.* Theranostics, 2021. **11**(12): p. 5863.
137. Rehm, FBH, Harmand, TJ, Yap, K, Durek, T, Craik, DJ, and Ploegh, HL, *Site-Specific Sequential Protein Labeling Catalyzed by a Single Recombinant Ligase.* Journal of the American Chemical Society, 2019. **141**(43): p. 17388-17393.
138. Sievers, F, Wilm, A, Dineen, D, Gibson, TJ, Karplus, K, Li, W, Lopez, R, McWilliam, H, Remmert, M, and Söding, J, *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Molecular Systems Biology, 2011. **7**(1): p. 539.

139. Livingstone, CD and Barton, GJ, *Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.* Bioinformatics, 1993. **9**(6): p. 745-756.
140. Waterhouse, AM, Procter, JB, Martin, DM, Clamp, M, and Barton, GJ, *Jalview Version 2—a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-1191.
141. Haywood, J, Schmidberger, JW, James, AM, Nonis, SG, Sukhoverkov, KV, Elias, M, Bond, CS, and Mylne, JS, *Structural basis of ribosomal peptide macrocyclization in plants.* eLife, 2018. **7**: p. e32955.
142. Merrifield, RB, *Solid phase peptide synthesis. I. The synthesis of a tetrapeptide.* Journal of the American Chemical Society, 1963. **85**(14): p. 2149-2154.
143. Boutureira, O and Bernardes, GJL, *Advances in Chemical Protein Modification.* Chemical Reviews, 2015. **115**(5): p. 2174-2195.
144. Hoyt, EA, Cal, PM, Oliveira, BL, and Bernardes, GJ, *Contemporary approaches to site-selective protein modification.* Nature Reviews Chemistry, 2019. **3**(3): p. 147-171.
145. White, CJ and Yudin, AK, *Contemporary strategies for peptide macrocyclization.* Nature Chemistry, 2011. **3**(7): p. 509.
146. Tapeinou, A, Matsoukas, MT, Simal, C, and Tselios, T, *Review cyclic peptides on a merry-go-round; towards drug design.* Peptide Science, 2015. **104**(5): p. 453-461.
147. Qian, Z, Dougherty, PG, and Pei, D, *Targeting intracellular protein–protein interactions with cell-permeable cyclic peptides.* Current Opinion in Chemical Biology, 2017. **38**: p. 80-86.
148. Pi, N, Gao, M, Cheng, X, Liu, H, Kuang, Z, Yang, Z, Yang, J, Zhang, B, Chen, Y, and Liu, S, *Recombinant butelase-mediated cyclization of the p53-binding domain of the oncoprotein MdmX stabilized protein conformation as a promising model for structural investigation.* Biochemistry, 2019.
149. Bi, X, Yin, J, Hemu, X, Rao, C, Tam, JP, and Liu, C-F, *Immobilization and Intracellular Delivery of Circular Proteins by Modifying a Genetically Incorporated Unnatural Amino Acid.* Bioconjugate Chemistry, 2018.
150. Nguyen, GK, Qiu, Y, Cao, Y, Hemu, X, Liu, CF, and Tam, JP, *Butelase-mediated cyclization and ligation of peptides and proteins.* Nature Protocols, 2016. **11**(10): p. 1977-1988.
151. Nguyen, GK, Hemu, X, Quek, JP, and Tam, JP, *Butelase-Mediated Macrocyclization of d-Amino-Acid-Containing Peptides.* Angewandte Chemie, 2016. **128**(41): p. 12994-12998.
152. Hemu, X, Qiu, Y, Nguyen, GK, and Tam, JP, *Total Synthesis of Circular Bacteriocins by Butelase 1.* Journal of the American Chemical Society, 2016. **138**(22): p. 6968-71.
153. Nguyen, GKT, Kam, A, Loo, S, Jansson, AE, Pan, LX, and Tam, JP, *Butelase 1: A Versatile Ligase for Peptide and Protein Macrocyclization.* Journal of the American Chemical Society, 2015. **137**(49): p. 15398-15401.
