

Cyclic Peptides Arising by Evolutionary Parallelism via Asparaginyl-Endopeptidase-Mediated Biosynthesis

Joshua S. Mylne,^a Lai Yue Chan,^a Aurelie H. Chanson,^a Norelle L. Daly,^a Hanno Schaefer,^b Timothy L. Bailey,^a Philip Nguyencong,^a Laura Cascales,^a and David J. Craik^{a,1}

^aInstitute for Molecular Bioscience, University of Queensland, Brisbane, Queensland 4072, Australia

^bOrganismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138

The cyclic miniprotein *Momordica cochinchinensis* Trypsin Inhibitor II (MCoTI-II) (34 amino acids) is a potent trypsin inhibitor (TI) and a favored scaffold for drug design. We have cloned the corresponding genes and determined that each precursor protein contains a tandem series of cyclic TIs terminating with the more commonly known, and potentially ancestral, acyclic TI. Expression of the precursor protein in *Arabidopsis thaliana* showed that production of the cyclic TIs, but not the terminal acyclic TI, depends on asparaginyl endopeptidase (AEP) for maturation. The nature of their repetitive sequences and the almost identical structures of emerging TIs suggest these cyclic peptides evolved by internal gene amplification associated with recruitment of AEP for processing between domain repeats. This is the third example of similar AEP-mediated processing of a class of cyclic peptides from unrelated precursor proteins in phylogenetically distant plant families. This suggests that production of cyclic peptides in angiosperms has evolved in parallel using AEP as a constraining evolutionary channel. We believe this is evolutionary evidence that, in addition to its known roles in proteolysis, AEP is especially suited to performing protein cyclization.

INTRODUCTION

Novel proteins arise typically by two processes: divergence of gene duplicates and recombination events that alter DNA sequences for existing proteins (Schmidt and Davies, 2007). One recombination-mediated evolutionary event is the internal expansion of genes to create a string of repeated protein domains. Although 10 to 20% of eukaryotic proteins contain domain repeats, their genesis is not well understood (Marcotte et al., 1999; Björklund et al., 2006; Schmidt and Davies, 2007). Nevertheless, several features seem to be shared by repetitive protein domains. Bioinformatic analyses have shown that protein repeats are often short, and after the first repeat is made, the addition of further repeats is more likely (Marcotte et al., 1999). Also, several domains can duplicate in any one instance, and protein expansion with repeating domains is believed to occur from within the middle of a series of repeats (Björklund et al., 2006).

Knottins are a class of peptides containing a disulfide bond knot that has two adjacent disulfide bonds threaded by a third (Chiche et al., 2004). This knot motif, also referred to as an inhibitor cystine knot, is common in a range of unrelated proteins, including protease inhibitors from plants, animal toxins, antimicrobial peptides, as well as some examples from signaling peptides,

such as the agouti peptides (Craik et al., 2001). Many knottins have been isolated from seeds of the angiosperm family Cucurbitaceae (squash family), but only two genes for their precursor proteins have been described in the literature (Ling et al., 1993). These are short and encode an endoplasmic reticulum (ER) signal, a small prodomain, and end with the mature peptide domain (Ling et al., 1993).


The seeds of *Momordica cochinchinensis* (Cucurbitaceae), a tropical liana also called spiny bitter melon or gac, contain a typical knottin called *Momordica cochinchinensis* Trypsin Inhibitor III (MCoTI-III) as well as two unusual knottins, MCoTI-I and MCoTI-II, that are macrocyclic and therefore lack carboxyl and amino termini (Hernandez et al., 2000; Felizmenio-Quimio et al., 2001; Heitz et al., 2001). MCoTI-II has been heavily studied; it is a potent trypsin inhibitor (TI) (Avrutina et al., 2005), is exceptionally stable in plasma assays (Avrutina et al., 2005), is capable of penetrating cells (Greenwood et al., 2007), is structurally able to tolerate substitutions and additions to its loops (Craik et al., 2010), and also can be produced using inteins in *Escherichia coli* (Camarero et al., 2007) or chemoenzymatically using trypsin columns (Thongyoo et al., 2007). These properties make MCoTI-II ideal as a scaffold that can be used to stabilize peptide drugs. It is also an excellent starting point for the design of novel protease inhibitors (Thongyoo et al., 2009).

Despite having quite different amino acid sequences, the knotted cyclic structure of MCoTI-I and MCoTI-II has caused them to be grouped with members of a large class of plant cyclic peptides called cyclotides (Craik et al., 1999; Göransson et al., 1999) found in the violet (Violaceae), coffee (Rubiaceae), bean (Fabaceae), and petunia (Solanaceae) families. The cyclotide founding member was kalata B1 (Gran, 1970), and these kalata-type cyclic peptides typically comprise 28 to 37 amino acids, have three disulfide bonds, and are typically encoded by

¹ Address correspondence to d.craik@imb.uq.edu.au.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: David J. Craik (d.craik@imb.uq.edu.au).

 Some figures in this article are displayed in color online but in black and white in the print edition.

 Online version contains Web-only data.

www.plantcell.org/cgi/doi/10.1105/tpc.112.099085

dedicated precursor proteins that have an ER signal, a prodomain, one to three mature peptide domains, and end with a hydrophobic tail (Jennings et al., 2001; Dutton et al., 2004; Nguyen et al., 2011; Poth et al., 2011).

MCoTI-I and MCoTI-II also have some similarities to the PawS-derived cyclic peptides of sunflower (*Helianthus annuus*). In sunflowers, there are two unusual prealbumins, PawS1 and PawS2, that, in addition to producing napin-like seed storage albumin, also release 12 to 14 amino acid cyclic peptides with a single disulfide bond (Mylne et al., 2011).

For their maturation, both the kalata-type and PawS-derived peptide classes require asparaginyl endopeptidase (AEP, also known as vacuolar processing enzyme or legumain), an endoprotease that cleaves on the C-terminal side of Asn, and to a lesser extent Asp (Hara-Nishimura et al., 1991; Hiraiwa et al., 1999). In *Arabidopsis thaliana*, there are four AEPs (At2g25940, At1g62710, At4g32940, and At3g20210) that are genetically redundant. In the *aep* quadruple null, the major phenotype is misprocessing of seed storage proteins, a consequence of the failed cleavage at seed storage protein Asn-Pro bonds (Shimada et al., 2003; Gruis et al., 2004). Our studies of sunflower PawS1 processing in an *Arabidopsis aep* quadruple null showed that AEP is required for proper cleavage at specific Asn as well as Asp residues (Mylne et al., 2011) and that the Asp residue of the peptide within PawS1 can only be ligated to a Gly residue. How MCoTI-II is biosynthesized is unknown, but its similarity to kalata-type cyclic peptides and the presence of an Asp-Gly and an Asn-Gly in its cyclic sequence suggest two possible ligation points if it shares the same AEP-dependent mechanism of maturation.

Here we describe the discovery of three genes from *M. cochinchinensis* that all encode the cyclic knottin MCoTI-II. Instead of one peptide per precursor protein, as known for previously characterized acyclic knottins, these genes seem to have undergone extensive internal expansion, with the largest gene encoding eight repeating protein units in tandem, consisting of seven cyclic knottins and a terminal acyclic knottin. We purified and sequenced three novel cyclic knottins and two (acyclic) knottins encoded by these precursor genes. The acyclic knottins are N-terminally pyrolylated and have a standard carboxylic acid group at the C terminus. The cyclic knottins are backbone cyclic, meaning they have no amino or carboxyl termini. The similarity of the acyclic knottin and cyclic knottin units suggests that these unusual cyclic knottins in *M. cochinchinensis* evolved by internal expansion from their terminal knottin. It is unusual that cyclic and acyclic topologies of otherwise structurally identical peptides are formed from a single polypeptide precursor; thus, we named the precursor genes *Two Inhibitor Peptide TOPologies* (*TIPTOP*).

RESULTS

Cloning of the Concatemeric *TIPTOP* Genes

To understand the biosynthetic route for cyclic knottins in *M. cochinchinensis*, we designed degenerate primers to amplify the gene encoding MCoTI-II. Based on initial sequence data, we designed specific primers for 5' and 3' RACE that amplified

several fragments with shared 5' and 3' untranslated region (UTR) sequences. PCR with primers for these regions amplified three products. Each product was a full-length transcript encoding an ER signal sequence, and each ended with an acyclic knottin domain. However, between these two domains was a repeating series of MCoTI-II or similar peptides flanked by prosequences 16 residues in length (Figure 1A; see Supplemental Figure 1 online). We named them *TIPTOP* (for *Two Inhibitor Peptide TOPologies*), because they contained TI-like peptides of cyclic and acyclic topologies. PCR with genomic DNA demonstrated that the *TIPTOP1* to *TIPTOP3* genes lack introns (see Supplemental Figure 2 online). The same primers amplified *TIPTOP2* from genomic DNA of *Momordica sphaeroidea*, predicted by molecular dating analyses to have shared a common ancestor with *M. cochinchinensis* ~3.94 million years ago (Schaefer and Renner, 2010).

In *TIPTOP* proteins, the cyclic knottins are almost identical to the acyclic knottins, but each cyclic knottin is typically flanked by Gly-Gly-Val on its proto-N terminus and Ser-Gly-Ser-Asp on its proto-C terminus (Figures 1B and 1C). This indicates that the cyclization reaction occurs between Gly at the proto-N terminus and Asp at the proto-C terminus, similar to the kalata-type and PawS-derived classes of cyclic peptide. Each cyclic knottin domain is preceded by an Asn residue (Figure 1B); therefore, the cyclic knottins probably use AEP to release both prototermini, as is the case with PawS1 (Mylne et al., 2011).

The *TIPTOP* proteins were compared with the known knottin precursors (Figure 1D). Apart from their repetitive nature, they are otherwise similar. All share an ER signal and a C-terminal knottin domain as well as a conserved region of unknown function that follows the ER signal (consensus of IELISDG). This suggests that the ancestral *TIPTOP* protein might have been a single-TI-domain protein that underwent a series of internal gene duplications. Using *TIPTOP* primers in RACE and with *M. cochinchinensis* genomic DNA, we could not amplify the gene encoding any single-TI-domain protein similar to those found in other Cucurbitaceae species (see Supplemental Figure 2 online).

Gene Confirmation through Peptide Analysis

To confirm these gene sequences, we examined the peptides deriving from them (Figure 2A). For simplicity, we hereafter use Arabic numerals instead of Roman numerals and drop the MCo prefix (e.g., MCoTI-II becomes TI-2). The number of encoded peptide domains differs for each gene (Figure 1A); *TIPTOP1* encodes an array of five peptides, starting with TI-1, followed by three TI-2 peptides, and terminating with acyclic TI-5. *TIPTOP2* encodes six peptides, starting with TI-1, three TI-2 units, TI-4, and terminating with acyclic TI-5. *TIPTOP3* is the largest of the genes and encodes eight peptides, including TI-8, followed by five TI-2 peptides, TI-7, and terminating with acyclic TI-6.

In addition to encoding TI-1 and TI-2, the *TIPTOP* genes encode three novel cyclic knottins and two novel knottins (Figure 2A). To confirm their presence in vivo, we examined crude seed extract by liquid chromatography-mass spectrometry (LC-MS). Masses that support all five novel peptides were observed (Figure 2B). To detect the terminating knottins TI-5 and TI-6, we had to adjust the mass for an N-terminal pyroglutamic acid also

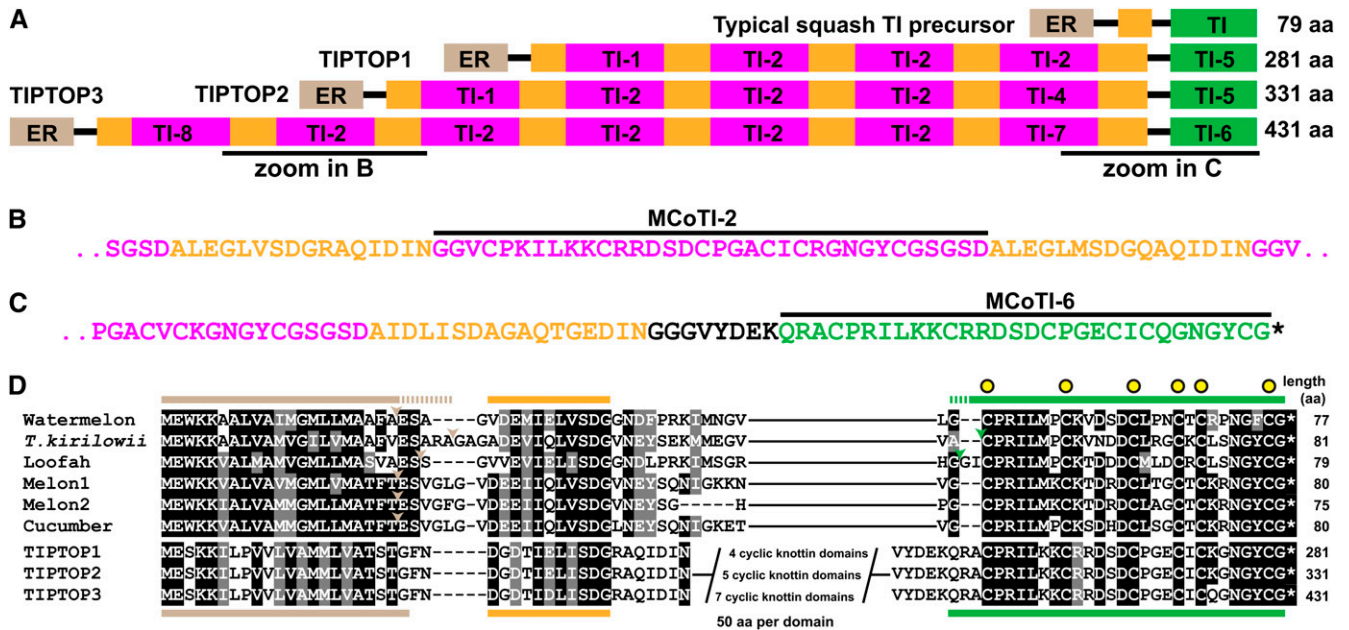


Figure 1. TIPTOP Proteins from *M. cochinchinensis*.

(A) Schematic of a typical squash TI precursor from the towel gourd (*Luffa cylindrica*) TGTI-II precursor compared with TIPTOP1-3 from *M. cochinchinensis*. aa, amino acids.

(B) Predicted sequence of a cyclic knottin domain from TIPTOP3 and its flanks.

(C) Region containing terminal knottin TI-6 from TIPTOP3.

(D) BOXSHADE alignment of six single-unit knottin precursors with TIPTOP1-3. See Methods for full details of the sources for the six single-unit knottin precursor sequences. This alignments shows that all nine predicted proteins share an ER signal sequence (brown, predicted cleavages shown with arrowheads), a conserved prodomain of unknown function (orange) and the terminal knottin domain (green, known cleavages shown with arrowheads).

seen previously for TI-3 (Hernandez et al., 2000). The previously reported TI-3 was not present in any of the *TIPTOP* genes. TI-5 is identical to TI-3 except for Gly-25, which in TI-3 is Glu. We could purify four of the five novel peptides and sequenced each by tandem mass spectrometry (MS/MS) after reduction, alkylation, and digestion with endoproteinase Glu-C, trypsin, or chymotrypsin (see Supplemental Figures 3 to 6 and Supplemental Table 1 online). We confirmed these four to have the expected sequence; TI-7 abundance was too low to be purified. It is worth noting that, for the new cyclic knottins TI-4 and TI-8, we obtained MS/MS fragmentation that crosses the Asp-Gly ligation point (see Supplemental Figures 3D and 6D online). Only subtle amino acid differences differentiate the new knottins from the three previously known (Hernandez et al., 2000).

Structural Analysis of the Knottin TI-5

Structures are available for TI-2 (1B9, 1HA9), but not its acyclic relatives. The sequences of TI-5 and TI-2 are similar (Figure 3A). We used NMR to determine the three-dimensional structure of TI-5 and found that, like TI-2, TI-5 is characterized by a cystine knot arrangement of the disulfide bonds and a β -sheet motif as the main element of secondary structure (Figure 3B). An analysis of the three-dimensional structures is provided in Supplemental Table 2 online. TI-5 and TI-2 overlay with a root-mean-square deviation of 0.55 Å over the backbone of residues 2 to 30,

highlighting their similarity. The N-terminal residues of TI-5 preceding the first Cys are slightly disordered, consistent with the disorder in loop 6 of TI-2 (Felizmenio-Quimio et al., 2001). This structural analysis of TI-5 confirmed that both peptides are almost identical apart from the Ser-Gly-Ser-Asp joining sequence, suggesting that closing the ring does not require structural rearrangement of the residues flanking the termini of the knottin.

