JBC Papers in Press. Published on May 23, 2017 as Manuscript M117.776955
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M117.776955

Two Proteins for the Price of One: Structural Studies of the Dual-Destiny Protein Preproalbumin with Sunflower Trypsin Inhibitor-1

Bastian Franke[‡], Amy M. James[§], Mehdi Mobli[¶], Michelle L. Colgrave[#], Joshua S. Mylne[§], K. Johan Rosengren[‡]

From The University of Queensland, [‡]School of Biomedical Sciences and [¶]Centre for Advanced Imaging, St Lucia, Brisbane, QLD 4072, Australia; [§]The University of Western Australia, School of Molecular Sciences & The ARC Centre of Excellence in Plant Energy Biology, Crawley, Perth, WA 6009, Australia; [#]CSIRO Agriculture and Food, St Lucia, Brisbane, QLD 4067, Australia

Running title: Structure and processing of PawS1

To whom correspondence should be addressed:
K. Johan Rosengren, The University of Queensland, School of Biomedical Sciences, St Lucia, Brisbane, QLD 4072, Australia; Telephone: +61 7 3365 1403; E-mail: j.rosengren@uq.edu.au

Keywords: Preproalbumin with SFTI-1 (PawS1), Sunflower Trypsin Inhibitor-1 (SFTI-1), Asparaginyl Endopeptidase (AEP), Protein processing, Seed storage albumin

ABSTRACT

Seed storage proteins are both an important source of nutrition for humans and essential for seedling establishment. Interestingly, unusual napin-type 2S seed storage albumin precursors in sunflower contain a sequence that is released as a macrocyclic peptide during post-translational processing. The mechanism by which such peptides emerge from linear precursors proteins has received increased attention, however the structural characterization of intact precursor proteins has been limited. Here we report the 3D NMR structure of the *Helianthus annuus* PreproAlbumin With Sunflower trypsin inhibitor-1 (PawS1), and provide new insights into the processing of this remarkable dual-destiny protein. In seeds PawS1 is matured by asparaginyl endopeptidases (AEP) into the cyclic peptide Sunflower Trypsin Inhibitor-1 (SFTI-1) and a heterodimeric 2S albumin. The structure of PawS1 revealed that SFTI-1 and the albumin are independently folded into well-defined domains separated by a flexible linker. PawS1 was cleaved *in vitro* with recombinant sunflower HaAEP1 and *in situ* using a sunflower seed extract in a way that resembled the expected *in vivo* cleavages. Recombinant HaAEP1 cleaved PawS1 at multiple positions and *in situ* its flexible linker was removed yielding fully mature heterodimeric albumin. Liberation and cyclization of SFTI-1, however, was inefficient suggesting specific seed conditions or components may be required for *in vivo* biosynthesis of SFTI-1. In summary, this study has revealed the 3D structure of a macrocyclic precursor protein and provided important mechanistic insights into the maturation of sunflower proalbumins into an albumin and a macrocyclic peptide.

Endoproteases are well-known for their role in protein degradation, but are also intimately involved in protein biosynthesis, where they mature sequences by releasing flanking fragments or separating subunits by internal cleavage events (1,2). An abundant class of proteins that are subject to substantial proteolytic maturation are seed storage proteins, which are both an important source of nutrition for humans and essential for the establishment of the seedling (3). Seed storage proteins are matured from precursors that are sent to the endoplasmic reticulum for folding and removal of their N-terminal signal sequence, before they are subject to one or more internal cleavages often followed by trimming to yield highly stable and long lasting proteins that can withstand desiccation (4,5). Several classes exist, with the major ones being the salt-soluble globulins, alcohol-soluble prolamins and water-soluble albumins.

The prototypic seed albumin is *Brassica napus* (canola) napin, which begins as a 178-residue protein that exits the endoplasmic reticulum as a 157 amino acid protein. It is further processed through the removal of two joining peptides and the C-terminal

residue by the action of endoproteases called asparaginyl endopeptidases, yielding the "mature" napin. In this final form napin exists as a heterodimer with two chains of 29 and 86 residues (6) that are cross braced by four disulfide bonds, which are conserved throughout the 2S albumin family. Some napin-type albumins are not processed into heterodimers, but remain as monomers, such as the abundant SESA3 (SFA8), SESA2-1, SESA2-2 and SESA20-2 from sunflower seeds (7,8). Whether they are matured into a heterodimer or remain monomeric, napin-type albumins adopt a conserved five-helix tertiary structure stabilized by the characteristic disulfide array (5).

The precursor proteins for napin-type albumins in the common sunflower (*Helianthus annuus*) are diverse. Some precursor proteins consist of two full albumin sequences instead of one (9), but one class in particular is unusual for the fact in addition to an albumin, they contain a macrocyclic peptide (10). *H. annuus* possesses two of these unusual albumin precursor genes encoding buried peptides. Preproalbumin with SunFlower Trypsin Inhibitor-1 (SFTI-1) named PawS1 and a closely related PawS2 are matured into typical heterodimeric albumins with one small and one large subunit, but also yield the 14-residue SFTI-1 (Figure 1) and the 12-residue SFT-L1, respectively (10). Both peptides are head-to-tail peptide macrocycles with potential applications in drug design (11).

This proteinaceous duality is ancient; *PawS1* genes and the corresponding peptide evidence found in related species infer that this unusual class of albumin precursor with PawS-Derived Peptides is over 20 million years old (12). The mechanism by which macrocyclic peptides emerge from linear precursors is the subject of much interest (13,14). The maturation of SFTI-1 has been studied with short synthetic peptide precursors using crude extracts of sunflower (15). Recombinantly produced asparaginyl endopeptidases (AEPs) also allow the reaction to be reconstituted *in vitro* (15). The proposed model for SFTI-1 is that the most highly expressed seed AEP, HaAEP1, cleaves an 18-residue peptide SFTI-GLDN from the PawS1 albumin. This 18-residue peptide is then subjected to a cleavage-coupled intramolecular trans-peptidation reaction, where after cleavage of the GLDN tail the amino-terminus reacts with the acyl intermediate, resulting in a 14-residue macrocycle and the release of the 4-residue GLDN tail (15).

Biosynthesis involving AEP-mediated macrocyclization is shared by other plant macrocyclic peptides, such as the much larger cyclotides or cyclic knottins (16), and accordingly AEPs that are highly efficient at macrocyclizing various synthetic substrates have been discovered and characterized from species containing such macrocycles (13,14). Gene-encoded macrocyclic peptides are not only found in plants, but throughout the kingdom of life, including bacteria, fungi and mammals (17). These peptides vary greatly in size from ~6 to ~70 amino acids and although they are all produced through post-translational processing of precursors, the nature of these precursors and the processing machineries required are diverse. Despite the interest in these systems the significance of the structural context in which the cyclic peptide sequence is embedded is poorly understood, and the structural characterization of intact precursor proteins for cyclic peptide families has been limited.

Here we present the solution NMR structure of the PawS1 proalbumin and use both recombinant sunflower HaAEP1 and crude sunflower seed extract to illustrate the maturation of PawS1 *in vitro*. We provide mechanistic insights into the structures of proalbumins and the consequences of the AEP cleavages, as well as the initial steps of biosynthesis of the ultrastable macrocycle SFTI-1.

Results

To study the structural features and the processing of the unusual albumin precursor PawS1 a synthetic gene that encoded the full SFTI-1 and albumin domains was designed. The construct consists of the C-terminal 116 residues of the 151-residue preproprotein, starting from Gly1 of SFTI-1 (Gly36 in full length PawS1) and ending with the C-terminal Ile of the PawS1 albumin. PawS1 was expressed successfully in the *E. coli* strain SHuffle and purified with yields of up to 150 µg/L soluble ^{15}N labeled or ^{13}C and ^{15}N double labeled protein obtained. The purity of PawS1 was assessed by reverse phase-high-performance liquid chromatography (RP-HPLC) and liquid chromatography-mass spectrometry (LC-MS) revealing ~95% purity (Figure 2). The average calculated mass for the oxidized double-labeled protein was 13,987 Da and the observed average mass for the protein based on LC-MS was 13,943 Da, representing ~94% incorporation of ^{15}N and ^{13}C.