154. Nguyen, GK, Cao, Y, Wang, W, Liu, CF, and Tam, JP, *Site-Specific N-Terminal Labeling of Peptides and Proteins using Butelase 1 and Thiodepsipeptide.* Angewandte Chemie, 2015. **127**(52): p. 15920-15924.
155. Tam James, P, Nguyen Giang, KT, Kam, A, and Loo, S, Proceedings of the 35th European Peptide Symposium, 2018: p. 3-7.
156. Bi, X, Yin, J, Nguyen Giang, KT, Rao, C, Halim Nurashikin Bte, A, Hemu, X, Tam James, P, and Liu, CF, *Enzymatic Engineering of Live Bacterial Cell Surfaces Using Butelase 1.* Angewandte Chemie, 2017. **129**(27): p. 7930-7933.
157. Bi, X, Yin, J, Zhang, D, Zhang, X, Balamkundu, S, Lescar, J, Dedon, PC, Tam, JP, and Liu, C-F, *Tagging Transferrin Receptor with a Disulfide FRET*

Probe To Gauge the Redox State in Endosomal Compartments. Analytical Chemistry, 2020. **92**(18): p. 12460-12466.
158. Harmand, T, Bousbaine, D, Chan, AI, Zhang, X, Liu, DR, Tam, JP, and Ploegh, HL, *One-pot dual labeling of an IgG 1 and preparation of C-to-C fusion proteins through a combination of Sortase A and Butelase 1.* Bioconjugate Chemistry, 2018.
159. Cao, Y, Nguyen, GKT, Chuah, S, Tam, JP, and Liu, C-F, *Butelase-Mediated Ligation as an Efficient Bioconjugation Method for the Synthesis of Peptide Dendrimers.* Bioconjugate Chemistry, 2016. **27**(11): p. 2592-2596.
160. Cao, Y, Nguyen, GK, Tam, JP, and Liu, C-F, *Butelase-mediated synthesis of protein thioesters and its application for tandem chemoenzymatic ligation.* Chemical Communications, 2015. **51**(97): p. 17289-17292.
161. Aimoto, S, *Polypeptide synthesis by the thioester method.* Peptide Science, 1999. **51**(4): p. 247-265.
162. Tam, JP, Xu, J, and Eom, KD, *Methods and strategies of peptide ligation*.* Peptide Science, 2001. **60**(3): p. 194-205.
163. Liu, C-F and Tam, JP, *Peptide segment ligation strategy without use of protecting groups.* Proceedings of the National Academy of Sciences, 1994. **91**(14): p. 6584-6588.
164. Tam, JP, Lu, YA, Liu, CF, and Shao, J, *Peptide synthesis using unprotected peptides through orthogonal coupling methods.* Proceedings of the National Academy of Sciences, 1995. **92**(26): p. 12485.
165. Wu, Y, Batyuk, A, Honegger, A, Brandl, F, Mittl, PR, and Plückthun, A, *Rigidly connected multispecific artificial binders with adjustable geometries.* Scientific Reports, 2017. **7**(1): p. 1-11.
166. Amos, S-BTA, Vermeer, LS, Ferguson, PM, Kozlowska, J, Davy, M, Bui, TT, Drake, AF, Lorenz, CD, and Mason, AJ, *Antimicrobial Peptide Potency is Facilitated by Greater Conformational Flexibility when Binding to Gram-negative Bacterial Inner Membranes.* Scientific Reports, 2016. **6**(1): p. 37639.
167. Goldstein, JL, Anderson, RGW, and Brown, MS, *Coated pits, coated vesicles, and receptor-mediated endocytosis.* Nature, 1979. **279**(5715): p. 679-685.
168. Russell-Jones, GJ, *The potential use of receptor-mediated endocytosis for oral drug delivery.* Advanced Drug Delivery Reviews, 1996. **20**(1): p. 83-97.
169. Qian, ZM, Li, H, Sun, H, and Ho, K, *Targeted Drug Delivery via the Transferrin Receptor-Mediated Endocytosis Pathway.* Pharmacological Reviews, 2002. **54**(4): p. 561.