Repeating Units Are Consistent with Gene Expansion

The most striking feature of the *TIPTOP* genes is their repetitive structure. Previous work has indicated that internal gene amplification originates from within the middle of repeating arrays (Björklund et al., 2006). This also seems to be the case with *TIPTOP* genes, which encode repeating TI-2 units in their mid-sections with variation in cyclic knottin sequences seen at the front and rear of the concatemers. We aligned the repeating domains of all three *TIPTOP* genes (see Supplemental Figure 7A online) and performed a phylogenetic analysis (see Supplemental Figure 7B and Supplemental Data Set 1 online). This analysis reinforced observations made at the peptide-coding level, with the first and last peptide domain of each gene clustering separately in an unrooted phylogram. The remaining repeat domains were so similar that their phylogenetic relationship could not be resolved (see Supplemental Figure 7B online). The similarity between the first peptide domain of *TIPTOP1*

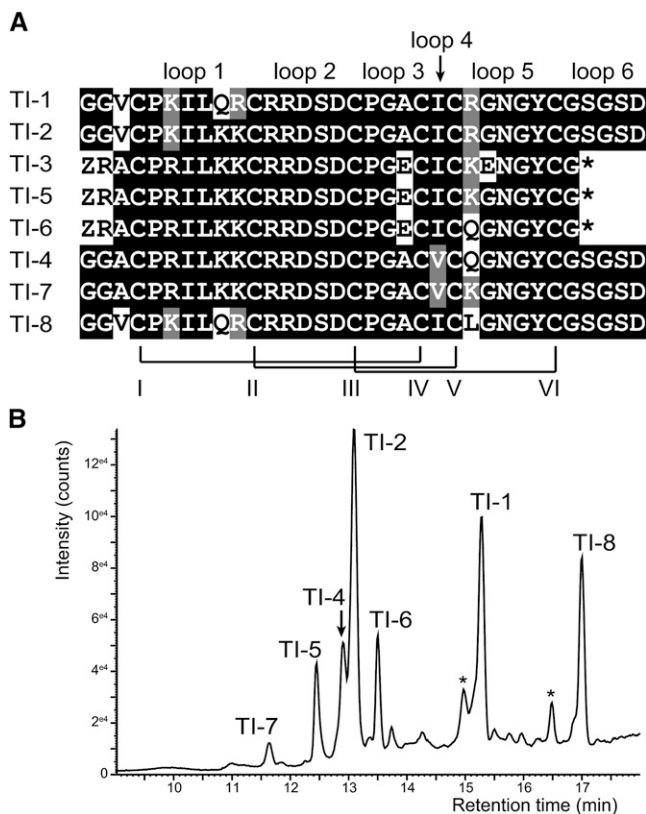


Figure 2. TIPTOP-Derived Knottins.

(A) Sequence alignment of new TI sequences (TI-4 to TI-8) with known sequences (TI-1 to TI-3). Asterisks indicate an acyclic peptide. TI-1, TI-2, TI-4, TI-7, and TI-8 are backbone cyclic. The disulfide connectivity determined by NMR for TI-2 and TI-5 is shown below the alignment.

(B) LC-MS profile of *M. cochinchinensis* peptide extract with sequenced knottins marked. The two peaks with asterisks we suspect contain isomers of identical mass to nearby peaks but with isoaspartyl bonds, a feature of these cyclic knottins observed during their initial characterization (Hernandez et al., 2000).

to *TIPTOP3* and the last peptide domain of *TIPTOP1* to *TIPTOP3* implies that the three *TIPTOP* genes are the product of gene duplication after our proposed expansion. We also used nucleotide and protein alignments of the repeating protein domains (see Supplemental Figures 1B and 1C online) to compare neighboring repeats (see Supplemental Figures 7C and 7D online). This approach reinforced that, for the first and last repeat, there is a strong link between genes, whereas for most of the middle domains, they are so similar to each other that it is impossible to establish any relationships.

The alignment of TIPTOP proteins with single-knottin precursors (Figure 1D) identifies the region where TIPTOP proteins diverge sharply in their predicted protein sequence. Specifically, this is between the conserved consensus sequence IELLISDG and the first Cys residue of the terminal knottin domain. The domains encoding cyclic and acyclic knottins are similar, and this is mirrored in the DNA that encodes them. The 82-base DNA sequence preceding the stop codon in each *TIPTOP* gene

shares 86.5% identity with the repeat unit before it. This region encodes 27 amino acids that share 92.6% identity with the cyclic peptide unit before it. This similarity strongly suggests the internal units encoding the cyclic knottins arose from a duplicated segment of DNA encoding the single ancestral knottin. For this to be the case, in addition to duplication, the duplicated segment would have had to acquire subsequent deletion or frame shifting of the DNA sequence encoding GVVDEKQRA of the knottin as well as addition of a sequence that encodes SGSD-ALEG at the C terminus of each cyclic knottin. Therefore, there is no single, simple duplication and adjacent placement that can be proposed to explain the appearance of cyclic knottins. The simplest scenario is that a DNA segment encoding the terminal knottin was tandemly duplicated and subjected to rearrangements and sequence additions around the flanks, and then this first cyclic knottin underwent additional internal duplication events.

TIPTOP Repeat Units Contain Low Folding Free Energy Sequences

Several repeat proteins from other species have been found to contain palindromic elements (Ogata et al., 2000; Claverie and Ogata, 2003). To ascertain whether *TIPTOP* repeats contained palindromes, we analyzed the DNA sequence of *TIPTOP2* using MEME (Bailey and Elkan, 1994), a program designed to detect repetitive DNA motifs and palindromes. Querying MEME to detect palindromes using default settings, we found that the top-scoring palindrome was a series of related 50-mer repeats within all the cyclic knottin domains as well as the terminal acyclic knottin domain (Figure 4A). The 50-mers are each imperfect palindromes that encode four Cys residues from loop 2 until the sixth Cys (Figures 4B and 4C). To establish whether these imperfect palindromes were statistically significant, we generated a histogram of folding free energies from randomly chosen 50-mer open reading frame (ORF) segments of *Arabidopsis* (see Supplemental Figure 8B online). We compared the folding free energy of the *TIPTOP2* 50-mers to the histogram (see Supplemental Figure 8C online). Although the folding free energies were seemingly low, we forced MEME to find palindromes that have accordingly low folding free energies, so when the P values were adjusted for multiple testing, we found that the folding free energies of these 50-mers were not statistically significant.

The palindromic nature of these repeats (albeit not statistically significant) encouraged us to explore whether *TIPTOP2* contained low folding free energy sequences. This time, we queried MEME to find repeats without forcing it to find palindromes. MEME returned a series of 113-mers that overlapped with the previous series of palindromic 50-mers (Figure 4D). To establish whether these repeat sequences were significant in structure, we compared their folding free energies to a probability histogram of folding free energies for random 113-mers (Figure 4E). The N-terminal 113-mer repeat has a predicted folding free energy that is statistically significant ($P < 0.05$) compared with random ORFs of similar length from *Arabidopsis* (Figure 4F). The folding free energies of the other five repeats are not significant at the 0.05 level; however, there is a clear trend of increasing predicted folding free energy from the N to C terminus.

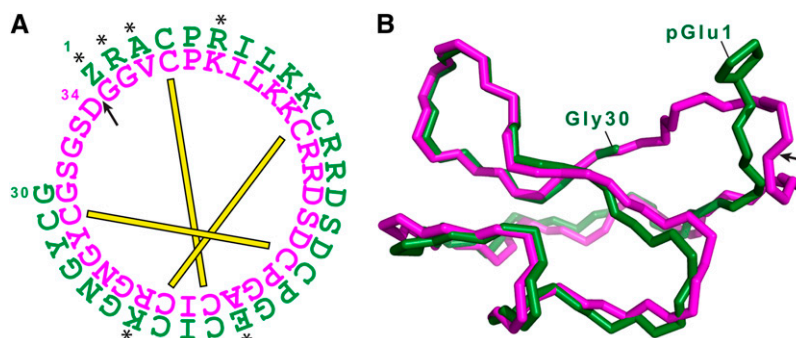


Figure 3. Sequence and Structural Alignment of Cyclic and Acyclic Knottins.

(A) Sequence of cyclic TI-2 (magenta) and acyclic TI-5 (green). The ligation point in TI-2 is marked with an arrow. Residues that differ between TI-2 and TI-5 are marked with asterisks. The three disulfide linkages are shown by connecting bars.

(B) Overlay of structural models for TI-2 (magenta, 1HA9) and the newly acquired TI-5 structure (green, 2LJS). Aside from the obvious ligating Ser-Gly-Ser-Asp sequence in TI-2, TI-5 has a root mean square deviation of 0.55 Å over the backbone residues 2 to 30. The ligation point in TI-2 is marked on its structure with an arrow. The N-terminal pyrrol ring of TI-5 is displayed (pGlu1).

AEP-Dependent Maturation of Cyclic Knottins from TIPTOP2

To test whether AEP is critical for the maturation of cyclic knottins, we expressed TIPTOP2 in *Arabidopsis* using the strong seed-specific promoter of *OLEOSIN* (Parmenter et al., 1995). We transformed the *OLEOSIN:TIPTOP2* construct into wild-type *Arabidopsis* and an *aep* null mutant that has lesions in all four *Arabidopsis* AEP genes (Kuroyanagi et al., 2005). The knottin TI-5 is not preceded by Asn or Asp, and so we expected it to be matured in an AEP-independent manner (Figure 5A). If TIPTOP2 is correctly processed in *Arabidopsis*, it will yield the peptides TI-1 (3480.97 D), TI-2 (3453.00 D), TI-4 (3410.88 D), and the N-terminally pyrolyated TI-5 (3306.86 D). Peptides were extracted from T₂ seeds and analyzed by matrix-assisted laser desorption/ionization (MALDI)-mass spectrometry (MS).

In a wild-type background, we detected all the predicted cyclic masses as well as a mass for TI-5 containing a pyroglutamic acid residue (Figure 5B), suggesting *Arabidopsis* could process all TIPTOP2-derived knottins correctly. For each peptide, we also detected +18-D masses consistent with unligated peptide and a +17-D mass consistent with nonpyrolyated TI-5 (i.e., a free N terminus). Presence of nonligated peptides was also reported in Mylne et al. (2011), when sunflower PawS1 was similarly expressed in *Arabidopsis* using the *OLEOSIN* promoter. The efficiency of cyclization in *Arabidopsis* varied between TIPTOP2-derived peptides. Assuming identical MS ionization strengths by cyclic and acyclic versions of the same peptide, the cyclization efficiency by *Arabidopsis* as judged by the peak height of cyclic to a combined peak height for cyclic and unligated peptide masses was ~45% for TI-1, ~70% for TI-2, and ~55% for TI-4 (Figure 5B). Pyrolyation efficiency for TI-5 by *Arabidopsis* was ~80%. The cyclization efficiencies from TIPTOP2 are much higher than that of SFTI-1, which is ~5% when expressed from an *OLEOSIN:PawS1* construct (Mylne et al., 2011).

The *OLEOSIN:TIPTOP2* construct in an *aep* null mutant background revealed none of the masses for either cyclic or unligated TI-1, TI-2, or TI-4 (Figure 5B), suggesting AEP is indeed required for their release from TIPTOP2. By contrast, the

mass for TI-5 remained detectable in the *aep* null mutant background (Figure 5B), confirming that AEP is not required to release this knottin from within TIPTOP2.

We confirmed the presence of all four TIPTOP2-encoded knottins in *Arabidopsis* seeds using LC-MS, using *M. cochinchinensis* knottins as controls (Figure 5C). In the *aep* null mutant background, LC-MS data confirmed the maturation of TI-5 and the absence of any detectable cyclic knottin. Therefore, TIPTOP2 requires AEP to mature its cyclic knottins, but maturation of the terminal knottin is AEP-independent.

Features Conserved between TIPTOP Proteins and Other Cyclic Peptide Precursors

Precursors for two other classes of plant cyclic peptide are known: the kalata-type and the PawS-derived peptides (Figure 6A). The kalata-type cyclic peptides are matured from precursors that differ greatly in their structure and have been found in the Rubiaceae (Craik, 2001; Jennings et al., 2001), Violaceae (Dutton et al., 2004), Fabaceae (Nguyen et al., 2011; Poth et al., 2011), and Solanaceae (Poth et al., 2012). The small cyclic peptides from sunflower are derived from bifunctional 2S seed storage albumin precursors (Mylne et al., 2011). The cyclic knottins are similar to kalata-type cyclic peptides in their size, number of disulfides, and knotted structure. They also have some similarities to sunflower SFTI-1—they are found in seeds and are TIs. Although the TIPTOP proteins that produce cyclic knottins are very different from those of the other two peptide classes (Figure 6A), alignment of the sequences flanking the mature peptide domain suggests they use a similar method of processing (Figure 6B).

In TIPTOP proteins, each cyclic knottin domain is preceded by Asn and ends with Asp, as is the case for SFTI-1 within PawS1, where cleavage at these sites was shown to require AEP (Mylne et al., 2011). Another TIPTOP feature conserved with peptides in the other two classes is that the first residue of each cyclic knottin is always Gly. The prototerminal Asn or Asp in all three classes of cyclic peptide domain is typically followed by a small residue at the P1' position, and at the P2' position, all have a Leu or Ile residue.

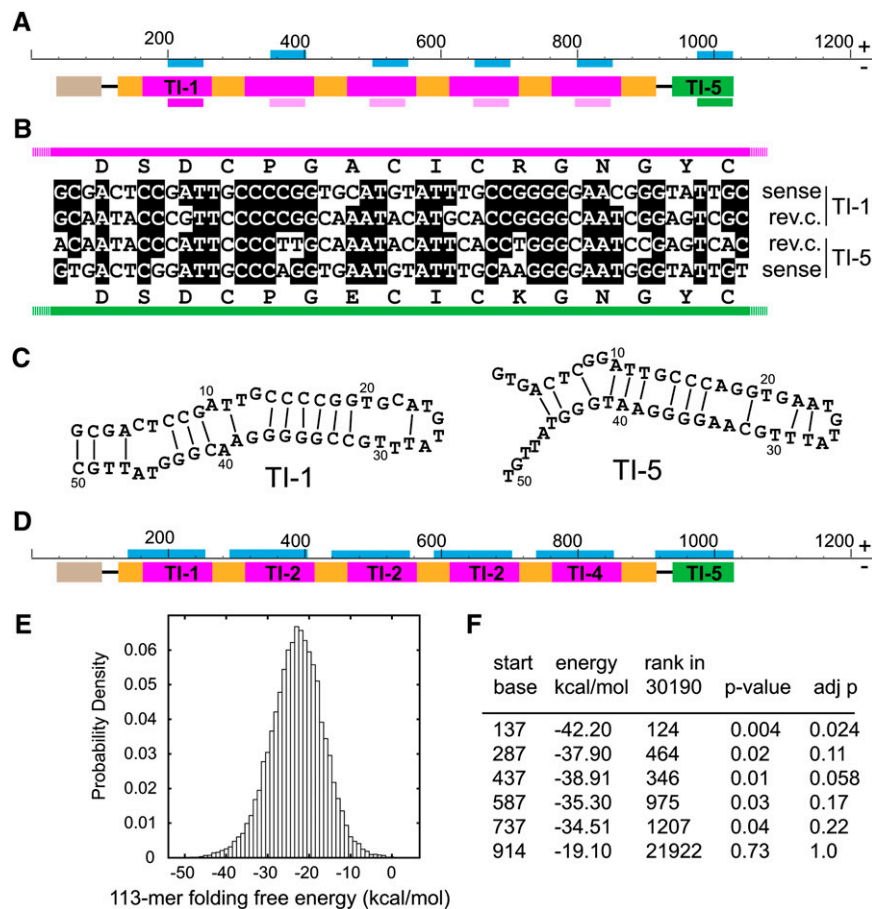


Figure 4. DNA Repeat Analysis Using *TIPTOP2* Reveals Imperfect Palindromes and Significant Low Energy Folding Structures.

(A) Reconstruction of the MEME raw output, showing the location of 50 base repeats found on the sense (+) and minus (−) strand. Below the MEME output, the equivalent regions are marked on the *TIPTOP2* protein schematic.

(B) When the regions encoding this sequence were compared with their reverse complements (rev.c.), it revealed the sequences are highly complementary.

(C) Putative DNA hairpins in TI-1 and TI-5 generated using CONTRAfold.

(D) Reconstruction of the MEME raw output when repeat size maxima was uncapped, showing the location of high-scoring 113-mer repeats.

(E) A histogram displaying the empirical probability of 113-mers with a given folding free energy, estimated using 30,190 random 113-mers extracted from unspliced *Arabidopsis* mRNA.

(F) A summary of the folding free energies and statistical significance of the 113-mer repeats shown in (D). The P value is the area under the histogram corresponding to free energies greater than or equal to the given value; the adjusted P value (adj p) is adjusted for six multiple tests, because we chose the repeat copy with the lowest free energy.

These three classes of plant cyclic peptides were all discovered based on bioactivities, but there seems to be a strong evolutionary bias toward a similar mode of processing involving AEP.