Resonance assignment and structure determination

Samples for solution NMR studies were prepared to concentrations of 3.6 mg/mL and all NMR data recorded at 25°C. The ^{1}H-^{15}N-HSQC spectrum showed excellent peak dispersion and sharp lines,

indicating a highly structured protein (Figure 3). A suite of standard triple resonance three dimensional (3D) experiments, including HNCACB, CBCA(CO)NH, HNCO, H(CC)(CO)NH-TOCSY, and 3D (H)CC(CO)NH-TOCSY were acquired and allowed complete sequential assignment of the protein backbone, except the amide of Gln98, and majority of side chain atoms. The side chain amide groups of the 26 glutamine residues were not assigned due to the severe peak overlap in the ^{1}H-^{15}N-HSQC spectrum (Figure 3). Assignments around the proline residues Pro8, Pro9, Pro13, Pro19, Pro50 and Pro107, which lack backbone amide protons, were confirmed using 3D ^{15}N and ^{13}C edited NOESY data.

The NMR secondary shifts (Figure 3) of the active site loop of SFTI-1, including the characteristic downfield shifted Hα protons of Cys3 and Cys11 resulting from the conformation of the disulfide across the β-sheet, were found to be consistent with those reported previously for native SFTI-1 (18). In contrast, some variations are seen at positions Gly1, Arg2, Phe12, Pro13 and Asp14. Structural changes around these residues are expected, as they are part of the loop where Gly1 and Asp14 are joined by the trans-peptidation reaction that leads to the head-to-tail cyclization of SFTI-1. The peptide bond proceeding Pro8 was confirmed to be in a *cis* configuration based on NOE patterns (19) and ^{13}C chemical shifts, like in SFTI-1, while all other proline residues were found to be in a *trans* conformation. Throughout the albumin domain the majority of Hα resonances are upfield shifted, and large stretches of negative secondary shift are consistent with extensive helical structure (Figure 3).

To determine the solution structure of PawS1, structural restraints were obtained from the NMR data. Inter-proton distances were generated from NOE intensities in 3D ^{15}N and ^{13}C edited NOESY spectra and backbone dihedral angle phi (φ), psi (ψ) and the side chain torsion angle chi1 ($\chi 1$) were derived from a TALOS-N analysis of the HN, Hα, Cα, Cβ, CO and N chemical shifts (20). In addition, deuterium exchange experiments were conducted using ^{15}N labeled protein to identify hydrogen-bonded amide protons. Where hydrogen bond acceptors could be identified for slow exchanging amide protons during the structure calculations, hydrogen bond restraints were also included. Initial calculations involved automatic assignment of the NOESY data and generation of structures using torsion angle dynamics within the program CYANA (21). In the final round of structure calculation a set of 50 structures was generated and refined in water within CNS (22), and the 20 structures with lowest energy were chosen to represent the solution structure of PawS1. Structural statistics for the ensemble demonstrate that the structure is of high quality and in good agreement with both experimental data and covalent geometry, as assessed by MolProbity (23) (Table 1).

Structural analysis

The 116-residue PawS1 structure is characterized by two structural entities separated by a flexible linker (Figure 4). The N-terminal entity is the 14-residue peptide SFTI-1 in which the active site loop has retained the shape of native SFTI-1 and the small β-sheet is conserved. This is followed by the four-residue linker peptide GLDN, which connects the SFTI-1 domain to the albumin domain. The linker does not adopt a preferred single conformation in solution, thus the relative positions of the SFTI-1 and albumin domains differ greatly within the structural ensemble (Figure 4). The albumin structure consists of four helices that are closely packed in a tertiary fold creating an extensive hydrophobic core. The small subunit (SSU) contains one helical segment (helix I – residues Ile26 to Thr37), which is arranged anti-parallel to a second helix (helix II – residues Pro50 to Glu65) located in the large subunit (LSU). The unstructured loop region comprising residues Thr37-Asn49 between helix I and helix II contain the seven-residue linker peptide LRMAVEN that link the two subunits at the precursor stage. The LSU contains two additional helices (helix III – residues Gln71 to Gln87 and helix IV – residues Gln94 to Gln109), which are of similar length and lie anti-parallel to each other, but across the face of helix I and helix II at an angle of about 45º. The hypervariable region, which varies in length and sequence amongst the 2S albumins, is located between helix III and helix IV, and in PawS1 comprise residues Gln88-Gln93.

PawS1 contains five disulfide bonds of which one is located in the SFTI-1 domain (Cys3-Cys11), where it is a key feature in stabilizing the β-sheet. The remaining four disulfides are located in the albumin domain. Two inter-chain disulfides connect the SSU and LSU (Cys22-Cys70 and Cys32-Cys59) and two intra-chain disulfides cross-brace the LSU (Cys60-Cys110 and Cys72-Cys114). This cysteine connectivity pattern, which is highly conserved amongst 2S albumins, is critical for stabilizing the compact fold. Cys22-Cys70 link the N-terminal part of the SSU to a loop between helix II and III, Cys32-Cys59 bridge helix I and II, Cys60-Cys110 bridge helix II and helix IV, and Cys72-Cys114 ties the C-terminus to helix III.

NMR relaxation analysis

As evident from the overlay of the ensemble both domains adopt highly ordered structures throughout most of the protein sequence, but noticeable disorder is seen in the linker regions comprising residues 15-18 (GLDN) and 43-49 (LRMAVEN) (Figure 4). Both linkers are classified as dynamic by TALOS-N based on their chemical shift pattern. To further confirm that these regions are flexible a heteronuclear ^{1}H-^{15}N steady state NOE relaxation experiments was recorded (Figure 5). From here it is clear that the elements of secondary structure are highly ordered with NOE values in the range of 0.8-0.9 consistent with a rigid peptide backbone. In contrast the linker segments show NOEs in the order 0.3-0.5, which is consistent with a substantial increase in flexibility. Notably the entire SFTI-1 domain shows significantly lower heteronuclear NOEs suggesting the linker "decouples" the peptide from the larger protein allowing it to adopt a faster overall motion more like a 1.5 kDa peptide than a 13 kDa protein. These data strongly suggest that not only is the linker flexible, the entire SFTI-1 domain, which is highly ordered, does not adopt a preferred orientation relative to the albumin domain.

***In vitro* digests of PawS1 with sunflower HaAEP1**

Previous processing studies into the biosynthesis of macrocyclic peptides by AEPs have exclusively used small peptides as substrates (13-15). In this study we wanted to investigate the processing of a native precursor substrate (PawS1) using a recombinant AEP. This requires a suitable buffer system that includes a redox system required for AEP activity, without reducing any of the five disulfide bonds in the protein, which would prevent successful maturation. In initial experiments using AEP preferred conditions with DTT as a reducing agent PawS1 disulfide bonds were found to be reduced by MS analysis. We instead chose milder conditions with an AEP activity buffer containing 0.6 mM/0.4 mM glutathione/glutathione-disulfide (24), where no reduction was observed by MS. To also rule out disulfide shuffling we compared two-dimensional ^{1}H-^{15}N-HSQC NMR spectra of PawS1 dissolved in 90% water /10% D_2O and PawS1 dissolved in 90% AEP activity buffer /10% D_2O over a two-week timeframe. No significant peak shifts or appearance of new peaks were observed in the ^{1}H-^{15}N-HSQC NMR spectra, confirming that the native structure was retained and no substantial disulfide shuffling or degradation due to the mild redox conditions occurred.