170. Tashima, T, *Effective cancer therapy based on selective drug delivery into cells across their membrane using receptor-mediated endocytosis.* Bioorganic & Medicinal Chemistry Letters, 2018. **28**(18): p. 3015-3024.
171. Posnett, DN, McGrath, H, and Tam, JP, *A novel method for producing anti-peptide antibodies. Production of site-specific antibodies to the T cell antigen receptor beta-chain.* Journal of Biological Chemistry, 1988. **263**(4): p. 1719-1725.
172. Tam, JP, *Synthetic peptide vaccine design: synthesis and properties of a high-density multiple antigenic peptide system.* Proceedings of the National Academy of Sciences, 1988. **85**(15): p. 5409-5413.
173. Sadler, K and Tam, JP, *Peptide dendrimers: applications and synthesis.* Reviews in Molecular Biotechnology, 2002. **90**(3): p. 195-229.
174. Tam, JP, Lu, Y-A, and Yang, J-L, *Antimicrobial dendrimeric peptides.* European Journal of Biochemistry, 2002. **269**(3): p. 923-932.
175. Hemu, X, Zhang, X, and Tam, JP, *Ligase-Controlled Cyclo-oligomerization of Peptides.* Organic Letters, 2019.
176. Matasci, N, Hung, L-H, Yan, Z, Carpenter, EJ, Wickett, NJ, Mirarab, S, Nguyen, N, Warnow, T, Ayyampalayam, S, and Barker, M, *Data access for the 1,000 Plants (1KP) project.* Gigascience, 2014. **3**(1): p. 2047-217X-3-17.

177. Gasteiger, E, Gattiker, A, Hoogland, C, Ivanyi, I, Appel, RD, and Bairoch, A, *ExPASy: the proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Research, 2003. **31**(13): p. 3784-3788.
178. Almagro Armenteros, JJ, Tsirigos, KD, Sønderby, CK, Petersen, TN, Winther, O, Brunak, S, von Heijne, G, and Nielsen, H, *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nature Biotechnology, 2019. **37**(4): p. 420-423.
179. Zvelebil, MJ, Barton, GJ, Taylor, WR, and Sternberg, MJE, *Prediction of protein secondary structure and active sites using the alignment of homologous sequences.* Journal of Molecular Biology, 1987. **195**(4): p. 957-961.
180. Lua, RC, Wilson, SJ, Konecki, DM, Wilkins, AD, Venner, E, Morgan, DH, and Lichtarge, O, *UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures.* Nucleic Acids Research, 2016. **44**(D1): p. D308-D312.
181. Madeira, F, Park, YM, Lee, J, Buso, N, Gur, T, Madhusoodanan, N, Basutkar, P, Tivey, AR, Potter, SC, and Finn, RD, *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Research, 2019. **47**(W1): p. W636-W641.
182. Ciccarelli, FD, Doerks, T, von Mering, C, Creevey, CJ, Snel, B, and Bork, P, *Toward automatic reconstruction of a highly resolved tree of life.* Science, 2006. **311**(5765): p. 1283-7.
183. Letunic, I and Bork, P, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* Nucleic Acids Research, 2021.
184. Suplatov, D, Sharapova, Y, Timonina, D, Kopylov, K, and Švedas, V, *The visualCMAT: A web-server to select and interpret correlated mutations/co-evolving residues in protein families.* Journal of Bioinformatics and Computational Biology, 2018. **16**(02): p. 1840005.
185. Schneider, TD and Stephens, RM, *Sequence logos: a new way to display consensus sequences.* Nucleic acids research, 1990. **18**(20): p. 6097-6100.
186. Crooks, GE, Hon, G, Chandonia, J-M, and Brenner, SE, *WebLogo: a sequence logo generator.* Genome Research, 2004. **14**(6): p. 1188-1190.
187. Do, CB, Mahabhashyam, MS, Brudno, M, and Batzoglou, S, *ProbCons: Probabilistic consistency-based multiple sequence alignment.* Genome Research, 2005. **15**(2): p. 330-340.