DISCUSSION

Here we have described the *TIPTOP* genes of *M. cochinchinensis*, each of which encode a string of backbone-cyclized knottins and end with the more usual acyclic knottin. The concatemeric arrangement of 150-base imperfect repeats in *TIPTOP* genes suggests that they have undergone extensive internal duplication. Previous studies of internal protein expansion, particularly for larger repeats, have attributed the origin of

the repeats to faulty recombination (Marcotte et al., 1999). Although the first repeat-causing event is not well understood, expansion of repeats often continues from within the middle of a series of repeat sequences (Björklund et al., 2006). The sequences of *TIPTOP* repeats are consistent with these observations. For example, the TIs in *TIPTOP3* are arranged 8-2-2-2-2-7-6, with the central repeats the most similar and with variants at the extremities (Figure 1; see Supplemental Figure 7 online).

At present, this consistency of *TIPTOP* repeat sequences with the features of known expanding proteins suggests genetic expansion. The squash TI class of knottins has been characterized thoroughly at the peptide level, and this class is so

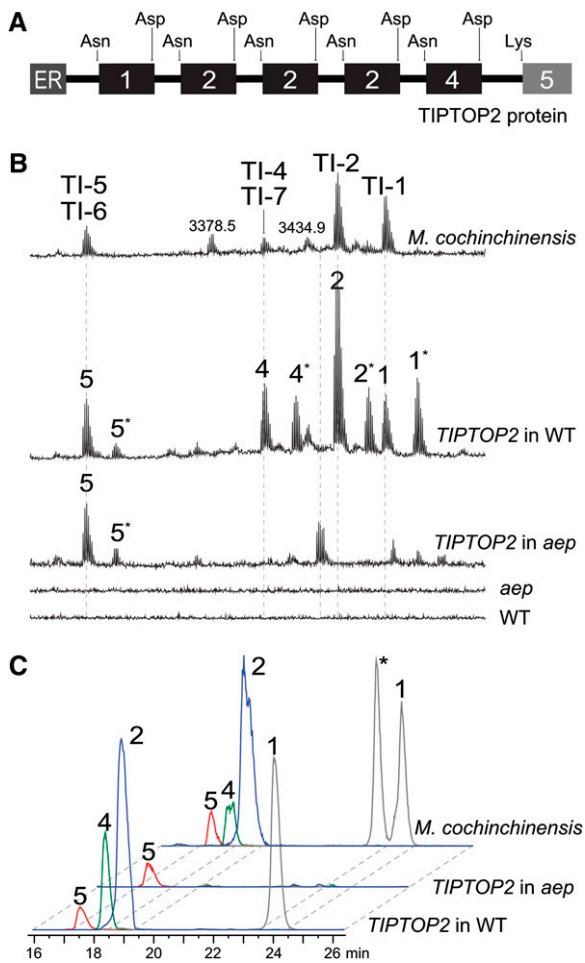


Figure 5. In Vivo Processing of TIPTOP2.

(A) Schematic of TIPTOP2 showing the Asn residue preceding each cyclic knottin domain, the terminal Asp of each cyclic knottin domain, and the Lys residue preceding the terminal knottin.

(B) MALDI-MS analysis of seed peptide extracts of either *M. cochinchinensis* or *Arabidopsis* containing *OLEOSIN:TIPTOP2* in either wild-type (WT) or *aep* null mutant backgrounds. The identity of *M. cochinchinensis* masses 3378.5 and 3434.9 are not known; those that match known peptides are labeled. The asterisks in the *OLEOSIN:TIPTOP2* in wild-type spectra denote misprocessed peptides. For TI-5, this mass is consistent with failure to pyrolyse (+17 D), whereas for TI-1, TI-2, and TI-4, the masses marked by 1*, 2*, and 4*, respectively are +18-D masses consistent with noncyclized peptide. For comparison, nontransgenic wild-type and *aep* null mutant profiles are shown. See Supplemental Figure 9 online for MALDI-MS spectra with a broader mass range.

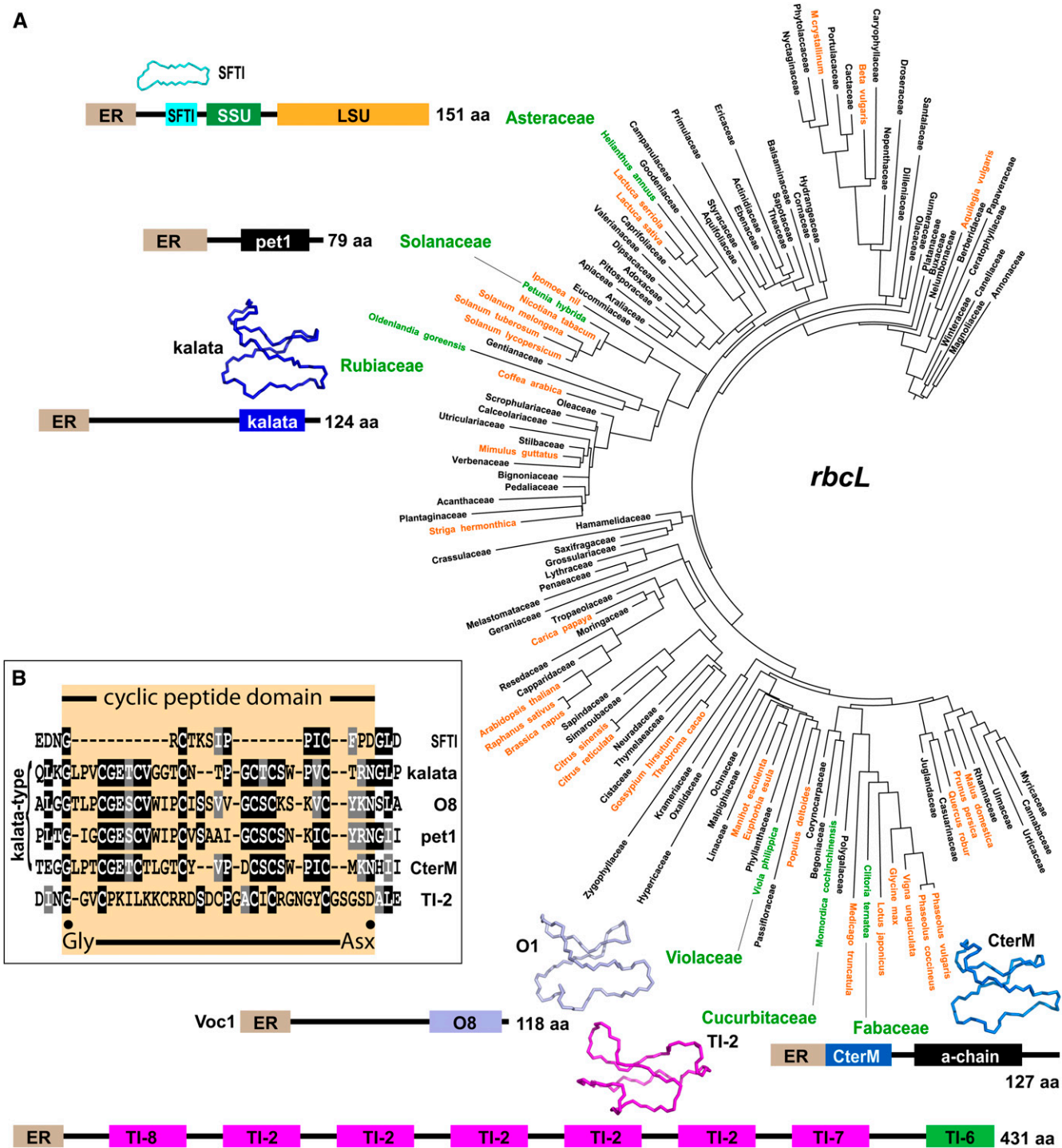
(C) Ions within the LC-MS data of the same extracts with ranges 827.2 to 827.4, 853.0 to 853.2, 863.5 to 863.7, and 870.5 to 870.7 D confirmed each peak in *Arabidopsis* matches its counterpart in *M. cochinchinensis*. The peak with the asterisk marks what we suspect is a TI-1 isomer. For fully annotated LC-MS traces, see Supplemental Figure 10 online. [See online article for color version of this figure.]

named because members are restricted to the squash (Cucurbitaceae) family. Showing that the lineage with the multiunit TIPTOPs diverged from other lineages that contain single-unit genes but no multiunit TIPTOPs would support our proposal that TIPTOPs evolved by expansion of a common ancestral single-unit squash TI precursor. *Momordica* is outside of a clade that includes most of cucurbit diversity, so the single-unit cDNA transcripts presented in Figure 1E from four other cucurbit species (tribes Trichosantheae and Benicaseae) cannot be used to confirm expansion, because these species' lineages are all more closely related to one another than to *Momordica*, and thus might have experienced contraction from a multiknottin ancestor. Consequently, we cannot experimentally confirm expansion over the alternative possibility of gene contraction in Cucurbitaceae after the *Momordica* lineage diverged.

The feature that separates the TIPTOP proteins from all other cyclic peptide precursors is that they contain almost identical matured proteins but include both cyclic and acyclic versions. These dual topologies arising from within one protein have some similarities to the PawS proteins, from which a cyclic peptide and a linear heterodimeric napin-type albumin arise. However, in PawS proteins, the two matured products differ greatly in size, sequence, three-dimensional structure, and function (Mylné et al., 2011).

We found that the acyclic *Momordica* TIs have a pyroglutamic acid at their N terminus. A pyroglutamic acid can arise by conversion of either Glu or Gln. The TIPTOP genes encode Gln at their knottin proto-N termini. Conversion of Gln to pyroglutamic acid is a deamination reaction that confers resistance to degradation by aminopeptidases (Schilling et al., 2008). Structurally, we showed that the amino and carboxyl termini of the TI-5 knottin are exposed. An advantage that cyclization offers over pyrolylation is it would provide resistance to amino- and carboxypeptidases. Even endoproteases have difficulty cleaving cyclic peptides containing internal disulfide bonds. In vitro and in vivo studies of disulfide-rich conotoxin drug leads (Clark et al., 2005; Clark et al., 2010) have shown that improved stability may be achieved by synthetic cyclization. Previous studies with TI-2 have shown that acyclic mimics have reduced trypsin inhibitory activity. TI-2 is 10 times more effective at inhibiting trypsin than an open chain TI-2 variant that lacks the ligating loop 6 (Avrutina et al., 2005). Of these two possibilities, it is unclear whether the evolved cyclic innovation in *Momordica* is providing cyclic knottins with greater in planta stability, higher activity, or perhaps both, compared with their linear homologs.

There are four known classes of gene-encoded backbone-cyclized peptides in the plant kingdom that are found in distantly related families. Of these, we believe only three are using AEP-mediated processing. Each of these classes arises from a different type of precursor. Most kalata-type cyclic peptides arise from dedicated precursors that may have one, two, or three peptide units (Jennings et al., 2001). The kalata-type cyclic peptides from the legume *Clitoria ternatea* (Nguyen et al., 2011; Poth et al., 2011) are encoded by a protein similar to pea (*Pisum sativum*) Pa1-albumin, except that the usual first of two albumin domains has been replaced by a kalata-type sequence. The PawS-derived cyclic peptides (Mylné et al., 2011) are embedded within unusual bifunctional proteins. The cyclic knottins here are



derived from a concatemeric protein that seems to have arisen from internal expansion of a precursor for a knottin. Despite these different peptides and precursors, there seem to be significant shared features between the TIPTOP proteins and both the kalata-type and PawS-derived classes, whose maturation and proposed cyclization is by AEP. These shared features include the proto-N-terminal Gly, proto-C-terminal Asp/Asn, and the trailing small P1' residue and P2' Leu.

The fourth class of gene-encoded backbone-cyclized peptides in the plant kingdom is found in members of the Caryophyllaceae and Rutaceae plant families and highlights that this AEP-mediated mechanism is unlikely to be the only way plants can produce cyclic peptides. Condie et al. (2011) revealed that five to nine residue segetalins lacking disulfides emerge from short precursors. Critically, the segetalins and their precursors do not contain conspicuous Asp or Asn residues, the target residues of AEP. Furthermore, segetalin precursors share none of the other properties seen for AEP-mediated cyclic peptides.

Involvement of AEP in cyclic peptide maturation has previously been shown for SFTI-1, which is derived from the unusual seed storage preproalbumin PawS1. Not only is AEP well established as being essential for seed storage albumin maturation (Shimada et al., 2003; Gruis et al., 2004), but when *PawS1* constructs were transformed into an *Arabidopsis aep* null mutant, the misprocessing detectable by MALDI-MS and proteomics analyses demonstrated that AEP was required to release the cyclic peptide domain at both prototermini (Mylne et al., 2011). The prototermini of the cyclic peptide in PawS1 are identical to those of the cyclic knottins within TIPTOP1-3. Each cyclic knottin domain is preceded by Asn and ends with Asp, both target residues for AEP (Hara-Hishimura et al., 1993; Hiraiwa et al., 1999). There are additional similarities when TIPTOPs are compared with PawS proteins and precursors of kalata-type cyclic peptides, namely the presence of a proto-N-terminal Gly in all cyclic peptide domains as well as the P1' and P2' residues trailing the cyclic peptide domains consisting of a small amino acid (often Gly) and an absolutely conserved Leu, respectively. This P2' Leu has been shown to be essential for the cyclization of kalata B1 associated with processing from its precursor protein Oak1 (Gillon et al., 2008) but seems not to be critical for PawS1 when expressed in *Arabidopsis* (Mylne et al., 2011). The significance of these observations is that all three classes of backbone cyclic peptide are using the same AEP-mediated mechanism for their maturation.

We hypothesized that the convergence on involvement of AEP may in part be caused by the reactive thioester acyl intermediate that AEP (a Cys protease) will form after cleavage at the scissile peptide bond. After this cleavage, the conditions at the enzyme active site would be entropically ideal for the acyl intermediate to react with the N terminus of an unmasked Gly, instead of with water (Gillon et al., 2008; Mylne et al., 2011). This mechanism requires sequential cleavages that first unmask the Gly at the proto-N terminus of the peptide before the cleavage at the proto-C-terminal Asp. For PawS2, the Gly is unmasked by ER signal removal, but in PawS1, it is believed to be unmasked by preferential cleavage at Asn, the preferred substrate of AEP. The TIPTOP proteins all share the same Asn/Asp, which we

believe permits sequential cleavage in PawS1. At the proto-C terminus of kalata B1, a role for AEP has been implied from transient transformation of tobacco (*Nicotiana tabacum*) with the kalata B1 precursor Oak1 with and without AEP-silencing constructs (Saska et al., 2007). However, which enzyme unmasks the Gly at the proto-N terminus is unknown. There is little, if any, sequence conservation at the residues preceding each kalata-type peptide domain, suggesting there may be several enzymes capable of removing the prodomain. The identity of the enzyme that releases the knottin at the end of each TIPTOP-predicted protein is equally uncertain. TI-5 and TI-6 are each preceded by Lys (Figure 1D), and TI-5 release is proven to be AEP-independent (Figure 5). In other squash TI precursors, the knottins are preceded by Gly or Ala (Ling et al., 1993). Unlike SFTI-1, which coopts AEP from its adjacent role in seed storage albumin maturation, the cyclic knottins derived from within the TIPTOP proteins have acquired a completely new requirement for AEP.

These commonalities in cyclic peptide-processing residues have been found for three structurally different classes of peptide. Not only do the peptides differ structurally, but they are also embedded within precursor proteins of very different architectures and in unrelated plant families (Figure 6). The kalata-type family of cyclic peptides are usually embedded in proteins that encode signal peptides and prodomains but no mature peptides other than kalata types, and these are from the phylogenetically distant families Rubiaceae (Jennings et al., 2001) and Violaceae (Dutton et al., 2004). An interesting exception is the kalata-type cyclic peptides found in the legume *C. ternatea*, which are encoded by a Pa1-like albumin protein, but one in which the Pa1b domain has been replaced (Nguyen et al., 2011; Poth et al., 2011). A very recent discovery is a new precursor structure that encodes kalata-type cyclic peptides in the Solanaceae (Poth et al., 2012). The TIPTOP proteins we found in *Momordica* (Cucurbitaceae) are distinct from those in kalata-bearing families Fabaceae and Violaceae (Figure 6). Six plant lineages that are phylogenetically quite distantly related have converged to use the same AEP-dependent processing to make three classes of cyclic peptides, and we believe this provides strong evidence of evolutionary parallelism.

In this context, the term parallelism refers to the independent evolution of the same derived trait via the same developmental changes, whereas convergent evolution refers to superficially similar traits that have a distinct developmental basis (Patterson, 1982; Yoon and Baum, 2004). These three peptide classes are all using AEP-mediated processing and have certain conserved residues. Importantly, the occurrence of parallelism shows that the path of evolution for a particular trait seems to be constrained to certain channels. We have proposed AEP is especially suited for performing ligation reactions; therefore, AEP might be the constraining evolutionary channel inferred by parallelism.