To demonstrate the albumin processing events we used this buffer system and carried out digests of PawS1 with recombinantly produced HaAEP1. When PawS1 was incubated with HaAEP1 at 37°C, it produced a product with a mass-to-charge ratio (*m/z*) of $1{,}930.9^{1+}$ corresponding to the SFTI-1 peptide with a tetrapeptide tail, herein described as SFTI-GLDN (Figure 6A). Once SFTI-GLDN was released from its precursor, a set of peaks at $11{,}344.3^{1+}$ was observed, corresponding to the albumin domain of PawS1 (Figure 6B). The mass shift of +18 Da indicated that HaAEP1 also cleaved the albumin domain between the two subunits, as expected to occur after the linker peptide LRMAVEN at Asn49 (Figure 6B). None of the masses appeared in the no-enzyme control indicating that these were the result of HaAEP1-dependent cleavage. Notably the majority of PawS1 remained intact after 92 h (Figure 6C), indicating a slow enzymatic reaction. To ensure that this was not due to low activity of the enzyme, the enzymatic reaction rate was tested using the small SFTI(D14N)-GL peptide substrate and found to be fast, as previously reported (15). Within 2 h the peptide SFTI(D14N)-GL was fully cleaved at Asn14 to produce acyclic-SFTI(D14N). Thus, the processing reactions are much slower in the bulky and sterically hindered PawS1 substrate.

Some non-productive events unrelated to the physiological processing were also observed under the experimental conditions. After 92 h, masses were observed at $12{,}712.5^{1+}$ and $12{,}730.1^{1+}$ in the high mass MALDI TOF-MS spectrum, corresponding to cleavage at the C-terminal Asn111 of PawS1. This cleavage was also noted to be HaAEP1-dependent as no corresponding masses were observed in the no-enzyme control and was concluded to be caused by partial reduction of the disulfide bond Cys72-Cys114 due to the redox system, followed by HaAEP1-mediated cleavage at Asn111. The mass shift of +18 Da suggests additional HaAEP1-mediated cleavage after the linker peptide LRMAVEN between the two subunits was also occurring. Furthermore, after 5 h, a signal appeared at $1{,}531.6^{1+}$ in the low mass MALDI TOF-MS spectrum of the no-enzyme control of PawS1, which corresponds to acyclic-SFTI. This reaction is a result of spontaneous hydrolysis at the Asp-Gly bond at low pH. Evidence for the reaction can also be seen in the high mass spectrum at *m/z* $11{,}743.9^{1+}$, which corresponds to PawS1 still carrying the linker peptide GLDN, but having lost the 14-residue SFTI-1 segment. Similar events were previously seen using the small AEP substrates (15).

In situ digests of PawS1 with sunflower extract

To further explore these cleavage events, and any subsequent processing relying on other enzymes, PawS1 was incubated *in situ* with a sunflower seed extract and the enzymatic reactions were monitored using mass spectrometry. After 2 h, a mixture of intact PawS1 with an *m/z* of 13,255.9^{1+} and a PawS1 variant that was cleaved after the linker peptide LRMAVEN, resulting in a mass shift of +18 Da (13,274.5^{1+}) was observed in the high mass MALDI TOF-MS spectrum. After 5 h, the reaction neared completion with the majority of the PawS1 starting material cleaved after the linker peptide LRMAVEN (Figure 7A). A peak at *m/z* 10,548.5^{1+} was observed corresponding to the mature PawS1 albumin with the LRMAVEN linker peptide removed and with the cleavage of the N-terminal SFTI-GLDN peptide at the N-terminus (Figure 7B), the cleavage of SFTI-GLDN from PawS1 was again however inefficient. In the low mass MALDI TOF-MS spectrum a low intensity signal was observed at *m/z* 1513.9^{1+}, matching cyclic SFTI-1. However, this was also present in the non-enzymatic control, consistent with SFTI-1 being present in the seed extract. It was thus not possible to confirm if any mature SFTI-1 was produced from the PawS1 protein *in situ*. If indeed SFTI-1 was produced the amount was very limited.

To study the structural consequence of the processing, ^{15}N labeled PawS1 was incubated with sunflower *in situ* extract and the reaction monitored by NMR spectroscopy. The initial AEP mediated cleavage of the linker peptide LRMAVEN, which resulted in the formation of two flexible unrestrained ends, was rapid and after 7 h had gone to completion. This reaction was evident from the movement of resonances originating from residues in the region between Cys32 and Cys59, most notably Ser38, Asp40, Leu43, Ala46, Glu48, which all reappeared in different positions in the ^{15}N-HSQC spectrum (Figure 8). The appearance of Asn49 at a new downfield shift was consistent with it becoming a new C-terminal residue. The reaction continued with the removal of the entire LRMAVEN linker peptide. Previous work has suggested that aspartic proteases are involved in the maturation of albumins and therefore could potentially be the factors responsible for this process (25). The appearance of a yet another new peak with characteristics of a C-terminus (Lys42) after 10 days in the C-terminal region of the ^{15}N-HSQC spectra, coupled with the disappearance of the peaks that moved initially, suggested the removal of this peptide and that the maturation to a heterodimeric albumin PawS1 was complete. However, consistent with the mass spectrometric analysis the second expected processing – the liberation of SFTI-GLDN from the protein was remarkably inefficient. At the less sensitive NMR level no evidence of any processing at the N-terminal region, or consequently the formation of SFTI-1 or variants was observed. Importantly despite the removal of the LRMAVEN linker the appearance of the ^{15}N-HSQC spectra is largely unaffected, confirming that the overall structure is not affected by this processing event.

Discussion

PawS1 – a protein with two faces and two fates

Despite the presence of five disulfide bonds, PawS1 could be expressed using the *E. coli* SHuffle strain into a fully oxidized and highly soluble protein in sufficient amounts for structural studies. The structure has two faces – the first is the albumin domain, which adopts a compact helical fold, rich in surface exposed glutamine and arginine residues. Of the four helical segments in PawS1, helix I was located in the SSU, while the three remaining helical segments were all located in the LSU. The cysteine connectivity is conserved amongst seed storage albumins and features two inter-chain disulfide bonds connecting the small and the large subunit (I-V and II-III) and two intra-chain disulfide bonds stabilizing the large subunit (IV-VII and VI-VIII).

To date, seven seed storage albumin structures have been resolved and deposited in the Protein Data Bank, namely BnIb from *Brassica napus* (oilseed rape) (26), RicC3 from *Ricinus communis* (castor bean) (27), SESA3 (also known as SFA8) from *Helianthus annuus* (sunflower) (28), rproBnIb from *Brassica napus* (29), Ara h 6 from *Arachis hypogaea* (peanut) (30), Mabinlin II from *Capparis masaikai* (31) and Ber e 1 from *Bertholletia excels* (Brazil nut) (32). Of these studied albumins SESA3 is monomeric, whereas the remaining proteins are heterodimeric. Despite this difference the fold is highly conserved. BnIb was produced in a recombinant form, rproBnIB, consisting of a single polypeptide, whereas in native BnIb the linker peptide Ser32-Glu33-Asn34 has been removed by proteolytic cleavage during processing. Despite this difference both proteins was shown to adopt very similar structures that are consistent with the structural architecture of RicC3 and SESA3 (29). Superimposing the albumin domain of our solution NMR structure of PawS1 on the sunflower albumin SESA3 highlight that PawS1 also adopts a similar helical-bundle fold (Figure 9A). However, a notable

difference is present in the small subunit (SSU). In SESA3, and all other structures, the SSU contains two helical segments, whereas PawS1 only contains one extended helix I segment. The N-terminal helix of the SSU in SESA3 is missing in PawS1, probably due to the segment between the cysteine residues in this region being three residues shorter, forcing a more extended conformation. C-terminally to the conserved SSU helical segment, both structures possess an unstructured loop of similar length, but with a different amino acid composition, linking this helix to the first helix of the LSU. In the LSU Helix III is significantly longer in PawS1, while Helix IV is of similar length to the structurally equivalent counterparts in SESA3.