188. Serra, A, Hemu, X, Nguyen, GK, Nguyen, NT, Sze, SK, and Tam, JP, *A high-throughput peptidomic strategy to decipher the molecular diversity of cyclic cysteine-rich peptides.* Scientific reports, 2016. **6**(1): p. 1-13.
189. Nonis, SG, Haywood, J, Schmidberger, JW, Bond, CS, and Mylne, JS, *Structural basis for a natural circular permutation in proteins.* bioRxiv, 2020: p. 2020.10.28.360099.
190. Wilkins, A, Erdin, S, Lua, R, and Lichtarge, O, *Evolutionary trace for prediction and redesign of protein functional sites*, in *Computational Drug Discovery and Design*. 2012, Springer. p. 29-42.
191. Lichtarge, O, Bourne, HR, and Cohen, FE, *An evolutionary trace method defines binding surfaces common to protein families.* Journal of Molecular Biology, 1996. **257**(2): p. 342-358.
192. Lichtarge, O, Yamamoto, KR, and Cohen, FE, *Identification of functional surfaces of the zinc binding domains of intracellular receptors.* Journal of Molecular Biology, 1997. **274**(3): p. 325-337.
193. Liu, Y, Gierasch, LM, and Bahar, I, *Role of Hsp70 ATPase domain intrinsic dynamics and sequence evolution in enabling its functional interactions with NEFs.* PLoS Computational Biology, 2010. **6**(9): p. e1000931.
194. Liu, Y and Bahar, I, *Sequence evolution correlates with structural dynamics.* Molecular Biology and Evolution, 2012. **29**(9): p. 2253-2263.

195. Edgar, RC, *MUSCLE: a multiple sequence alignment method with reduced time and space complexity.* BMC Bioinformatics, 2004. **5**(1): p. 1-19.
196. Mihalek, I, Res, I, and Lichtarge, O, *A family of evolution-entropy hybrid methods for ranking protein residues by importance.* Journal of Molecular Biology, 2004. **336**(5): p. 1265-1282.
197. Luque, I and Freire, E, *Structural stability of binding sites: Consequences for binding affinity and allosteric effects.* Proteins: Structure, Function, and Bioinformatics, 2000. **41**(S4): p. 63-71.
198. Hara-Hishimura, I, Takeuchi, Y, Inoue, K, and Nishimura, M, *Vesicle transport and processing of the precursor to 2S albumin in pumpkin.* The Plant Journal, 1993. **4**(5): p. 793-800.
199. Yamada, K, Shimada, T, Nishimura, M, and Hara-Nishimura, I, *A VPE family supporting various vacuolar functions in plants.* Physiologia Plantarum, 2005. **123**(4): p. 369-375.
200. Gillon, AD, Saska, I, Jennings, CV, Guarino, RF, Craik, DJ, and Anderson, MA, *Biosynthesis of circular proteins in plants.* The Plant Journal, 2008. **53**(3): p. 505-515.
201. Tam, JP, Chan, N-Y, Liew, HT, Tan, SJ, and Chen, Y, *Peptide asparaginyl ligases—renegade peptide bond makers.* Science China Chemistry, 2020. **63**(3): p. 296-307.
202. Slazak, B, Jacobsson, E, Kuta, E, and Göransson, U, *Exogenous plant hormones and cyclotide expression in Viola uliginosa (Violaceae).* Phytochemistry, 2015. **117**: p. 527-536.
203. Konarev, AV, Anisimova, IN, Gavrilova, VA, Rozhkova, VT, Fido, R, Tatham, AS, and Shewry, PR, *Novel proteinase inhibitors in seeds of sunflower (Helianthus annuus L.): polymorphism, inheritance and properties.* Theoretical and Applied Genetics, 2000. **100**(1): p. 82-88.
204. Mahatmanto, T, Mylne, JS, Poth, AG, Swedberg, JE, Kaas, Q, Schaefer, H, and Craik, DJ, *The Evolution of Momordica Cyclic Peptides.* Molecular Biology and Evolution, 2014. **32**(2): p. 392-405.