For the three cyclic peptide classes using this AEP-mediated processing, the founding peptide member for each class was discovered based on bioactivity, which does not bias discovery toward any particular biosynthetic mechanism. The PawS-derived peptide ring SFTI-1 was first identified by in-gel trypsin inhibition assays with the common sunflower *H. annuus* (Luckett et al., 1999). The founding member of the large kalata-type family

was discovered as the uterotonic component of extracts of a traditional Congo medicine derived from a tea made from *Oldenlandia affinis* (Gran, 1970). The cyclic knottins TI-1 and TI-2 were discovered for TI activity in the traditional Chinese medicinal plant *M. cochinchinensis* (Hernandez et al., 2000). Despite their independent discovery, different precursors, and structural diversity, all three classes depend on AEP for their maturation. By International Union of Biochemistry and Molecular Biology definition EC 3.4, a protease catalyzes peptide bond hydrolysis, but based on thermodynamic reversibility, it can also catalyze peptide bond formation. More than 70 years ago, Max Bergmann used several proteases, including chymotrypsin, to induce bond formation in vitro (Bergmann and Fruton, 1938). More recently and relevant to our case, jack bean (*Canavalia ensiformis*) AEP was demonstrated to perform a transpeptidation reaction in vitro (Min and Jones, 1994). For structurally constrained peptide substrates, the N terminus is held close to the scissile bond during cleavage, favoring ligation that leads to cyclic peptides. Although plants typically contain hundreds of proteases (García-Lorenzo et al., 2006), evolution has produced cyclic peptides several times via AEP processing, suggesting that this type of protease is especially favorable for performing protein cyclization.

METHODS

Plant Material

Momordica cochinchinensis fruits were obtained from a commercial vendor of Vietnamese produce at the Footscray (Melbourne, Australia) markets. The seeds were removed from fruits and washed. The seed coat was removed initially with sandpaper and then, once the coat was broken in a large enough area, was peeled away. The mature embryos were washed briefly in water and dried on a paper towel.

Genomic DNA Extraction

100 mg of *M. cochinchinensis* leaf tissue was ground under liquid nitrogen to a fine powder, transferred to a 1.5-mL tube, and resuspended in 1 mL of cetyltrimethylammonium bromide buffer (140 mM sorbitol, 220 mM Tris-HCl, pH 8.0, 22 mM EDTA, pH 8.0, 800 mM sodium chloride, 1% sarkosyl, 0.8% cetyltrimethylammonium bromide). We incubated the mixture at 65°C for 15 min with occasional mixing. We added 0.4 mL of chloroform and mixed by inversion. The sample was centrifuged for 10 min at 17,000g, and the supernatant was transferred to a fresh tube. We added 0.7 mL of cold isopropyl alcohol and incubated the sample at -20°C for 30 min. After 30 min of centrifugation at top speed at 4°C, the dried pellet was dissolved in 0.3 mL of TE buffer (10 mM Tris, 5 mM EDTA, pH 8.0) and extracted with phenol:chloroform. The DNA was precipitated from the aqueous phase by addition of 0.1 volume of 3 M sodium acetate (pH 5.5) and two volumes of ethanol. The sample was mixed, incubated at -20°C for 1 h, and centrifuged at top speed at 4°C for 30 min. The pellet was washed with 1 mL of 70% ethanol, left to dry, and then dissolved in a final volume of 50 µL of TE buffer.

RNA Extraction

Three dehusked *M. cochinchinensis* seeds were ground under liquid nitrogen with glass beads to a fine powder. Before thawing, 0.3 mL of tissue powder was resuspended in 0.25 mL of acidic phenol (pH 4.3; Sigma-Aldrich) and 0.5 mL of 0.1 M Tris, pH 8.0, 5 mM EDTA, 0.1 M

sodium chloride, 0.5% SDS, 1% 2-mercaptoethanol that had been preheated to 65°C. This mixture was vortexed for 20 min before addition of 0.25 mL of chloroform for an additional 10 min. After centrifugation at 17,000g, the supernatant was transferred to a new tube for a second extraction with one volume of 1:1 phenol:chloroform. The mixture was centrifuged again at 17,000g, and its supernatant was precipitated with 2.5 volumes of ethanol, 0.1 volume of 3 M sodium acetate, and incubation at -80°C for 15 min. The nucleic acid pellet was dissolved in 0.5 mL of water, and RNA was precipitated by addition of 0.5 mL of 4 M lithium chloride and incubation overnight at 4°C. After centrifugation at 17,000g for 10 min at 4°C, the RNA pellet was washed with 1 mL of 80% ethanol, dried, and resuspended in 60 µL of water.

Cloning of *TIPTOP* Genes

Genomic DNA (1 µg) was digested overnight with *Mcr*BC, which cleaves DNA containing methylcytosine. The digest was purified by QIAquick spin column (Qiagen) to remove digested, low-molecular-weight DNA. Because TI-2 is cyclic, its biological ligation point is impossible to know without cloning the gene. However, knowledge of cyclic peptide processing allowed us to previously postulate the sequence order TI-2 might have in its precursor protein (Daly et al., 2006). We designed degenerate primers to TI-2; namely PN02 (see Supplemental Table 3 online for all primer sequences) to the sense sequence of Gly-Gly-Val-Cys-Pro-Lys and PN10 to the reverse complement of the sequence Ile-Cys-Arg-Gly-Asn-Gly. The PCR products from this initial reaction were used in a second, nested PCR reaction with primers PN03 to the sense sequence of Val-Cys-Pro-Lys-Ile-Leu-Lys and PN11 to the reverse complement of the sequence Ala-Cys-Ile-Cys-Arg-Gly. The largest of several PCR products from this nested reaction encoded a full TI-2 encoding unit flanked by two partial TI-2 encoding units with the primers at its termini, suggesting the TI-2 precursor was multidomain and contained at least three units of TI-2. This DNA sequence was used to design a suite of specific primers for 5' and 3' RACE.

We extracted RNA with 1:1 phenol:chloroform and selective precipitation of RNA by lithium chloride. 500 ng of total RNA was used to create 5' and 3' RACE libraries using the SMARTer RACE cDNA Amplification Kit (Clontech). The RACE libraries were amplified with specific primers designed against the gene fragment. Products were cloned from the PCR reactions of the 5' RACE library with JM368 and 3' RACE libraries with JM369 and JM371. The 5' and 3' RACE products were cloned into pGEM-T (Promega), sequenced, and aligned. Although it was clear from polymorphisms that more than one gene was being amplified, the 5' UTR and 3' UTRs were identical. To amplify full-length clones, we designed JM429 to the most 5' region and a reverse primer JM430 immediately upstream from the polyA sequence in 3' RACE clones. PCR amplification of the aforementioned 5' cDNA library with these primers produced three products. Complete sequencing through all repeating sequence in a single pass required design of primers inside the ORF JM437 and JM438. The three products each encoded a full ORF that included TI-2 as well as other novel peptides. The three genes were named *TIPTOP1*, *TIPTOP2*, and *TIPTOP3*, and they encode five, six, and eight peptides, respectively. For each *TIPTOP* gene, at least five independent clones were obtained. Independent cloning events were ensured by the observed loss of one to three nucleotides at the very 5' end of either of the cloning primers in sequenced products.

PCR amplification using the JM377 and JM378 primer pair with genomic DNA produced the same three *TIPTOP1* to *TIPTOP3* products, revealing that these three *TIPTOP* genes lack introns (see Supplemental Figure 2A online). We detected faint PCR product bands below *TIPTOP1* when genomic DNA (gDNA) was used as the PCR template. Upon cloning these faint bands, we found these DNAs encoded *TIPTOP* genes with fewer peptide units, but the DNA sequences matched one of *TIPTOP1* to *TIPTOP3*, indicating these were truncated *TIPTOP* products produced as

an artifact of PCR. This artifactual nature of this lower-molecular-weight laddering by PCR was especially noticeable (see Supplemental Figure 2B online) when a plasmid containing *TIPTOP2* was PCR-amplified with primers JM439 and JM440 to make an *OLEOSIN:TIPTOP2* construct (see Seed-Specific Expression of *TIPTOP2* in *Arabidopsis* for details). PCR using JM377 and JM378 with cDNA (see Supplemental Figure 2A online) could not amplify additional *TIPTOP* genes.

Alignment of Squash TI Precursors and *TIPTOP1* to *TIPTOP3*

The alignment shown in Figure 1D contains the predicted protein sequences for *M. cochinchinensis* *TIPTOP1* to *TIPTOP3* (HQ853490 to HQ853492). They are compared with six sequences from other Cucurbitaceae. The sequence labeled watermelon (*Citrullus lanatus* subsp *vulgaris*) is from translation of a cDNA sequence filed under GenBank accession number AI563213. The sequence labeled *Trichosanthes kirilowii* came from retranslation of a *T. kirilowii* cDNA GenBank accession number X82230, which encodes the knottin TGTI-II (Ling et al., 1993). The predicted protein in GenBank is lacking 16 upstream amino acids encoded by an earlier in-frame start codon. The additional 16 amino acids add a conserved ER signal sequence. The sequence labeled Loofah (*Luffa aegyptiaca*) came from a retranslation of cDNA GenBank accession number M98055. The predicted protein in M98055 similarly lacks 16 amino acids of the ER signal sequence. The protein sequence for Melon1 is supported by 57 *Cucumis melo* expressed sequence tags, including JG526994 (see Supplemental Table 4 online for the full list). The protein sequence for Melon2 is supported by 69 *C. melo* expressed sequence tags, including JG532730 (see Supplemental Table 5 online). The sequence labeled Cucumber (*Cucumis sativus*) comes from translation of a *C. sativus* cDNA sequence filed under GenBank accession CK758797.

Peptide Extraction and NMR Analysis

The *M. cochinchinensis* tissue used for RNA was also used to extract crude peptides, which were analyzed by LC-MS (Chan et al., 2009). The LC-MS data were analyzed for predicted masses to identify their retention times. The knottins TI-5 and TI-6 had identical predicted masses but one candidate peak in the LC-MS. Individual peptides were purified as described elsewhere (Chan et al., 2009). The peptide concentrations of pure fractions were quantified using a Nanodrop UV-spectrometer (NanoDrop Technologies) and prepared for sequencing by MS/MS. Briefly, 0.5 mg of peptide was dissolved in 0.5 mL of 100 mM ammonium bicarbonate (pH 8.1) and reduced with 25 μ L of 100 mM DTT followed by incubation at 60°C under nitrogen gas for 30 min. A sample of 0.25 mL received 0.125 mL of 40 mg/mL tosyl phenylalanyl chloromethyl ketone-treated bovine trypsin (Sigma-Aldrich) or 0.125 mL of 40 μ g/mL chymotrypsin and 0.125 mL of 40 μ g/mL endoproteinase Glu-C and was incubated at 37°C for 3 h before quenching each digest with 10 μ L of 0.5% formic acid. Samples were desalted using C18 ZipTips (Millipore) and eluted with 80% (v/v) acetonitrile 0.5% (v/v) formic acid. The digest fragments were examined by MALDI-time of flight MS. A Nanospray QSTAR Pulsar I QqTOF mass spectrometer (Applied Biosystems) was used to sequence all novel knottins by selecting doubly charged and triply charged precursor ions from an initial time of flight-MS scan, followed by MS/MS on each selected product ion. A capillary voltage of 900 V was used with a collision energy between 10 to 50 V, depending on the charge and size of the ions. Analyst QS 1.5 software was used for the processing and acquisition of data. For TI-5, it was necessary to analyze by NMR to distinguish it from TI-6. NMR analysis of TI-5 and TI-6 was performed using total correlation spectroscopy and nuclear Overhauser effect spectroscopy spectra. For TI-5, we obtained a solution structure using 1 mg of purified peptide in 90% H₂O/10% D₂O (v/v). Analysis of the spectra was also used to distinguish TI-5 from TI-6 based on the Lys and Gln peaks. Structures of TI-5 were calculated using CYANA (Ikeya et al., 2006) and CNS (Brünger et al.

1997). A set of 50 structures was calculated, and the 20 lowest-energy structures were selected for further analysis, followed by structure analysis using the programs PROCHECK_NMR (Laskowski et al., 1996) and PROMOTIF (Hutchinson and Thornton, 1996) to generate statistical analysis.

Repeat Analysis of *TIPTOP2*

The *TIPTOP2* cDNA sequence (*TIPTOP2* is intronless) was submitted to the MEME server (Bailey and Elkan, 1994) at <http://meme.sdsc.edu/meme/cgi-bin/meme.cgi> using default settings, but with the “find palindromes” box checked. A high-scoring hit was the 50-mer regions shown in Figure 4A, which are imperfect palindromes. The reason that each 50-mer seems to be either on the plus or minus strand and not both was because the combined block diagrams in MEME do not display any motif occurrences that overlap. The putative hairpin models in Figure 4C were generated with CONTRAfold (<http://contra.stanford.edu/contrafold>) (Do et al., 2006) in default mode, except with “allow all possible base pairs” selected.

To establish the significance of the 50-mer repeats, we compared the folding free energies of each repeat to a histogram of random 50-mer folding free energies. Genomic resources for *M. cochinchinensis* are not available; therefore, to generate the histogram, we downloaded all unspliced ORFs for *Arabidopsis* from the Regulatory Sequence Analysis Tools website (<http://rsat.ulb.ac.be>) without masking repeats. We then extracted a random segment of size 50 bases from each of the unspliced ORFs whose length was at least 50 bases. This resulted in 35,176 DNA sequences. We then ran the RNAfold (Hofacker et al., 1994) algorithm to predict minimum energy secondary structures for 50-mer repeats found by MEME in the *TIPTOP2* gene. To estimate the free energy of binding of a random ORF segment of the same length as the *TIPTOP2* repeats, we then ran RNAfold on each of the 35,176 randomly chosen ORF segments. We plotted a histogram of the folding free energies of the randomly chosen ORF segments. In this plot (see Supplemental Figure 8B online), the y axis is the fraction of random ORFs with a given predicted folding free energy, and the x axis is the folding free energy in kcal/mol.

To examine the *TIPTOP2* repeats more closely, we ran MEME using the command “meme -dna -revcomp -mod anr -maxw 200 -minsites 2 -maxsites 20 -nmotifs 1 data/tiptop2.fasta.” This specifies that MEME look for de novo repeats of length up to 200 bases that occur between two and 20 times in the single sequence in *TIPTOP2*. This approach does not bias MEME’s search toward or away from DNA palindromes, which is important in the folding energy analysis described in the next paragraph. MEME found a repeat of 113 bases that occurs six times in the *TIPTOP2* coding sequence.

To establish the significance of the 113-mer repeats, we compared the folding free energies of each repeat to a histogram of random 113-mer folding free energies in a similar approach to the 50-mers as detailed above. We generated a histogram from unspliced ORFs for *Arabidopsis* using a random segment of size 113 bases from each of the unspliced ORFs whose length was at least 113 bases. This resulted in 30,190 DNA sequences of length 113 bases, which we ran through RNAfold and compared the folding free energies to those of each 113-mer *TIPTOP2* repeat.

Seed-Specific Expression of *TIPTOP2* in *Arabidopsis*

The *TIPTOP2* cDNA was amplified using JM439, which added a *Clal* site followed by the translation initiation sequence ACA to the *TIPTOP2* start ATG. At the 3’ end of *TIPTOP2*, the primer JM440 bound to the end of the *TIPTOP2* 3’ UTR and added a *BamHI* site. *TIPTOP2* was put under control of the *OLEOSIN* promoter and was expressed in wild-type and *aeo Arabidopsis* backgrounds as described elsewhere (Mylne et al., 2011). Peptides were extracted by grinding 100 T₂ seeds under liquid nitrogen with glass beads, adding 0.25 mL of methanol and 0.25 mL of dichloromethane and separating the phases with 0.1 mL of 0.05% trifluoroacetic acid, then mixing for 2 min at room temperature. After 5 min of

centrifugation at 17,000g, the aqueous phase was diluted fivefold to 10-fold in 50% acetonitrile 0.1% trifluoroacetic acid and was spotted for analysis by MALDI-MS.

Phylogenetic Analyses

To reconstruct the angiosperm phylogeny in Figure 6, we downloaded 187 *rbcl* sequences (see Supplemental Data Set 2 online) for 157 angiosperm families plus two outgroup gymnosperms from GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) and built a nucleotide alignment using MacClade v. 4.08 (<http://www.macclade.org>). In most cases, we included just one species to represent an entire family. We added all commonly used angiosperm model organisms based on the plant species listed in The Gene Index Project (<http://compbio.dfci.harvard.edu/tgi/plant.html>). A maximum likelihood (Felsenstein, 1973) tree search was performed using RAxML v.7.2.6 (Stamatakis et al., 2008) on the CIPRES cluster (<http://www.phylo.org/>). Based on the Akaike information criterion (Akaike, 1974) as implemented in jModeltest (Posada, 2008), we selected the general time-reversible + Γ model (six general time-reversible substitution rates, assuming gamma rate heterogeneity), with model parameters estimated over the duration of specified runs. We did not infer bootstrap values to assess statistical branch support, because the tree is only needed to visualize the general distribution of the discussed proteins across the angiosperms and not to test specific relationships between plant families. In all cases with only one species representing the family, the tips are labeled with the family name. The species used for placement of each family is detailed in Supplemental Table 4 online.