Despite the conserved fold, PawS1 and SESA3 have distinct characters. PawS1 is rich in nitrogen-containing amino acids, including 26 glutamines and six asparagines. In contrast SESA3 is rich in sulfur containing amino acids, including 16 methionines in addition to its eight cysteines. Many of the SESA3 methionine residues are buried in the protein center (28), creating a distinctly different hydrophobic core compared to PawS1, in which the core is dominated by isoleucine, leucine and valine residues. Methionine residues in SESA3 are also found exposed on the surface, and the difference in the number of polar versus hydrophobic residues between PawS1 and SESA3 is reflected in their retention time on RP-HPLC, with SESA3 being the most late eluting of the sunflower albumins (9). What has been referred to as the hypervariable loop region (Asn72-Met79) contains in SESA3 four hydrophobic residues (Met75, Trp76, Ile77 and Met79), which have been shown to create a solvent accessible hydrophobic patch on the surface of the albumin. This combination of hydrophobicity and flexibility might contribute to its good emulsifying properties to form highly stable emulsions with oil/water mixtures (28). The corresponding region of PawS1 (Gln88-Gln93) is more hydrophilic and contains three glycine and three glutamine residues. The 26 glutamine residues in PawS1 are otherwise distributed throughout the entire structural surface.

The second face of PawS1 is the small N-terminal peptide region that becomes SFTI-1. Superimposing the solution structure of SFTI-1 on the SFTI-1 domain of PawS1 reveals an identical structural architecture of the β-sheet and trypsin inhibitory binding loop (Figure 9B). Although no trypsin inhibition assays have been performed with the full length precursor, the identical arrangement of the binding loop in both structures suggests that PawS1 itself may also be able to inhibit trypsin like SFTI-1. The role of the cyclic backbone for the bioactivity of SFTI-1 has been investigated and it was shown that opening the backbone in the cyclization loop only causes only a small decrease in inhibitory activity (18). In acyclic-SFTI the Gly1 and Asp14 termini remained close together and the hydrogen bonding network was largely intact (18). We observe that also in PawS1 the cyclization loop has adopted turn features that mimic the native SFTI-1, despite the lack of a cyclic backbone, and these features bring Gly1 and Asp14 together, which may be important for the backbone to be efficiently cyclized.

HaAEP1 can mediate post-translational processing of PawS1

The role of AEPs in albumin processing *in vivo* is well established. The most highly expressed AEP in sunflower seeds is HaAEP1, which can be recombinantly expressed, and thus we utilized this to reconstitute the first step of the reaction that produces SFTI-1 and matures the seed storage albumin. Previous studies into the processing mechanisms of this system have been using small peptide substrates and only focused on the biosynthesis of the cyclic peptide SFTI-1 (15). Here we show for the first time the processing events that take place in the bigger native precursor PawS1. HaAEP1 prefers Asn residues (15), and consistent with this, in our experiments it cleaves PawS1 at Asn18 and releases the SFTI-1 pro-peptide SFTI-GLDN. The linker peptide GLDN is located between the SFTI-1 domain and the albumin domain and is confirmed by our relaxation studies to be flexible. This flexibility is likely pivotal for enzyme access, as sterical hindrance would prevent the AEP enzyme from performing the cleavage reaction. Flexibility can also be seen in the loop region that contains the linker peptide LRMAVEN between helix I (SSU) and helix II (LSU), which will allow access to the AEP cleavage site Asn49. The sunflower preproalbumins for SESA2-1, SESA2-2 and SESA20-2, unlike PawS1 and PawS2, do not contain an Asn or Asp in this loop region and consequently cannot be processed but remain monomeric (7). The sunflower albumin SESA3 does contain an AEP-cleavage candidate Asn31 in this region, however structural characterization by NMR spectroscopy revealed that it is located towards the end of helix Ib (28) where the helical structure likely prevents sunflower HaAEP1 from cleaving SESA3 into a heterodimeric seed storage albumin.

Treatment with a sunflower seed extract *in situ* can fully mature PawS1

To further investigate the enzymatic cleavage events taking place with PawS1, a sunflower seed extract and NMR spectroscopy and mass spectrometry were used to monitor the processing *in situ.* The extract was able to rapidly process the linker between the SSU and LSU, and this cleavage was completed after 5-7 h, with the subsequent removal of the linker being completed over a few days. In contrast, only limited cleavage occurred after Asn18 to release SFTI-GLDN and produce the fully matured PawS1 albumin, and furthermore the production of mature cyclic SFTI-1 was too low to be confirmed by NMR or mass spectrometry. An inefficiency of *in situ* macrocyclization of SFTI-1 has been reported previously using seed extracts and the small enzyme substrate SFTI-GLDN (15). Only one in seven *in situ* reaction products from SFTI-GLDN produced cyclic SFTI-1, with the remaining six of seven reaction products being acyclic-SFTI. Bernath-Levin *et al*. (2015) found a degradative activity in sunflower seeds that led to the proposal of a breakdown pathway that would mask catalytic inefficiency *in vivo* by reducing the disulfide bonds and then degrading any acyclic-SFTI (15). However, the most likely limiting factor for the production of SFTI-1 in this study was the inefficient release of the SFTI-GLDN peptide segment, making the amount of smaller substrate available for further processing stoichiometrically unfavorable.

Why is SFTI-1 processing from PawS1 so inefficient in *in vitro* and *in situ* experiments?

Although we have presented evidence for the expected processing events of PawS1 using a recombinant AEP and a sunflower extract, the processes were surprisingly inefficient. This is in contrast to recent work on cyclotide processing using either extracted or recombinantly expressed AEPs, which has shown that substrates can be rapidly processed and cyclized to completion *in vitro* (13,14). There are some fundamental differences between these studies. First, AEP is a cysteine protease relying on reducing conditions for function, which is not always compatible with disulfide-rich substrates. Rather than the stronger reducing conditions previously reported we here use a milder redox system to prevent unfolding of the PawS1 protein during incubation. The stability under these conditions was found to be acceptable, but nonetheless evidence for reduction of one disulfide bond was observed coupled with subsequent off-target processing of Asn111. Secondly, all previous studies have used small artificial substrates focusing on the final step of processing – the trans-peptidation reaction. Here we use the full PawS1 processing studying the earlier steps of processing. For the *in vitro* studies we used recombinant HaAEP1. *HaAEP1* mRNA expression levels in seeds are 100 times higher than *HaAEP2* and *HaAEP3* and 500 times higher than *HaAEP4* and *HaAEP5* suggesting that, although expression levels are not directly related to protein content, HaAEP1 is a dominant AEP in sunflower seeds (15). Incubation ratios of sunflower HaAEP1 and PawS1 were calculated assuming an activity of 100% for expressed HaAEP1. However, due to the lack of an AEP enzyme inhibitor, no detailed characterization of the enzyme was performed to determine the active component of the expressed HaAEP1 fraction. Instead a small peptide was used as an enzyme control for substrate cleavage under the same redox conditions, and it was converted to near completion within 2 h, confirming AEP activity. In contrast less than half of PawS1 is cleaved at Asn49 after 92 h, thus clearly the slower processing of PawS1 is substrate-specific and suggest that despite the flexibility in the key regions, the overall size of PawS1 presents an obstacle that limits AEP access. The Asn49 cleavage site between the SSU and LSU is much more efficiently targeted than the Asn18 cleavage site between SFTI-1 and the albumin. This difference can be explained from the PawS1 structure suggesting that the conformation and protrusion from the core of the longer LRMAVEN loop makes it considerably more accessible than the shorter GLDN linker, which is masked by the flanking SFTI-1 and albumin domains. Cyclic knottin precursors generally contain multiple knottin domains separated by linker segments in dedicated precursor proteins (16,33), and these proteins may also require different conditions for efficient processing. Nonetheless, it is interesting to speculate in that the processing *in vivo* is aided by additional co-factors in the seed compartment that ensures efficient production of SFTI-1.