For the phylogenetic analysis of the *TIPTOP* DNA repeats, we again performed a maximum likelihood tree search using RAxML v.7.2.6 on the CIPRES cluster. We aligned the repeats using the pairwise aligning function in MacClade v. 4.08 (for sequences and alignment, see Supplemental Data Set 2 online) and chose the general time-reversible + Γ model (six general time-reversible substitution rates, assuming gamma rate heterogeneity), with model parameters estimated over the duration of specified runs. The best maximum likelihood tree was obtained using the rapid bootstrap algorithm (RAxML option: -f a).

Accession Numbers

Sequence data from this article can be found in the EMBL/GenBank data libraries under accession numbers HQ853490 for *M. cochinchinensis* *TIPTOP1*, HQ853491 for *TIPTOP2*, HQ853492 for *TIPTOP3*, and JN819554 for *M. sphaeroidea* *TIPTOP2* gDNA. Arabidopsis Genome Initiative locus identifiers referred to include At2g25940 (AEP1, α -vacuolar processing enzyme [VPE]); At1g62710 (AEP2, β -VPE); At4g32940 (AEP3, γ -VPE); At3g20210 (AEP4, δ -VPE); and At2g25890 (OLEOSIN). The atomic coordinates of TI-5 were deposited at Protein Data Bank (2LJS), and NMR restraints were deposited at Biological Magnetic Resonance Data Bank (17,956).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Alignment of *TIPTOP* Repeating Domains.

Supplemental Figure 2. PCR with cDNA and gDNA Favor Amplification of *TIPTOP1* to *TIPTOP3*.

Supplemental Figure 3. MS Data for TI-4.

Supplemental Figure 4. MS Data for TI-5.

Supplemental Figure 5. MS Data for TI-6.

Supplemental Figure 6. MS Data for TI-8.

Supplemental Figure 7. Analysis of the *TIPTOP1* to *TIPTOP3* Repeats.

Supplemental Figure 8. The Statistical Analysis of the Folding Free Energy of 50-mers Indicates the Imperfect Palindromes in *TIPTOP2* Are Not Significant.

Supplemental Figure 9. A Wider Mass Range for the MALDI Spectra Shown Zoomed in for Figure 5B.

Supplemental Figure 10. Fully Labeled LC-MS Profile of Peptide Extracts from *TIPTOP2* Expressing *Arabidopsis*.

Supplemental Figure 11. Angiosperm Phylogeny Based on *rbcl* Sequences.

Supplemental Table 1. MS/MS Product Ions for TI-4, TI-5, TI-6, and TI-8.

Supplemental Table 2. NMR and Refinement Statistics for MCoTI-V.

Supplemental Table 3. Primer Sequences Used in This Study.

Supplemental Table 4. *Cucumis* Sequences Supporting the Protein Sequence for Melon1.

Supplemental Table 5. *Cucumis* Sequences Supporting the Protein Sequence for Melon2.

Supplemental Table 6. Species Used for Placement of Families for Angiosperm Phylogeny.

Supplemental Data Set 1. NEXUS Format Text File of the Sequences and Alignment Used for the Phylogenetic Analysis of *TIPTOP* Repeats Shown in Supplemental Figure 7 Online.

Supplemental Data Set 2. NEXUS Format Text File of the *rbcl* Sequences and Alignment Used to Generate the Angiosperm Phylogeny Shown in Figure 6 and Supplemental Figure 11 Online.

ACKNOWLEDGMENTS

We thank Ikuko Hara-Nishimura for *aep* null mutant seeds and Amy Argyros for technical assistance. This study was supported by a National Health and Medical Research Council grant (APP1009267). J.S.M. is an Australian Research Council Queen Elizabeth II Fellow (DP0879133) and The John S. Mattick Fellow. N.L.D. is a Queensland Smart State Fellow. D.J.C. is a National Health and Medical Research Council Professorial Fellow.

AUTHOR CONTRIBUTIONS

J.S.M. and D.J.C. designed research; J.S.M., L.Y.C., A.H.C., N.L.D., P.N., and L.C. performed research; J.S.M., L.Y.C., and N.L.D. analyzed data; H.S. provided materials and phylogenetic analyses; T.L.B. performed folding free energy analyses; J.S.M. and D.J.C. wrote the article.

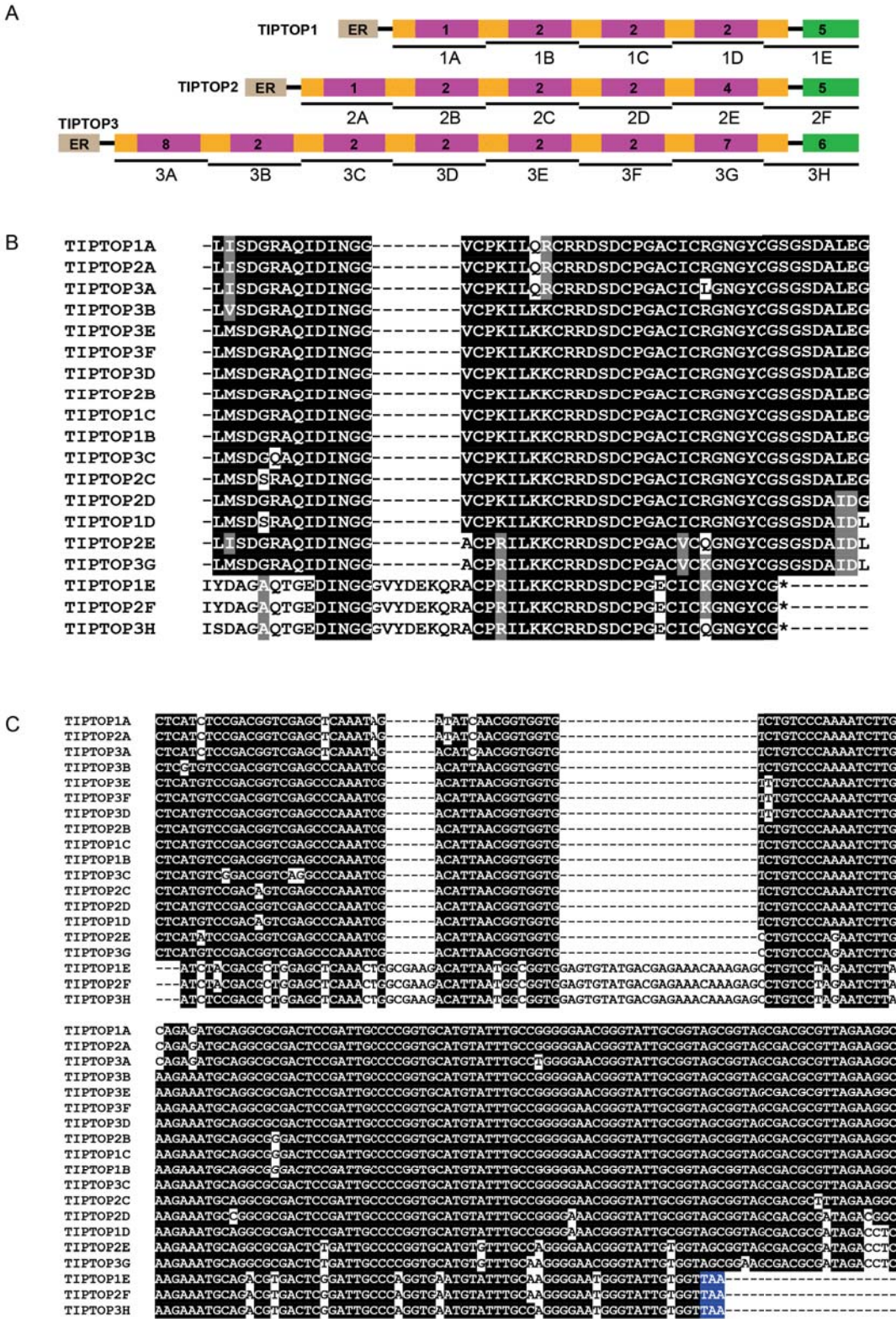
Received April 12, 2012; revised May 28, 2012; accepted June 29, 2012; published July 20, 2012.

REFERENCES

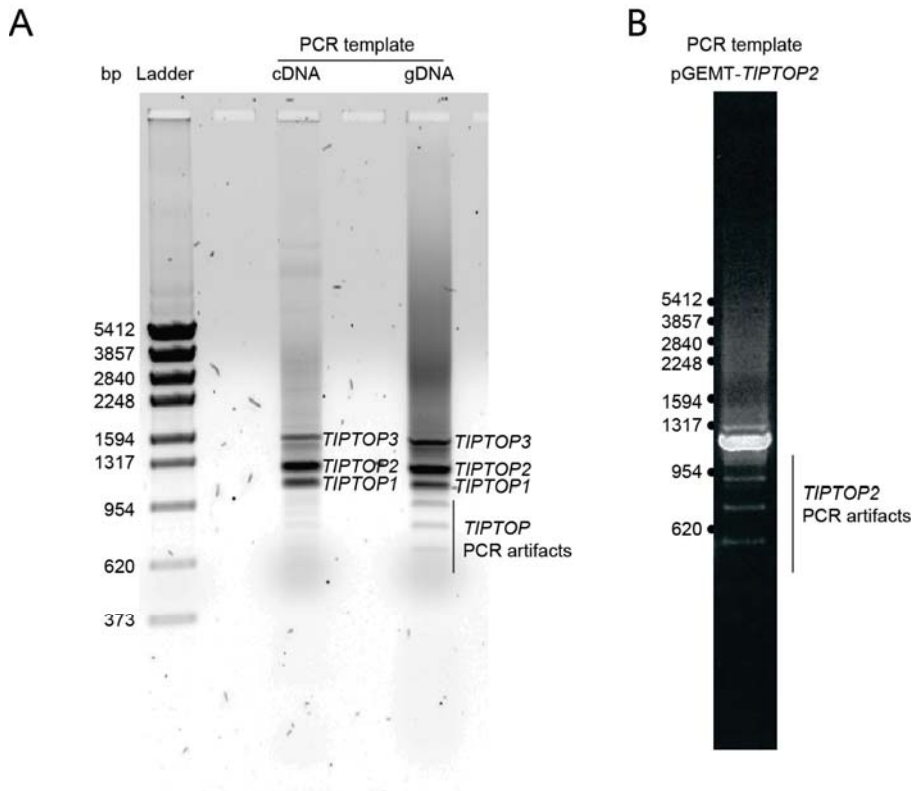
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**: 716–723.
- Avrutina, O., Schmoldt, H.-U., Gabrijelcic-Geiger, D., Le Nguyen, D., Sommerhoff, C.P., Diederichsen, U., and Kolmar, H. (2005). Trypsin inhibition by macrocyclic and open-chain variants of the squash inhibitor MCoTI-II. *Biol. Chem.* **386**: 1301–1306.

- Bailey, T.L., and Elkan, C.** (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, eds (Menlo Park, CA: AAAI Press), pp. 28–36.
- Bergmann, M., and Fruton, J.S.** (1938). Some synthetic and hydrolytic experiments with chymotrypsin. *J. Biol. Chem.* **124**: 321–329.
- Björklund, Å.K., Ekman, D., and Elofsson, A.** (2006). Expansion of protein domain repeats. *PLOS Comput. Biol.* **2**: e114.
- Brünger, A.T., Adams, P.D., and Rice, L.M.** (1997). New applications of simulated annealing in X-ray crystallography and solution NMR. *Structure* **5**: 325–336.
- Camarero, J.A., Kimura, R.H., Woo, Y.-H., Shekhtman, A., and Cantor, J.** (2007). Biosynthesis of a fully functional cyclotide inside living bacterial cells. *ChemBioChem* **8**: 1363–1366.
- Chan, L.Y., Wang, C.K., Major, J.M., Greenwood, K.P., Lewis, R.J., Craik, D.J., and Daly, N.L.** (2009). Isolation and characterization of peptides from *Momordica cochinchinensis* seeds. *J. Nat. Prod.* **72**: 1453–1458.
- Chiche, L., Heitz, A., Gelly, J.C., Gracy, J., Chau, P.T., Ha, P.T., Hernandez, J.F., and Le-Nguyen, D.** (2004). Squash inhibitors: From structural motifs to macrocyclic knottins. *Curr. Protein Pept. Sci.* **5**: 341–349.
- Clark, R.J., Fischer, H., Dempster, L., Daly, N.L., Rosengren, K.J., Nevin, S.T., Meunier, F.A., Adams, D.J., and Craik, D.J.** (2005). Engineering stable peptide toxins by means of backbone cyclization: Stabilization of the alpha-conotoxin MIII. *Proc. Natl. Acad. Sci. USA* **102**: 13767–13772.
- Clark, R.J., Jensen, J., Nevin, S.T., Callaghan, B.P., Adams, D.J., and Craik, D.J.** (2010). The engineering of an orally active conotoxin for the treatment of neuropathic pain. *Angew. Chem. Int. Ed. Engl.* **49**: 6545–6548.
- Claverie, J.-M., and Ogata, H.** (2003). The insertion of palindromic repeats in the evolution of proteins. *Trends Biochem. Sci.* **28**: 75–80.
- Condie, J.A., Nowak, G., Reed, D.W., Balsevich, J.J., Reaney, M.J. T., Arnison, P.G., and Covello, P.S.** (2011). The biosynthesis of Caryophyllaceae-like cyclic peptides in *Saponaria vaccaria* L. from DNA-encoded precursors. *Plant J.* **67**: 682–690.
- Craik, D.J.** (2001). Plant cyclotides: Circular, knotted peptide toxins. *Toxicon* **39**: 1809–1813.
- Craik, D.J., Daly, N.L., Bond, T., and Waive, C.** (1999). Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif. *J. Mol. Biol.* **294**: 1327–1336.
- Craik, D.J., Daly, N.L., and Waive, C.** (2001). The cystine knot motif in toxins and implications for drug design. *Toxicon* **39**: 43–60.
- Craik, D.J., Mylne, J.S., and Daly, N.L.** (2010). Cyclotides: Macrocyclic peptides with applications in drug design and agriculture. *Cell. Mol. Life Sci.* **67**: 9–16.
- Daly, N.L., Clark, R.J., Plan, M.R., and Craik, D.J.** (2006). Kalata B8, a novel antiviral circular protein, exhibits conformational flexibility in the cystine knot motif. *Biochem. J.* **393**: 619–626.
- Do, C.B., Woods, D.A., and Batzoglou, S.** (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**: e90–e98.
- Dutton, J.L., Renda, R.F., Waive, C., Clark, R.J., Daly, N.L., Jennings, C.V., Anderson, M.A., and Craik, D.J.** (2004). Conserved structural and sequence elements implicated in the processing of gene-encoded circular proteins. *J. Biol. Chem.* **279**: 46858–46867.
- Felizmenio-Quimio, M.E., Daly, N.L., and Craik, D.J.** (2001). Circular proteins in plants: Solution structure of a novel macrocyclic trypsin inhibitor from *Momordica cochinchinensis*. *J. Biol. Chem.* **276**: 22875–22882.
- Felsenstein, J.** (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.* **22**: 240–249.
- García-Lorenzo, M., Sjödin, A., Jansson, S., and Funk, C.** (2006). Protease gene families in *Populus* and *Arabidopsis*. *BMC Plant Biol.* **6**: 30.
- Gillon, A.D., Saska, I., Jennings, C.V., Guarino, R.F., Craik, D.J., and Anderson, M.A.** (2008). Biosynthesis of circular proteins in plants. *Plant J.* **53**: 505–515.
- Göransson, U., Luijendijk, T., Johansson, S., Bohlin, L., and Claeson, P.** (1999). Seven novel macrocyclic polypeptides from *Viola arvensis*. *J. Nat. Prod.* **62**: 283–286.
- Gran, L.** (1970). An oxytocic principle found in *Oldenlandia affinis* DC. *Medd. Nor. Farm. Selsk.* **12**: 173–180.
- Greenwood, K.P., Daly, N.L., Brown, D.L., Stow, J.L., and Craik, D. J.** (2007). The cyclic cystine knot miniprotein MCoTI-II is internalized into cells by macropinocytosis. *Int. J. Biochem. Cell Biol.* **39**: 2252–2264.
- Gruis, D., Schulze, J., and Jung, R.** (2004). Storage protein accumulation in the absence of the vacuolar processing enzyme family of cysteine proteases. *Plant Cell* **16**: 270–290.
- Hara-Nishimura, I., Inoue, K., and Nishimura, M.** (1991). A unique vacuolar processing enzyme responsible for conversion of several proprotein precursors into the mature forms. *FEBS Lett.* **294**: 89–93.
- Hara-Hishimura, I., Takeuchi, Y., Inoue, K., and Nishimura, M.** (1993). Vesicle transport and processing of the precursor to 2S albumin in pumpkin. *Plant J.* **4**: 793–800.
- Heitz, A., Hernandez, J.F., Gagnon, J., Hong, T.T., Pham, T.T., Nguyen, T.M., Le-Nguyen, D., and Chiche, L.** (2001). Solution structure of the squash trypsin inhibitor MCoTI-II. A new family for cyclic knottins. *Biochemistry* **40**: 7973–7983.
- Hernandez, J.F., Gagnon, J., Chiche, L., Nguyen, T.M., Andrieu, J.P., Heitz, A., Trinh Hong, T., Pham, T.T., and Le Nguyen, D.** (2000). Squash trypsin inhibitors from *Momordica cochinchinensis* exhibit an atypical macrocyclic structure. *Biochemistry* **39**: 5722–5730.
- Hiraiwa, N., Nishimura, M., and Hara-Nishimura, I.** (1999). Vacuolar processing enzyme is self-catalytically activated by sequential removal of the C-terminal and N-terminal propeptides. *FEBS Lett.* **447**: 213–216.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P.** (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**: 167–188.
- Hutchinson, E.G., and Thornton, J.M.** (1996). PROMOTIF—a program to identify and analyze structural motifs in proteins. *Protein Sci* **5**: 212–220.
- Ikeya, T., Terauchi, T., Güntert, P., and Kainosho, M.** (2006). Evaluation of stereo-array isotope labeling (SAIL) patterns for automated structural analysis of proteins with CYANA. *Magn. Reson. Chem.* **44**: S152–S157.
- Jennings, C., West, J., Waive, C., Craik, D., and Anderson, M.** (2001). Biosynthesis and insecticidal properties of plant cyclotides: The cyclic knotted proteins from *Oldenlandia affinis*. *Proc. Natl. Acad. Sci. USA* **98**: 10614–10619.
- Kuroyanagi, M., Yamada, K., Hatsugai, N., Kondo, M., Nishimura, M., and Hara-Nishimura, I.** (2005). Vacuolar processing enzyme is essential for mycotoxin-induced cell death in *Arabidopsis thaliana*. *J. Biol. Chem.* **280**: 32914–32920.
- Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M.** (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR.* **8**: 477–486.
- Ling, M.H., Qi, H.Y., and Chi, C.W.** (1993). Protein, cDNA, and genomic DNA sequences of the towel gourd trypsin inhibitor. A squash family inhibitor. *J. Biol. Chem.* **268**: 810–814.