Why is SFTI-1 hidden in a seed storage albumin?

Macrocyclic peptides have now been described in plants, bacteria, fungi and even mammals (17). They vary greatly in size, character, physiological function and biosynthetic origin. Although a trend is emerging where the mature sequences are expressed as precursors with both N- and C-terminal extensions, thus requiring both cleavage and ligation events to produce mature products, the nature of the precursors,

their size, and the point of cyclization are diverse (33-35). The significance of these different expression systems for the folding and production is less clear as the precursors are poorly studied. Intriguingly there are no direct interactions between the SFTI-1 domain and the albumin in PawS1, rather they are separate individual structural entities. Consequently despite their symbiotic existence from a structural perspective it is unlikely that the albumin domain plays any role in the proper folding or processing of SFTI-1. The advantage of the genetic location of SFTI-1 may instead solely be related to the benefits of being able to hijack the albumin's seed expression profile, folding and processing machinery on the pathway via the ER and through the vacuolar system (10). Whether similar advantages have also driven the evolution of other classes of macrocyclic peptides to suit their physiological locations and functions, remain to be seen. Recently precursor variants of kalata B1 carrying both a short C-terminal tail and different lengths of an N-terminal sequence that is repeated before each cyclotide domain in the precursor protein were studied. The N-terminal sequence was found to be intrinsically unstructured, but appears to mediate self-association of precursors under NMR conditions, although the physiological significance of this is unclear (36).

In summary this study provides the 3D structure of a macrocyclic precursor protein and new mechanistic insights into the maturation of sunflower proalbumins into an albumin and SFTI-1. The structural characterization of PawS1 using triple resonance NMR experiments revealed a structure consisting of two well-defined entities connected by a flexible linker GLDN. A second flexible linker LRMAVEN separates the two subunits of the albumin. These linkers can be targeted by AEPs to produce the cleaved heterodimeric albumin PawS1 and the short peptide SFTI-GLDN, the starting point for further processing into cyclic SFTI-1. The site separating the albumin subunits is significantly more accessible for processing, being much more readily cleaved by both recombinant HaAEP1 and the *in situ* seed extract. The extract also contains secondary processing enzymes that remove the LRMAVEN sequence, leading to the fully matured albumin observed in seeds. The formation of cyclic SFTI-1 from PawS1 *in situ* could not be confirmed, likely due to both the inefficiency of liberation of SFTI-GLDN and the inefficiency of its biosynthesis from the short substrate, which has been reported previously (15). This is an intriguing observation and may suggest that additional auxiliary components or enzymes that are not present or active under the *in situ* conditions are involved in the biosynthesis of SFTI-1.

Experimental Procedures

Cloning of PawS1

A synthetic gene for PawS1 with *E. coli* codon optimization (GeneArt) was produced such that it encoded the 151-residue PawS1 minus the 21-residue signal peptide and a 14-residue spacer between the signal peptide and SFTI-1. The coordinates of the PawS1 protein were PawS1[Gly36-Ile151], with PawS1 Gly36 being the Gly1 of SFTI-1. The SFTI-1 Gly1 (i.e. PawS1[Gly36]) was made the P1′ residue for a TEV protease cleavage site. A restriction site was included that allowed the ORF to be cloned into pQE30 (QIAGEN) that adds an N-terminal sequence including a 6-His tag. The final sequence of the recombinant PawS1 protein encoded in pQE30 includes a N-terminal tag (MRGSHHHHHH), a flexible linker coded by BamHI (GS) a TEV cleavage site (ENLYFQ) followed by SFTI-1 (GRCTKSIPPICFPD) and its tail (GLDN) that connects SFTI-1 to the albumin domain (P_{54}...I_{151}). Thus:
MRGSHHHHHHGSENLYFQGRCTKSIPPICFPDGLDNP_{54}...I_{151}.

Recombinant expression of ^{13}C and ^{15}N labeled PawS1

The pQE30-PawS1 construct and the repressor plasmid pREP4 (QIAGEN) were co-transformed into NEB Shuffle® (New England Biolabs), an *E. coli* strain engineered to promote disulfide bond formation in its cytoplasm (37). SHuffle *E. coli* containing the pQE30-PawS1 construct was grown at 30°C at 200 rpm in LB media supplemented with ampicillin (100 μg/ml). When optical density reached OD_{600} ~ 0.8, cells were centrifuged at 5,000 x *g* for 15 min at 25°C. Cell pellets were resuspended in minimal medium (one quarter of the volume of LB used to grow cells). The minimal medium was made as previously described and contained ^{15}N labeled ammonium chloride and *D*-glucose-$^{13}C_6$ (Sigma-Aldrich) (38). The labeled culture was incubated for 1 h at 30°C at 200 rpm to promote cell recovery, and then cooled to 16°C for 1 h before adding iso-propyl β-D-1-thiogalactopyranoside at a final concentration of 0.4 mM. The culture was incubated for 18 h in the dark at 16°C at 200 rpm, before being harvested by centrifugation at 6,000 x *g* for 15 min at 4°C. Cell pellets were stored at -80°C until required.

Purification of PawS1

The cell pellet was resuspended in lysis buffer 50 mM Tris(hydroxymethyl)aminomethane, 300 mM sodium chloride, 10 mM imidazole, pH 8.0 followed by lysis by sonication. The soluble lysate was retained after centrifugation at 15,000 rpm for 30 min at 4°C. The His_6-TEV-PawS1 fusion protein was purified by passing the supernatant over a Ni Sepharose HisTrap HP 5 ml column (GE Healthcare Life Sciences) using an ÄKTA Start chromatography system (GE Healthcare Life Sciences), followed by a washing step with 50 mM Tris(hydroxymethyl)aminomethane, 300 mM sodium chloride, 60 mM imidazole, pH 8.0. The PawS1 fusion protein was eluted with 50 mM Tris(hydroxymethyl)aminomethane, 300 mM sodium chloride, 500 mM imidazole, pH 8.0. Flow rate was kept constant at 1 ml/min at all times. The imidazole-rich buffer was exchanged using a PD-10 column (GE Healthcare Life Sciences) and replaced with a buffer compatible with TEV (50 mM Tris(hydroxymethyl)aminomethane, 100 mM sodium chloride, 2 mM ethylenediaminetetraacetic acid, 0.6 mM/0.4 mM glutathione/glutathione-disulfide, pH 8.0). TEV protease was added in a ratio of 1:20 (w:w) and the cleavage reaction was allowed to proceed at room temperature for 16 h with gentle agitation.