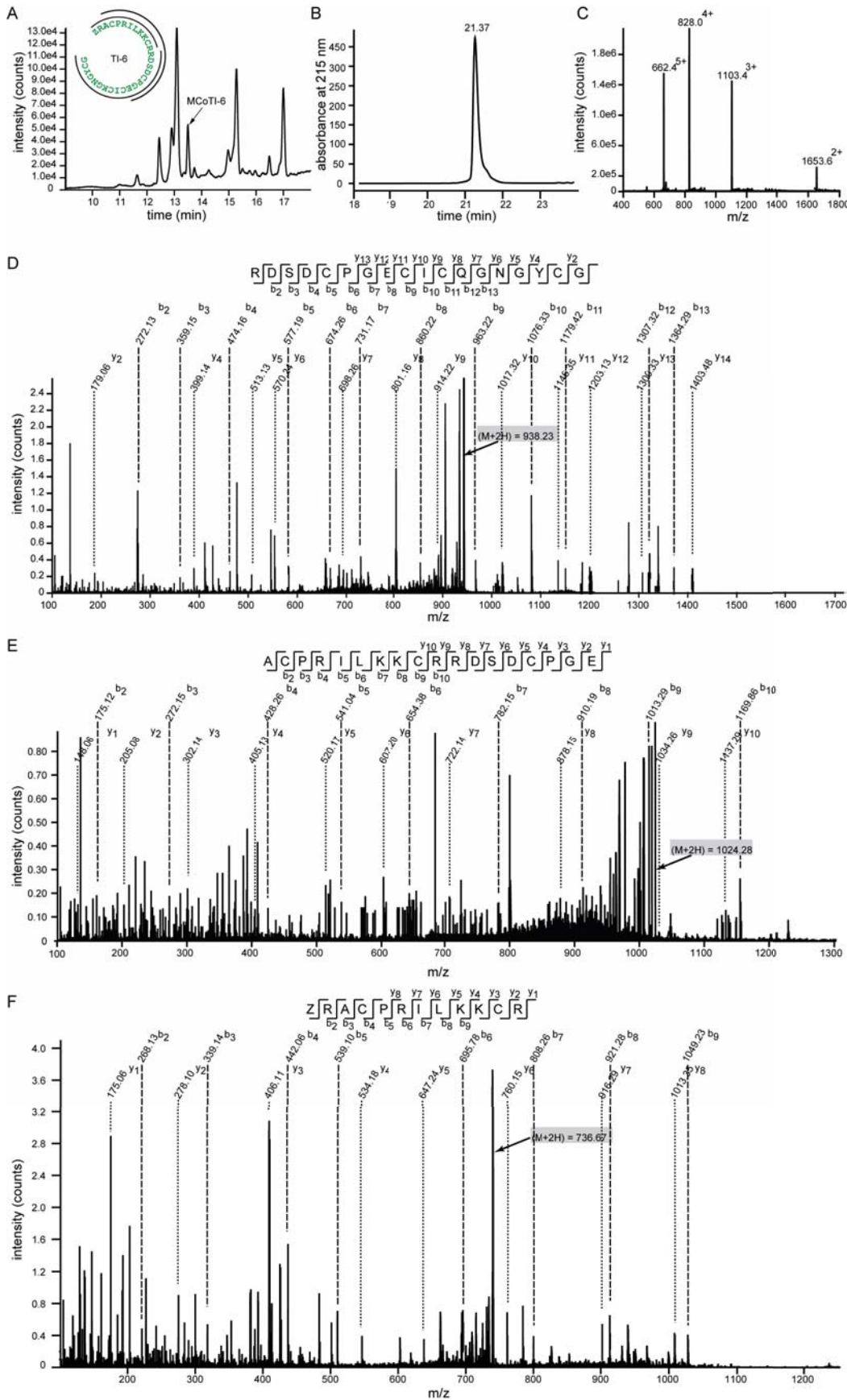
- Luckett, S., Garcia, R.S., Barker, J.J., Konarev, A.V., Shewry, P.R., Clarke, A.R., and Brady, R.L. (1999). High-resolution structure of a potent, cyclic proteinase inhibitor from sunflower seeds. *J. Mol. Biol.* **290**: 525–533.
- Marcotte, E.M., Pellegrini, M., Yeates, T.O., and Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**: 151–160.
- Min, W., and Jones, D.H. (1994). In vitro splicing of concanavalin A is catalyzed by asparaginyl endopeptidase. *Nat. Struct. Biol.* **1**: 502–504.
- Mylne, J.S., Colgrave, M.L., Daly, N.L., Chanson, A.H., Elliott, A.G., McCallum, E.J., Jones, A., and Craik, D.J. (2011). Albumins and their processing machinery are hijacked for cyclic peptides in sunflower. *Nat. Chem. Biol.* **7**: 257–259.
- Nguyen, G.K., Zhang, S., Nguyen, N.T., Nguyen, P.Q., Chiu, M.S., Hardjo, A., and Tam, J.P. (2011). Discovery and characterization of novel cyclotides originated from chimeric precursors consisting of albumin-1 chain a and cyclotide domains in the Fabaceae family. *J. Biol. Chem.* **286**: 24275–24287.
- Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.-E., Raoult, D., and Claverie, J.-M. (2000). Selfish DNA in protein-coding genes of *Rickettsia*. *Science* **290**: 347–350.
- Parmenter, D.L., Boothe, J.G., van Rooijen, G.J., Yeung, E.C., and Moloney, M.M. (1995). Production of biologically active hirudin in plant seeds using oleosin partitioning. *Plant Mol. Biol.* **29**: 1167–1180.
- Patterson, C. (1982). Morphological characters and homology. In *Problems of Phylogenetic Reconstruction*, K. Joysey and A. Friday, eds (London: Academic Press), pp. 21–74.
- Posada, D. (2008). jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol.* **25**: 1253–1256.
- Poth, A.G., Colgrave, M.L., Lyons, R.E., Daly, N.L., and Craik, D.J. (2011). Discovery of an unusual biosynthetic origin for circular proteins in legumes. *Proc. Natl. Acad. Sci. USA* **108**: 10127–10132.
- Poth, A.G., Mylne, J.S., Grassl, J., Lyons, R.E., Millar, A.H., Colgrave, M.L., and Craik, D.J. (June 14, 2012). Cyclotides associate with leaf vasculature and are the products of a novel precursor in *Petunia* (Solanaceae). *J. Biol. Chem.* <http://dx.doi.org/10.1074/jbc.M112.370841>.
- Saska, I., Gillon, A.D., Hatsugai, N., Dietzgen, R.G., Hara-Nishimura, I., Anderson, M.A., and Craik, D.J. (2007). An asparaginyl endopeptidase mediates in vivo protein backbone cyclization. *J. Biol. Chem.* **282**: 29721–29728.
- Schaefer, H., and Renner, S.S. (2010). A three-genome phylogeny of *Momordica* (Cucurbitaceae) suggests seven returns from dioecy to monoecy and recent long-distance dispersal to Asia. *Mol. Phylogenet. Evol.* **54**: 553–560.
- Schilling, S., Wasternack, C., and Demuth, H.-U. (2008). Glutaminyl cyclases from animals and plants: A case of functionally convergent protein evolution. *Biol. Chem.* **389**: 983–991.
- Schmidt, E.E., and Davies, C.J. (2007). The origins of polypeptide domains. *Bioessays* **29**: 262–270.
- Shimada, T., et al. (2003). Vacuolar processing enzymes are essential for proper processing of seed storage proteins in *Arabidopsis thaliana*. *J. Biol. Chem.* **278**: 32292–32299.
- Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**: 758–771.
- Thongyoo, P., Bonomelli, C., Leatherbarrow, R.J., and Tate, E.W. (2009). Potent inhibitors of β -tryptase and human leukocyte elastase based on the MCoTI-II scaffold. *J. Med. Chem.* **52**: 6197–6200.
- Thongyoo, P., Jalent, A.M., Tate, E.W., and Leatherbarrow, R.J. (2007). Immobilized protease-assisted synthesis of engineered cysteine-knot microproteins. *ChemBioChem* **8**: 1107–1109.
- Yoon, H.-S., and Baum, D.A. (2004). Transgenic study of parallelism in plant morphological evolution. *Proc. Natl. Acad. Sci. USA* **101**: 6524–6529.



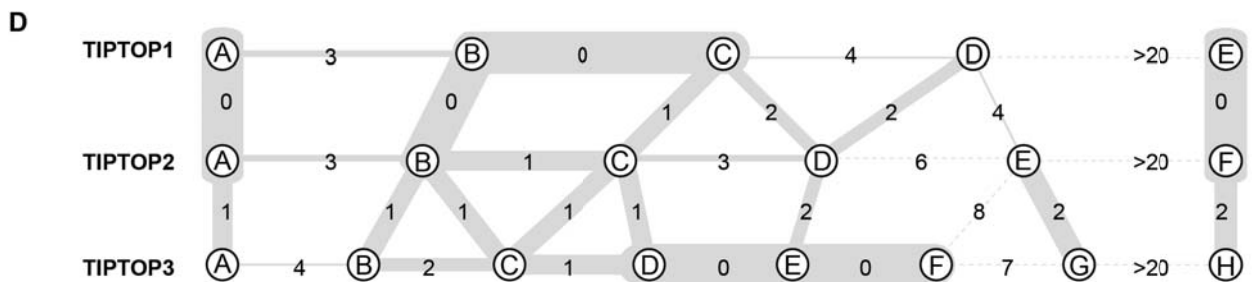
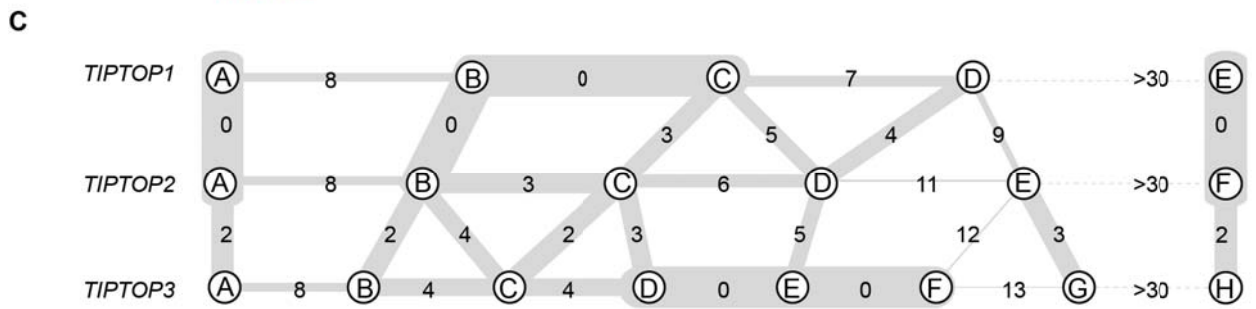
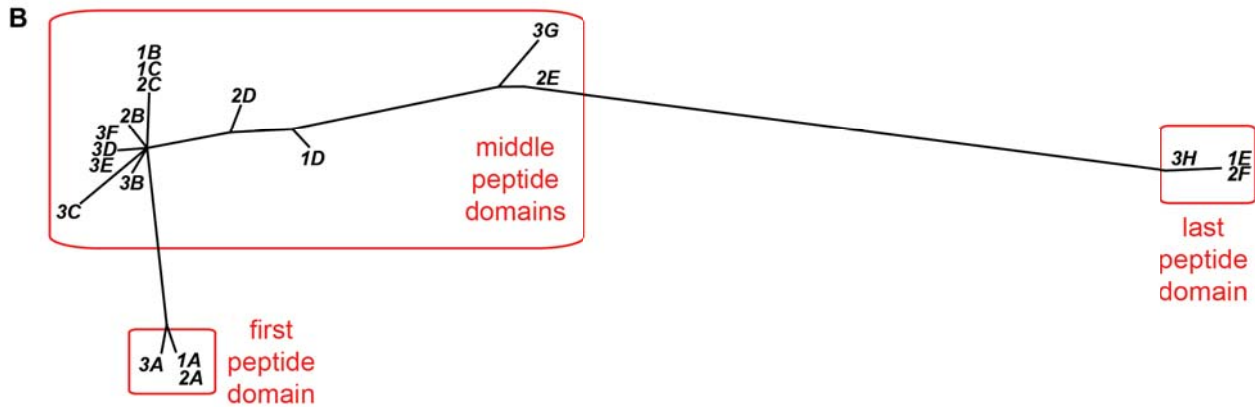
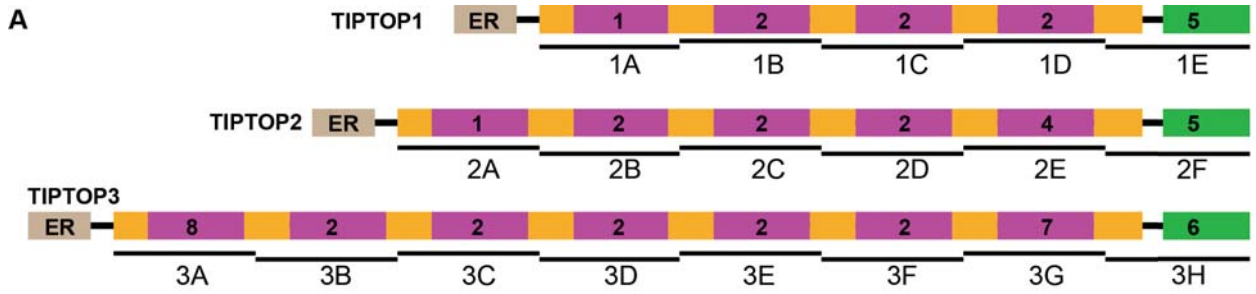
Supplemental Figure 1: Alignment of TIPTOP repeating domains. **A.** Labeling scheme for TIPTOP1-3 repeats. **B.** CLUSTALW alignment of predicted protein sequence for repeating units. This alignment was used to generate Supplemental Figure 7C. **C.** CLUSTALW alignment of DNA sequence for repeating units. This alignment was used to generate Supplemental Figure 7D.