The cleaved PawS1 protein was desalted using a PD-10 column (GE Healthcare Life Sciences) and separated further by RP-HPLC on a Shimadzu Prominence system using an analytical Grace Vydac C18 column (250 mm x 4.6 mm, 5 µm, 300 Å) at a 1% gradient in which buffer A was 0.05% trifluoroacetic acid and buffer B was 90% acetonitrile 0.05% trifluoroacetic acid. To determine the average mass of the labeled PawS1, the protein was analyzed by liquid chromatography mass spectrometry (LC-MS) on a SCIEX API 2000 LC-MS/MS electrospray mass spectrometer. A volume of 10 µL was injected at a flow rate of 0.1 mL/min with buffer A to buffer B ratio of 30:70. Buffer A consisted of 0.1% formic acid and buffer B contained 0.1% formic acid in 90% acetonitrile. Specifically LC-MS/MS instrument settings were as Declustering Potential (DP) 88 V; Focusing Potential (FP) 220 V; Entrance Potential (EP) 8 V; Q1 MS (Q1). The protein concentration was determined using a Direct Detect® Infrared Spectrometer (Merck Millipore). Purified PawS1 was lyophilized and stored at -20°C until required.

NMR spectroscopy

The structure of PawS1 was determined using heteronuclear NMR. Samples for NMR contained 0.3 mL of ^{15}N or $^{13}C/^{15}N$-labeled protein in 90% water and 10% D_2O (v/v) at pH 4.6 at a final concentration of 3.6 mg/mL and were added to susceptibility-matched 5 mm outer-diameter microtubes (Shigemi Inc.). All spectra were acquired at 25°C on an Avance 600 MHz spectrometer equipped with a cryoprobe (Bruker BioSpin). Resonance assignments were obtained using two-dimensional (2D) ^{1}H-^{15}N-heteronuclear single quantum coherence (HSQC), 2D ^{1}H-^{13}C-HSQC, 3D HNCACB, 3D CBCA(CO)NH, 3D HNCO, 3D HBHA(CO)NH, 3D H(CC)(CO)NH-TOCSY, and 3D (H)CC(CO)NH-TOCSY. 3D ^{13}C and ^{15}N-edited HSQC-NOESY spectra with a mixing time of 120 ms were recorded for structural information and a 2D ^{1}H-^{15}N-NOE relaxation experiment with saturation turned on or off during a 3 s relaxation delay was recorded to assess flexibility. Lyophilized protein was dissolved in D_2O and deuterium exchange was monitored by recording 2D ^{1}H-^{15}N-HSQC data sets. All data were recorded using linear sampling. 2D data were processed using Topspin 3.0 (Bruker Biospin) and 3D data were processed with the Rowland NMR Toolkit (http://rnmrtk.uchc.edu/rnmrtk/RNMRTK.html).

Spectral assignment and structure calculations

NMR spectra were analysed, peak picked and resonances assigned manually using CcpNmr Analysis 2.4.2 (39). Inter-proton distance restraints were derived from nuclear Overhauser effect spectroscopy (NOESY) cross peak heights from 3D ^{13}C-aliphatic, ^{13}C-aromatic, and ^{15}N-NOESY-HSQC spectra acquired using a mixing time of 120 ms. Protein backbone torsion angles φ, ψ and side chain torsion angle $\chi 1$ were predicted based on the chemical shift assignments of HN, Hα, Cα, Cβ and N resonances using the artificial neural network based program TALOS-N (20). For amide protons that were found to be slow exchanging hydrogen bond restraints were included if acceptors were unambiguously identified in preliminary structures. Backbone hydrogen bonds between amides and carbonyls included: Phe12-Arg2, Gln33-Leu29, His35-His31, Leu36-Cys32, Thr37-Gln33, Gln52-Asn49, Gln53-Pro50, Leu58-Gln54, Cys60-Leu56, Gln62-Leu58, Leu63-Cys59, Gln64-Cys60, Val66-Leu63, Cys70-Glu67, Ala74-Cys70, Ile75-Cys72, Gln77-Glu73, Val78-Ala74, Val79-Lys76, Glu80-Lys76, Gln81-Gln77, Ala82-Val78, Gln83-Val79, Gln85-Gln81, Gln87-Gln83, Gln95-Gly92, Val100-Val96, Lys102-Met99, Ala103-Val100, Gln104-Lys101, Met105-Lys102, Leu106-Ala103, Cys110-Leu106, Leu112-Pro107.

Initial structures were generated using Cyana 3.97 allowing iterative automatic assignment of the NOESY data (21). The final structures were calculated and refined in explicit water within CNS 1.21 using protocols from the RECOORD database (22,40). From the final round of structural calculations, 20 structures of a total of 50 were chosen based on lowest energy, and their covalent geometries analyzed using the structure validation web-service MolProbity (23). MolProbity calculates a score combining all-atom analysis such as steric interactions inside the model and geometric analysis such as Ramachandran and rotamer outliers. MOLMOL (41) and PyMOL were used to display and analyze the structures. Coordinates for the PawS1 structure were deposited in Biological Magnetic Resonance Bank and Protein Data Bank and given the accession codes: 30209 and 5U87.

Recombinant expression of sunflower HaAEP1

The $HaAEP1_{28-491}$ sequence optimized for expression in *E. coli* (GeneArt) fused to an N-terminal 6-His tag (MGRHHHHHHGS) in place of its signal peptide was cloned into pQE30 (QIAGEN). pQE30-SHuffle (NEB) *E. coli* containing HaAEP1 was grown at 30°C at 200 rpm in LB media supplemented with ampicillin (100 μg/ml). Upon reaching an OD_{600} of 0.8-1.0, the temperature was reduced to 16°C and incubated with shaking overnight. Cells were centrifuged at 5,000 x *g* for 10 min at 18°C. Cell pellets were frozen at -80°C prior to resuspending by sonication in lysis buffer (1 M sodium chloride, 50 mM Tris-Cl pH 8.0). The soluble lysate was isolated by centrifugation at 10,000 rpm for 15 min at 4°C. The His_6-TEV-HaAEP1 fusion protein was purified by incubating with Ni-NTA resin (BioRad) at 4°C overnight with mild agitation. The resin was washed with lysis buffer and eluted in elution buffer (50 mM Tris-Cl pH 8.0, 100 mM sodium chloride, 300 mM imidazole). HaAEP1 was activated by dialysis overnight at 4°C into 20 mM sodium acetate pH 5.5, 100 mM sodium chloride, 1 mM ethylenediaminetetraacetic acid, 5 mM dithiothreitol followed by a second dialysis into 20 mM sodium acetate pH 8.0, 100 mM sodium chloride, 1 mM ethylenediaminetetraacetic acid, 0.5 mM dithiothreitol. Aliquots were frozen in liquid nitrogen and stored at -80°C.

Stability studies of PawS1 in AEP activity buffer

To explore the stability of PawS1, 400 μg of ^{15}N labeled PawS1 were dissolved in 500 μL containing 90% AEP activity buffer (100 mM sodium acetate, 5 mM ethylenediaminetetraacetic acid, 0.6 mM/0.4 mM glutathione/glutathione-disulfide, pH 5.0) /10% D_2O (v/v) and two-dimensional ^{1}H-^{15}N-HSQC NMR spectrum was recorded at various time points over a two-week period. The ^{1}H-^{15}N-HSQC data sets of the PawS1 dissolved in 90% AEP activity buffer /10% D_2O (v/v) were then compared with the ^{1}H-^{15}N-HSQC data sets of PawS1 dissolved in 90% water and 10% D_2O (v/v).

***In vitro* digests of PawS1 with sunflower HaAEP1**

Recombinant HaAEP1 at a final concentration of 4 μg/mL was incubated with either 74 μM SFTI(D14N)-GL or 9.5 μM PawS1 for activity measurements in AEP activity buffer at 37°C. SFTI(D14N)-GL at a final concentration of 37 μM in AEP activity buffer without HaAEP1 and PawS1 at a final concentration of 4.8 μM in AEP activity buffer without HaAEP1 were used as negative controls. At 0 h, 2 h, 5 h, 21 h and 92 h, 5 μL aliquots were removed and desalted using ZipTip pipette filters (Merck Millipore) before being combined in a ration 1:1 with α-cyano-4-hydroxycinnamic acid (CHCA) matrix (Bruker Daltonics) onto a 384 MTP polished steel target plate, ready for MALDI-TOF MS analysis. The concentration of the CHCA matrix was 0.7 mg/mL.

***In situ* digests of PawS1 with sunflower seed extract**

Ten kernels of sunflower seeds (purchased from Coles, Australia) were frozen in liquid nitrogen and crushed using a mortar and pestle. The crushed seed meal was re-suspended in AEP activity buffer at a ratio of 0.1 mL AEP activity buffer per de-hulled sunflower seed kernel. The re-suspended meal was vortexed for 10 min and centrifuged at 10,000 x *g* for 2 min. The supernatant was transferred into a clean tube and mixed with an equal amount of *n*-hexane to remove lipids. The mixture was spun at 10,000 x *g* for 10 sec and the *n*-hexane was removed and the extract was transferred to an Amicon Ultra-0.5 mL 10 kDa centrifugal filter (Merck Millipore) and topped up with AEP activity buffer. The sunflower seed extract was centrifuged at 14,000 x *g* for 10 min. Sunflower seed extract was added in a ratio of 2 parts : 5 parts (volume extract : mass protein) to unlabeled and ^{15}N labeled PawS1. The mixture was incubated at 37°C for 0 h, 2 h and 5 h with 5 μL aliquots taken at each time point and desalted using ZipTip pipette filters (Merck Millipore) before being combined with CHCA matrix (Bruker Daltonics) onto a 384 MTP polished steel target plate, ready for MALDI-TOF MS analysis.

For ^{15}N NMR studies 70 kernels of sunflower seeds were crushed as described above. After supernatant was transferred into a clean tube and mixed with an equal amount of *n*-hexane to remove lipids, no Amicon Ultra-0.5 mL centrifugal filter was used in ^{15}N NMR study and sunflower seed extract was filtered through a 0.45 μm filter instead. 400 μg of ^{15}N labeled PawS1 were dissolved in 450 μL *in situ* sunflower seed extract containing AEP activity buffer and 50 μL D_2O (v/v) and ^{1}H-^{15}N- HSQC NMR measurements were undertaken at 37°C at various time points over 10 days.

MALDI-TOF MS

To monitor the processing of PawS1, 1.2 μL of desalted peptides/proteins were mixed with 1.2 μL of matrix and 2.4 μL were spotted onto a 384 MTP polished steel target plate. The matrix was prepared according to the Bruker Daltonics protocol using the CHCA dried droplet method for polished steel target plates. The Bruker peptide calibration standard (Bruker Daltonics) was used for calibrating the mass range 800-4,000 Da (low mass) and the Bruker protein calibration standard I (Bruker Daltonics) was used for calibrating the mass range 5,000-14,000 Da (high mass). Samples were analyzed on a Bruker UltrafleXtreme™ MALDI-TOF/TOF mass spectrometer (Bruker Daltonics) with laser intensity at 20% for low mass and 30% for high mass in the *in vitro* assay and 50% for low mass and 40% for high mass in the *in situ* assay. 5,000 shots were summed per MS analysis.

Acknowledgments

This work and BF were supported by Australian Research Council grant DP120103369 to JSM and KJR. MM, JSM and KJR were supported by Australian Research Council Future Fellowships FT110100925, FT120100013 and FT130100890, respectively. The authors acknowledge the access to the 600 MHz spectrometer at the Biological NMR facility at the Institute for Molecular Bioscience and CSIRO Agriculture and Food for access to the Bruker UltrafleXtreme MALDI-TOF/TOF mass spectrometer.

Conflict of Interest

The authors declare no conflicts of interest.

Author contributions

KJR and JSM conceived the study. BF performed all experiments with the help of MM (NMR data recording, processing and analysis), MLC (mass spectrometry recording and analysis), KJR (structure calculations and analysis). The recombinant expression of HaAEP1 was done by AMJ. BF, JSM and KJR wrote the manuscript with contributions from all authors.

References

1. Shimada, T., Yamada, K., Kataoka, M., Nakaune, S., Koumoto, Y., Kuroyanagi, M., Tabata, S., Kato, T., Shinozaki, K., Seki, M., Kobayashi, M., Kondo, M., Nishimura, M., and Hara-Nishimura, I. (2003) Vacuolar Processing Enzymes Are Essential for Proper Processing of Seed Storage Proteins in *Arabidopsis thaliana*. *J. Biol. Chem.* **278**, 32292-32299
2. 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
3. Shewry, P. R., and Casey, R. (1999) *Seed proteins*, Kluwer Academic, Dordrecht Boston
4. Otegui, M. S., Herder, R., Schulze, J., Jung, R., and Staehelin, L. A. (2006) The Proteolytic Processing of Seed Storage Proteins in *Arabidopsis* Embryo Cells Starts in the Multivesicular Bodies. *Plant Cell* **18**, 2567-2581
5. Mylne, J. S., Hara-Nishimura, I., and Rosengren, K. J. (2014) Seed storage albumins: biosynthesis, trafficking and structures. *Funct. Plant Biol.* **41**, 671-677

6. Ericson, M. L., Rödin, J., Lenman, M., Glimelius, K., Josefsson, L. G., and Rask, L. (1986) Structure of the Rapeseed 1.7 S Storage Protein, Napin, and Its Precursor. *J. Biol. Chem.* **261**, 14576-14581
7. Franke, B., Colgrave, M. L., Mylne, J. S., and Rosengren, K. J. (2016) Mature forms of the major seed storage albumins in sunflower: A mass spectrometric approach. *J. Proteomics* **147**, 177-186
8. Kortt, A. A., Caldwell, J. B., Lilley, G. G., and Higgins, T. J. (1991) Amino acid and cDNA sequences of a methionine-rich 2S protein from sunflower seed (*Helianthus annuus* L.). *Eur. J. Biochem.* **195**, 329-334
9. Jayasena, A. S., Franke, B., Rosengren, K. J., and Mylne, J. S. (2016) A tripartite approach identifies the major sunflower albumins. *Theor. Appl. Genet.* **129**, 613-617
10. 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
11. Craik, D. J., Fairlie, D. P., Liras, S., and Price, D. (2013) The Future of Peptide-based Drugs. *Chem. Biol. Drug Des.* **81**, 136-147
12. Elliott, A. G., Delay, C., Liu, H., Phua, Z., Rosengren, K. J., Benfield, A. H., Panero, J. L., Colgrave, M. L., Jayasena, A. S., Dunse, K. M., Anderson, M. A., Schilling, E. E., Ortiz-Barrientos, D., Craik, D. J., and Mylne, J. S. (2014) Evolutionary Origins of a Bioactive Peptide Buried within Preproalbumin. *Plant Cell* **26**, 981-995
13. Nguyen, G. K., Wang, S., Qiu, Y., Hemu, X., Lian, Y., and Tam, J. (2014) Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. *Nat. Chem. Biol.* **10**, 732-738
14. Harris, K. S., Durek, T., Kaas, Q., Poth, A. G., Gilding, E. K., Conlan, B. F., Saska, I., Daly, N. L., van der Weerden, N. L., Craik, D. J., and Anderson, M. A. (2015) Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. *Nat. Commun.* **6**, 10199
15. Bernath-Levin, K., Nelson, C., Elliott, A. G., Jayasena, A. S., Millar, A. H., Craik, D. J., and Mylne, J. S. (2015) Peptide Macrocyclization by a Bifunctional Endoprotease. *Chem. Biol.* **22**, 571-582
16. Mylne, J. S., Chan, L. Y., Chanson, A. H., Daly, N. L., Schaefer, H., Bailey, T. L., Nguyencong, P., Cascales, L., and Craik, D. J. (2012) Cyclic peptides arising by evolutionary parallelism via asparaginyl-endopeptidase-mediated biosynthesis. *Plant Cell* **24**, 2765-2778
17. Craik, D. J. (2006) Chemistry. Seamless proteins tie up their loose ends. *Science (New York, N.Y.)* **311**, 1563-1564
18. Korsinczky, M. L., Schirra, H. J., Rosengren, K. J., West, J., Condie, B. A., Otvos, L., Anderson, M. A., and Craik, D. J. (2001) Solution structures by 1H NMR of the novel cyclic trypsin inhibitor SFTI-1 from sunflower seeds and an acyclic permutant. *J. Mol. Biol.* **311**, 579-579
19. Wüthrich, K. (1986) *NMR of proteins and nucleic acids*, John Wiley & Sons, New York
20. Shen, Y., and Bax, A. (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J. Biomol. NMR* **56**, 227-241
21. Güntert, P. (2004) Automated NMR structure calculation with CYANA. *Methods in Molecular Biology (Clifton, N.J.)* **278**, 353-378
22. Brunger, A. T. (2007) Version 1.2 of the Crystallography and NMR system. *Nat. Protoc.* **2**, 2728-2733
23. Chen, V. B., Arendall, r. W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12-21