Supplemental Figure 2: PCR with cDNA and gDNA favor amplification of *TIPTOP1-3* **A.** Ethidium bromide stained gel with the resultant PCR reactions using either cDNA or gDNA as template and primers JM377 and JM378 that bind the 5' UTR and 3' UTR of *TIPTOP1-3* (see Cloning of *TIPTOP* genes in Methods section). The main bands are *TIPTOP1-3*. There are some feint bands below, which upon cloning were found to be either *TIPTOP* PCR artifacts or unrelated 'junk' sequences. The *TIPTOP* PCR artifacts were best demonstrated in **B.** where a pGEM-Teasy plasmid with *TIPTOP2* was used for PCR with primers JM439 and JM440 which bind at the ends of *TIPTOP2*. Although it gave a single intense band of the correct size, there was laddering below it with *TIPTOP* PCR artifacts.

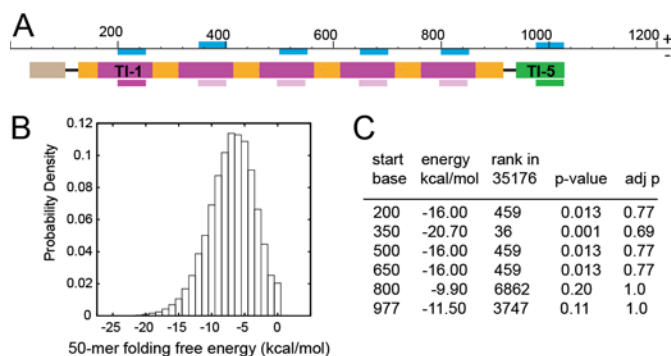


Supplemental Figure 5: Mass spectrometry data for TI-6. **A.** TI-6 is highlighted in the *M. cochinchinensis* LC-MS profile. Inset in this panel is the TI-6 sequence with the MS/MS coverage we show in lower panels marked upon the sequence. **B.** An analytical HPLC trace of purified TI-6 with > 95% purity. This sample was used for the MS/MS sequencing. **C.** TI-6 was sequenced using MS/MS. After reduction and digestion of TI-6, several fragments were observed and sequenced (please refer to Supplemental Table 1 for more details on digested fragments). Panels **D**, **E**, and **F** show assigned MS/MS spectra for three precursor ions of TI-6, which are 938.23²⁺, 1024.28²⁺, and 736.67²⁺, respectively. Both b- and y-series ions were followed and the mass difference between two ions was used to deduce the sequence.

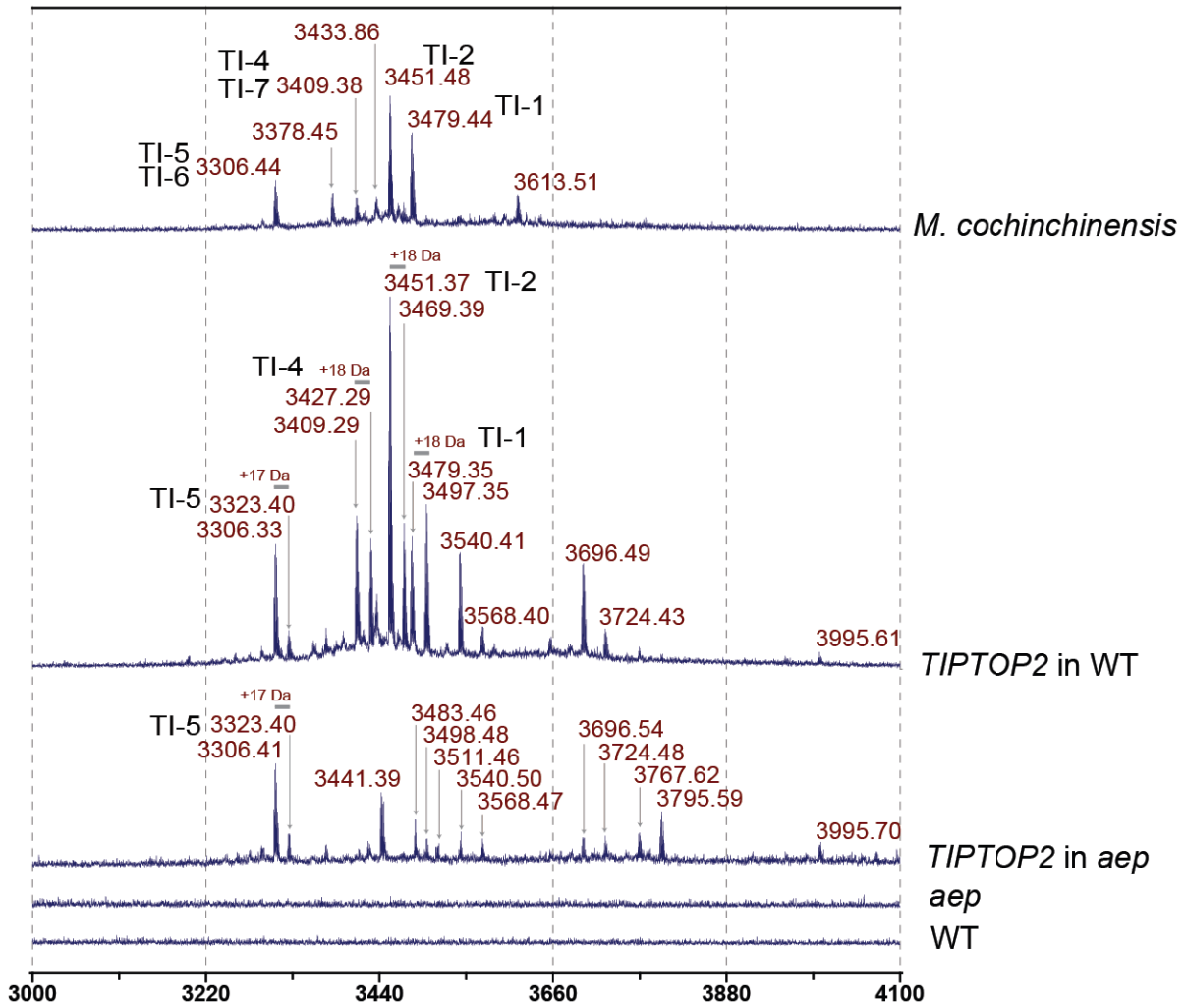


Supplemental Figure 7: Analysis of the *TIPTOP1-3* repeats. **A.** Labeling scheme for *TIPTOP1-3* repeats. **B.** Unrooted phylogram for the *TIPTOP* DNA repeats. The alignment used to produce this phylogeny is included as Supplemental Table 3. The first peptide domains from *TIPTOP1*, *TIPTOP2* and *TIPTOP3* are more closely related to each other than to the rest and this is also the case for the last peptide domain from the three genes. This suggests they have a common ancestor so these three genes are likely to be the result of duplication after they possessed an expanded structure. The remaining peptide domains are too closely related to each other to separate them confidently. **C.** A summary of the DNA sequence differences between neighboring *TIPTOP* repeats. Using the DNA alignment in Supplemental Figure 1C, differences were counted. Neighbors were joined with lines denoting the number of differences and the line thickness correlating to similarity (thickest lines for

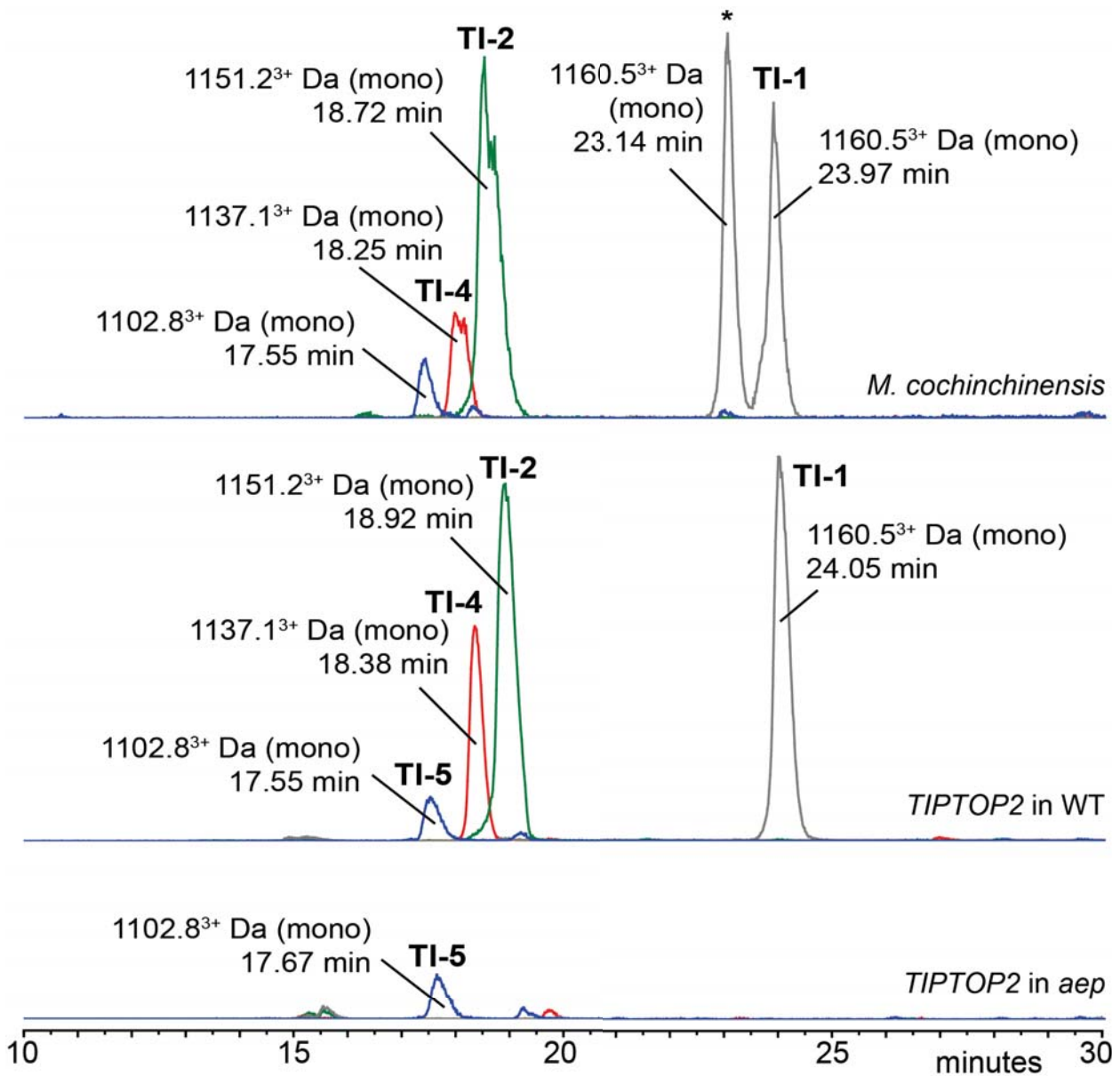
no differences). **D.** As for C above but using the protein alignment in Supplemental Figure 1B. These two analyses show the TI-2 encoding repeats are more closely related to each other than the terminal knottin domains, the “A” domains (encode TI-8, TI-1), TIPTOP2E (TI-4) and TIPTOP3G (TI-7).



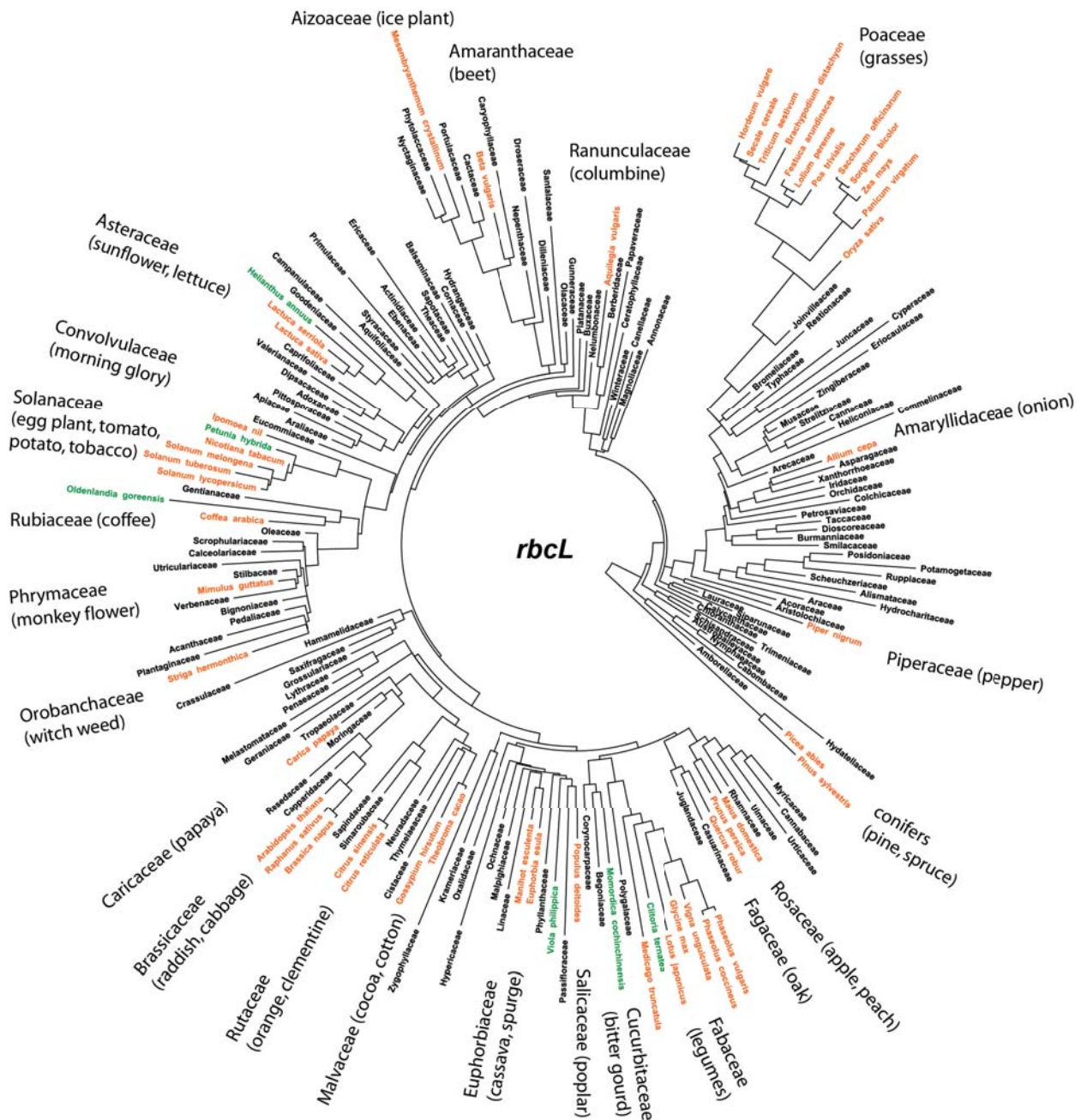
Supplemental Figure 8: The statistical analysis of the folding free energy of 50-mers indicates the imperfect palindromes in *TIPTOP2* are not significant. **A.** As for Figure 4A, a reconstruction of the MEME raw output, showing the location of 50 base repeats found on the sense (+) and minus (-) strand. Below the MEME output, the equivalent regions are marked on the *TIPTOP2* protein schematic. **B.** A histogram displaying the empirical probability of 50-mers with a given folding free energy, estimated using 35,176 random 50-mers extracted from unspliced *Arabidopsis* mRNA. **C.** A summary of the folding free energies and statistical significance of the 50-mers shown in panel A. The p-value is the area under the histogram corresponding to free energies greater than or equal to the given value; the adjusted p-value (adj p) is adjusted for 1,187 multiple tests - the number of places in the *TIPTOP2* sequence where a 50-mer palindrome could start.



Supplemental Figure 9: A wider mass range for the MALDI spectra shown ‘zoomed in’ for Figure 5B. Where known, the identity of ions has been included. In “*TIPTOP2* in WT” *Arabidopsis* and “*TIPTOP2* in *aep*” null mutant background there are higher molecular weight ions that are clearly *TIPTOP2*-derived based on their absence from WT and *aep*, but as they appear in both backgrounds they are AEP-independent (i.e. 3540, 3568, 3696, 3724 and 3995). These ions are likely to represent *TIPTOP2* mis-processing events as they are not detected in *M. cochinchinensis*.



Supplemental Figure 10: Fully labeled LC-MS profile of peptide extracts from *TIPTOP2*-expressing *Arabidopsis* extracted for ions within the ranges 827.2-827.4, 853.0-853.2, 863.5-863.7 and 870.5-870.7 Da. Each extracted ion count is fully labeled with its retention time and the monoisotopic mass of the triply charged ion. The asterisked peak with an identical mass, but earlier retention time to that of TI-1 we suspect is a TI-1 isomer with a iso-aspartyl bond. Iso-aspartyl bonds for these peptides have been observed previously (Hernandez et al., 2000).



Supplemental Figure 11: Angiosperm phylogeny (best Maximum likelihood tree produced using RAxML (Stamatakis et al., 2008) based on *rbcL* sequences. Species known to contain AEP-mediated cyclic peptides and their precursors are in green. A range of model plants are included and their names are highlighted in orange and their family added (with common names in brackets). For taxa representing each family see Methods. The alignment used to produce this phylogeny is included as Supplemental Table 4.

Supplemental Table 1: MS/MS product ions for TI-4, TI-5, TI-6, and TI-8.

peptide	peptide sequence	enzyme ^a	theoretical mass (Da)	experimental mass (Da)	Δ mass (Da) ^b
TI-4	D14-R6	TE	2534.89	2534.84	0.05
	R13-R6	TE	2690.99	2690.97	0.02
	R12-R6	TE	2950.11	2950.20	0.09
	C29-L8	CE	1291.56	1291.52	0.04
	Y28-K9	CE	2158.91	2158.86	0.05
TI-5	D14-G30	TE	1719.59	1719.33	0.26
	C11-G30	TE	2134.81	2134.78	0.03
	Z1-L8	CE	957.04	957.43	0.39
	Z1-E20	CE	2330.64	2330.34	0.30
	K9-G30	CE	2391.00	2390.88	0.12
TI-6	Z1-K10	TE	1471.33	1471.34	0.01
	A3-E20	TE	2045.99	2046.56	0.57
	R13-G30	TE	1875.66	1875.46	0.20
	D14-G30	TE	1719.56	1719.36	0.20
	K9-E20	CE	1392.62	1392.66	0.04
	K9-Y28	CE	2230.93	2230.72	0.21
TI-8	S15-K6	TE	2414.99	2415.30	0.31
	I7-D14	TE	1058.58	1058.58	0.00
	G25-L8	CE	1682.74	1682.92	0.18
	Q9-Y28	CE	2185.92	2185.38	0.54

^a Enzymes used included trypsin/endo-GluC(TE) and chymotrypsin/endo-GluC (CE).

^b The absolute difference between the theoretical and experimental mass.

Supplemental Table 2: NMR and refinement statistics for MCoTI-V.

NMR distance & dihedral constraints	
Distance constraints	
Total NOE	310
Intra-residue	83
Sequential ($ i-j = 1$)	52
Medium-range ($ i-j < 4$)	51
Long-range ($ i-j > 5$)	124
Total dihedral angle restraints	12
Structure Statistics	
Violations (mean and s.d.)	
Distance constraints (Å)	0.02 ± 0.003
Dihedral angle constraints (°)	0.2 ± 0.18
Max. dihedral angle violation (°)	3
Max. distance constraint violation (Å)	0.3

Deviations from idealized geometry	
Bond lengths (Å)	0.004 ± 0.0001
Bond angles (°)	0.84 ± 0.01
Impropers (°)	0.48 ± 0.04
Average pairwise r.m.s.d.** (Å)	
Backbone	0.59 ± 0.18
Heavy	1.51 ± 0.28
Backbone (residues 4-28)	0.35 ± 0.09
Heavy (residue 4-28)	1.26 ± 0.27
Ramachandran statistics	
% in most favoured region	85.9
% in additionally allowed region	14.1
% in generously allowed region	0

**Pairwise r.m.s.d. was calculated among 20 refined structures.

Supplemental Table 3: Primer sequences used in this study. JM439 adds a *Cl*I site (atcgat underlined) and the translation initiation sequence ACA to the *TIPTOP2* start ATG. JM440 binds the end of the *TIPTOP2* 3' UTR and adds a *B*amHI site (ggatcc, underlined).

Primer	Sequence
PN02	5'-GGA TCC AYG GNG GNG TNT GYC CAN AR-3'
PN10	5'-GGA TCC ANC CRT TNC CNC KRC ADA TR-3'
PN03	5'-GGA TCC GNG TNT GYC CAN ARA THY TNA AR-3'
PN11	5'-GGA TCC TTN CCN CKR CAD ATR CAN GC-3'
JM368	5'-GTC GGA CAC GAG GCC TTC TA-3'
JM369	5'-GAC GGT CGA GCC CAA ATC-3'
JM371	5'-GAC AGT CGA GCC CAA ATC-3'
JM429	5'-AAA ATA AAC AAG GAA GAA AAC GTC TTG CTA GAG A-3'
JM430	5'-GAA AAC AAC ACT CAT ATT CTC ACT TT-3'
JM437	5'-ATG GAG AGC AAG AAG ATT-3'
JM438	5'-ATT CTA GGA CAG GCT CTT T-3'
JM439	5'- aa <u>atc gat</u> ACA ATG GAG AGC AAG AAG ATT CTT CCG-3'
JM440	5'- <u>ttg gat ccG</u> AAA ACA ACA CTC ATA TTC TCA CT-3'

Supplemental Table 4: *Cucumis* sequences supporting the protein sequence for Melon1. The protein sequence for Melon1 is supported by 57 *Cucumis melo* ESTs

JG526994	JG504732	JG523175	JG518635	JG502482	JG506590	JG540879	JG537077	JG528501
JG527549	JG520687	JG519886	JG504089	JG501438	JG501036	JG523246	JG534118	JG521353
JG519372	JG504902	JG503400	JG503193	JG500645	JG532717	JG533693	JG531611	JG520293
JG507328	JG507118	JG505260	JG502812	JG501595	JG503002	JG534791	JG529309	JG528525
JG525736	JG503058	JG499587	JG501979	JG532768	JG531254	JG522404	JG506463	JG506052
JG502123	JG504184	JG521870	JG506746	JG506193	JG523317	JG530344	JG530725	JG523759
JG533556	JG506444	JG506421						

Supplemental Table 5. *Cucumis* sequences supporting the protein sequence for Melon2. The protein sequence for Melon2 is supported by 69 *Cucumis melo* ESTs:

JG532730	JG506147	JG552011	JG551982	JG536123	JG536702	JG531542	JG531354	JG524923
JG524465	JG522249	JG507111	JG506111	JG503182	JG505033	JG518638	JG536437	JG534533
JG530290	JG529188	JG528022	JG525518	JG525064	JG520715	JG520085	JG507346	JG507299
JG505266	JG504709	JG502990	JG530476	JG519852	JG506058	JG527433	JG504641	JG536676
JG524872	JG505758	JG507234	JG507453	JG549720	JG532869	JG525467	JG523301	JG506333
JG534526	JG536539	JG526351	JG505548	JG504701	JG503423	JG528531	JG521615	JG504606
JG522141	JG505680	JG551995	JG482657	JG551764	JG503023	JG490786	JG481332	JG551800
JG489241	JG488871	JG487098	JG486299	JG489655	JG488612			

Supplemental Table 6. Species used for placement of families for angiosperm phylogeny in the following format: Family: Species (can be >1 species per family where one family has several model plants).

Acanthaceae: <i>Justicia americana</i>	Acoraceae: <i>Acorus americanus</i>
Actinidiaceae: <i>Actinidia chinensis</i>	Adoxaceae: <i>Adoxa moschatellina</i>
Aizoaceae: <i>Mesembryanthemum crystallinum</i>	Alismataceae: <i>Alisma plantago aquatica</i>
Amaranthaceae: <i>Beta vulgaris</i>	Amaryllidaceae: <i>Allium cepa</i>
Amborellaceae: <i>Amborella trichopoda</i>	Annonaceae: <i>Annona senegalensis</i>
Apiaceae: <i>Apium graveolens</i>	Aquifoliaceae: <i>Ilex repanda</i>
Araceae: <i>Arum maculatum</i>	Araliaceae: <i>Aralia spinosa</i>
Arecaceae: <i>Phoenix canariensis</i>	Aristolochiaceae: <i>Aristolochia promissa</i>
Asparagaceae: <i>Asparagus officinalis</i>	Asteraceae: <i>Helianthus annuus</i> , <i>Lactuca sativa</i> , <i>Lactuca serriola</i>
Austrobaileyaceae: <i>Austrobaileya scandens</i>	Balsaminaceae: <i>Impatiens amplexicaulis</i>
Begoniaceae: <i>Begonia metallica</i>	Berberidaceae: <i>Berberis thunbergii</i>
Bignoniaceae: <i>Tecomaria capensis</i>	Brassicaceae: <i>Arabidopsis thaliana</i> , <i>Brassica napus</i> , <i>Raphanus sativus</i>
Bromeliaceae: <i>Bromelia plumieri</i>	Burmanniaceae: <i>Thismia rodwayi</i>
Buxaceae: <i>Buxus sempervirens</i>	Cabombaceae: <i>Cabomba caroliniana</i>
Cactaceae: <i>Pereskia aculeata</i>	Calceolariaceae: <i>Calceolaria spec</i>
Calycanthaceae: <i>Calycanthus floridus</i>	Campanulaceae: <i>Campanula trachelium</i>
Canellaceae: <i>Canella winterana</i>	Cannabaceae: <i>Cannabis sativa</i>
Cannaceae: <i>Canna indica</i>	Capparaceae: <i>Capparis spinosa</i>
Caprifoliaceae: <i>Lonicera orientalis</i>	Caricaceae: <i>Carica papaya</i>
Caryophyllaceae: <i>Stellaria media</i>	Casuarinaceae: <i>Casuarina equisetifolia</i>
Ceratophyllaceae: <i>Ceratophyllum demersum</i>	Chloranthaceae: <i>Chloranthus japonicus</i>
Cistaceae: <i>Cistus crispus</i>	Colchicaceae: <i>Colchicum speciosum</i>
Commelinaceae: <i>Commelina communis</i>	Convolvulaceae: <i>Ipomoea nil</i>
Cornaceae: <i>Cornus mas</i>	Corynocarpaceae: <i>Corynocarpus laevigata</i>
Crassulaceae: <i>Crassula marnierana</i>	Cucurbitaceae: <i>Momordica cochinchinensis</i>
Cyperaceae: <i>Rhynchospora nervosa</i>	Dilleniaceae: <i>Dillenia indica</i>
Dioscoreaceae: <i>Dioscorea polygonoides</i>	Dipsacaceae: <i>Dipsacus sativus</i>
Droseraceae: <i>Drosera spatulata</i>	Ebenaceae: <i>Diospyros virginiana</i>
Ericaceae: <i>Erica tetralix</i>	Eriocaulaceae: <i>Eriocaulon microcephalum</i>
Eucommiaceae: <i>Eucommia ulmoides</i>	Euphorbiaceae: <i>Euphorbia esula</i> , <i>Manihot esculenta</i>
Fabaceae: <i>Clitoria ternatea</i> , <i>Vigna unguiculata</i> ,	Fagaceae: <i>Quercus robur</i>

- Phaseolus vulgaris*, *Phaseolus coccineus*, *Glycine max*,
Lotus japonicus, *Medicago truncatula*
Gentianaceae: *Gentiana verna*
Goodeniaceae: *Scaevola frutescens*
Gunneraceae: *Gunnera manicata*
Heliconiaceae: *Heliconia latispatha*
Hydrangeaceae: *Hydrangea macrophylla*
Hypericaceae: *Hypericum perforatum*
Joinvilleaceae: *Joinvillea ascendens*
Juncaceae: *Juncus effusus*
Lauraceae: *Cinnamomum camphora*
Lythraceae: *Punica granatum*
Malpighiaceae: *Malpighia glabra*
Melastomataceae: *Clidemia rubra*
Musaceae: *Musella lasiocarpa*
Nelumbonaceae: *Nelumbo lutea*
Neuradaceae: *Neurada procumbens*
Nymphaeaceae: *Nuphar lutea*
Olacaceae: *Heisteria concinna*
Orchidaceae: *Apostasia stylidioides*
Oxalidaceae: *Oxalis dillenii*
Passifloraceae: *Passiflora biflora*
Penaeeae: *Olinia emarginata*
Phrymaceae: *Mimulus guttatus*
Phytolaccaceae: *Phytolacca americana*
Pinaceae: *Pinus sylvestris*
Pittosporaceae: *Pittosporum japonicum*
Platanaceae: *Platanus occidentalis*
- Polygalaceae: *Polygala cruciata*
Posidoniaceae: *Posidonia oceanica*
Primulaceae: *Lysimachia azorica*
Resedaceae: *Reseda alba*
Rhamnaceae: *Rhamnus cathartica*
Rubiaceae: *Coffea arabica*, *Oldenlandia goreensis*
Rutaceae: *Citrus reticulata*, *Citrus sinensis*
Santalaceae: *Viscum album*
Sapotaceae: *Manilkara zapota*
Scheuchzeriaceae: *Scheuchzeria palustris*
Scrophulariaceae: *Verbascum thapsus*
Siparunaceae: *Siparuna brasiliensis*
Solanaceae: *Petunia hybrida*, *Solanum lycopersicum*,
Solanum tuberosum, *Solanum melongena*, *Nicotiana
tabacum*
Strelitziaceae: *Strelitzia nicolai*
- Geraniaceae: *Geranium cinereum*
Grossulariaceae: *Ribes aureum*
Hamamelidaceae: *Hamamelis virginiana*
Hydatellaceae: *Trithuria submersa*
Hydrocharitaceae: *Vallisneria spec*
Iridaceae: *Gladiolus buckerveldii*
Juglandaceae: *Juglans nigra*
Krameriaceae: *Krameria lanceolata*
Linaceae: *Linum perenne*
Magnoliaceae: *Magnolia kobus*
Malvaceae: *Theobroma cacao*, *Gossypium hirsutum*
Moringaceae: *Moringa oleifera*
Myricaceae: *Morella cerifera*
Nepenthaceae: *Nepenthes alata*
Nyctaginaceae: *Bougainvillea glabra*
Ochnaceae: *Quiina pteridophylla*
Oleaceae: *Olea europaea*
Orobanchaceae: *Striga hermonthica*
Papaveraceae: *Hypecoum imberbe*
Pedaliaceae: *Sesamum indicum*
Petrosaviaceae: *Petrosavia stellaris*
Phyllanthaceae: *Phyllanthus flexuosus*
Piceaceae: *Picea abies*
Piperaceae: *Piper nigrum*
Plantaginaceae: *Plantago lanceolata*
Poaceae: *Oryza sativa*, *Panicum virgatum*, *Zea mays*,
Saccharum officinarum, *Sorghum bicolor*, *Festuca
arundinacea*, *Lolium perenne*, *Poa trivialis*, *Triticum
aestivum*, *Hordeum vulgare*, *Secale cereale*,
Brachypodium distachyon
Portulacaceae: *Portulaca grandiflora*
Potamogetaceae: *Potamogeton perfoliatus*
Ranunculaceae: *Aquilegia vulgaris*
Restionaceae: *Restio tetraphyllus*
Rosaceae: *Malus domestica*, *Prunus persica*
Ruppiaceae: *Ruppia maritima*
Salicaceae: *Populus deltoides*
Sapindaceae: *Koelreuteria paniculata*
Saxifragaceae: *Saxifraga mertensiana*
Schisandraceae: *Illicium parviflorum*
Simaroubaceae: *Ailanthus altissima*
Smilacaceae: *Smilax glauca*
Stilbaceae: *Halleria lucida*
Styracaceae: *Styrax americanus*

Taccaceae: *Tacca chantieri*

Thymelaeaceae: *Thymelaea hirsuta*

Trochodendraceae: *Trochodendron aralioides*

Typhaceae: *Typha latifolia*

Urticaceae: *Urtica dioica*

Valerianaceae: *Valeriana officinalis*

Violaceae: *Viola philippica*

Xanthorrhoeaceae: *Xanthorrhoea hostilis*

Zygophyllaceae: *Zygophyllum sessilifolium*.

Theaceae: *Camellia japonica*

Trimeniaceae: *Trimenia moorei*

Tropaeolaceae: *Tropaeolum majus*

Ulmaceae: *Ulmus alata*

Utriculariaceae: *Utricularia biflora*

Verbenaceae: *Verbena officinalis*

Winteraceae: *Drimys winteri*

Zingiberaceae: *Zingiber gramineum*

Cyclic Peptides Arising by Evolutionary Parallelism via Asparaginyl-Endopeptidase-Mediated Biosynthesis

Joshua S. Mylne, Lai Yue Chan, Aurelie H. Chanson, Norelle L. Daly, Hanno Schaefer, Timothy L. Bailey, Philip Nguyencong, Laura Cascales and David J. Craik
Plant Cell; originally published online July 20, 2012;
DOI 10.1105/tpc.112.099085

This information is current as of August 7, 2012

Supplemental Data	http://www.plantcell.org/content/suppl/2012/07/03/tpc.112.099085.DC1.html
Permissions	https://www.copyright.com/ccc/openurl.do?sid=pd_hw1532298X&issn=1532298X&WT.mc_id=pd_hw1532298X
eTOCs	Sign up for eTOCs at: http://www.plantcell.org/cgi/alerts/ctmain
CiteTrack Alerts	Sign up for CiteTrack Alerts at: http://www.plantcell.org/cgi/alerts/ctmain
Subscription Information	Subscription Information for <i>The Plant Cell</i> and <i>Plant Physiology</i> is available at: http://www.aspb.org/publications/subscriptions.cfm