24. Klint, J. K., Senff, S., Saez, N. J., Seshadri, R., Lau, H. Y., Bende, N. S., Undheim, E. A., Rash, L. D., Mobli, M., and King, G. F. (2013) Production of recombinant disulfide-rich venom peptides for structural and functional analysis via expression in the periplasm of *E. coli*. *PLoS ONE* **8**, e63865
25. Hiraiwa, N., Kondo, M., Nishimura, M., and Hara-Nishimura, I. (1997) An aspartic endopeptidase is involved in the breakdown of propeptides of storage proteins in protein-storage vacuoles of plants. *Eur. J. Biochem.* **246**, 133-141
26. Rico, M., Bruix, M., González, C., Monsalve, R. I., and Rodríguez, R. (1996) 1H NMR assignment and global fold of napin BnIb, a representative 2S albumin seed protein. *Biochemistry* **35**, 15672-15682
27. Pantoja-Uceda, D., Bruix, M., Giménez-Gallego, G., Rico, M., and Santoro, J. (2003) Solution structure of RicC3, a 2S albumin storage protein from *Ricinus communis*. *Biochemistry* **42**, 13839-13847
28. Pantoja-Uceda, D., Shewry, P. R., Bruix, M., Tatham, A. S., Santoro, J., and Rico, M. (2004) Solution Structure of a Methionine-Rich 2S Albumin from Sunflower Seeds: Relationship to Its Allergenic and Emulsifying Properties. *Biochemistry* **43**, 6976-6986
29. Pantoja-Uceda, D., Palomares, O., Bruix, M., Villalba, M., Rodríguez, R., Rico, M., and Santoro, J. (2004) Solution structure and stability against digestion of rproBnIb, a recombinant 2S albumin from rapeseed: relationship to its allergenic properties. *Biochemistry* **43**, 16036-16045
30. Lehmann, K., Schweimer, K., Reese, G., Randow, S., Suhr, M., Becker, W.-M., Vieths, S., and Rösch, P. (2006) Structure and stability of 2S albumin-type peanut allergens: implications for the severity of peanut allergic reactions. *Biochem. J.* **395**, 463-472
31. Li, D.-F., Jiang, P., Zhu, D.-Y., Hu, Y., Max, M., and Wang, D.-C. (2008) Crystal structure of Mabinlin II: A novel structural type of sweet proteins and the main structural basis for its sweetness. *J. Struct. Biol.* **162**, 50-62
32. Rundqvist, L., Tengel, T., Zdunek, J., Björn, E., Schleucher, J., Alcocer, M. J. C., and Larsson, G. (2012) Solution Structure, Copper Binding and Backbone Dynamics of Recombinant Ber e 1-The Major Allergen from Brazil Nut. *PLoS ONE* **7**, e46435
33. Jennings, C., West, J., Waine, C., Craik, D., and Anderson, M. (2001) Biosynthesis and Insecticidal Properties of Plant Cyclotides: The Cyclic Knotted Proteins from *Oldenlandia affinis*. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10614-10619
34. Himeno, K., Rosengren, K. J., Inoue, T., Perez, R. H., Colgrave, M. L., Lee, H. S., Chan, L. Y., Henriques, S. T., Fujita, K., Ishibashi, N., Zendo, T., Wilaipun, P., Nakayama, J., Leelawatcharamas, V., Jikuya, H., Craik, D. J., and Sonomoto, K. (2015) Identification, Characterization, and Three-Dimensional Structure of the Novel Circular Bacteriocin, Enterocin NKR-5-3B, from *Enterococcus faecium*. *Biochemistry* **54**, 4863-4876
35. Tang, Y.-Q., Yuan, J., Ösapay, G., Ösapay, K., Tran, D., Miller, C. J., Ouellette, A. J., and Selsted, M. E. (1999) A Cyclic Antimicrobial Peptide Produced in Primate Leukocytes by the Ligation of Two Truncated α-Defensins. *Science* **286**, 498-502
36. Daly, N. L., Gunasekera, S., Clark, R. J., Lin, F., Wade, J. D., Anderson, M. A., and Craik, D. J. (2016) The N-terminal pro-domain of the kalata B1 cyclotide precursor is intrinsically unstructured. *Biopolymers* **106**, 825-833
37. Lobstein, J., Emrich, C. A., Jeans, C., Faulkner, M., Riggs, P., and Berkmen, M. (2012) SHuffle, a novel *Escherichia coli* protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. *Microb. Cell Fact.* **11**, 56
38. Marley, J., Lu, M., and Bracken, C. (2001) A method for efficient isotopic labeling of recombinant proteins. *J. Biomol. NMR* **20**, 71-75
39. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L.,

Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins: Struct. Funct. Bioinf.* **59**, 687-696
40. Nederveen, A. J., Doreleijers, J. F., Vranken, W., Miller, Z., Spronk, C. A. E. M., Nabuurs, S. B., Güntert, P., Livny, M., Markley, J. L., Nilges, M., Ulrich, E. L., Kaptein, R., and Bonvin, A. M. J. J. (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **59**, 662-672
41. Koradi, R., Billeter, M., and Wüthrich, K. (1996) MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51-55

Table 1: Structural statistics for the NMR structure of PawS1.

Distance restraints	
Total NOE	2659
Intra-residue	589
Inter-residue	2070
Sequential ($\|i-j\| = 1$)	930
Medium range ($1<\|i-j\| \leq 4$)	685
Long range ($\|i-j\| \geq 5$)	455
Hydrogen bond restraints	68
Dihedral-angle restraints	208
φ	94
ψ	87
χ_1	27
Atomic RMSD [Å] [a]	
Mean global backbone overlaid over SFTI-1	0.23 ± 0.06
Mean global heavy overlaid over SFTI-1	0.71 ± 0.17
Mean global backbone overlaid over albumin	0.53 ± 0.09
Mean global heavy overlaid over albumin	1.10 ± 0.11
Molprobity statistics [b]	
Ramachandran favoured [%]	92.26 ± 1.98
Ramachandran outliers [%]	0.53 ± 0.59
Poor rotamers [%]	2.92 ± 1.49
Favored rotamers [%]	85.52 ± 2.24
Clash score [c]	10.63 ± 1.65 (69%)
MolProbity score [d]	2.32 ± 0.15 (57%)

[a] RMSD values were calculated over residues 3-11 for SFTI-1 and over residues 22-38 and 54-114 for albumin.

[b] http://molprobity.biochem.duke.edu.

[c] Clashes is the number of steric overlaps >0.4 Å per 1000 atoms.

[d] 100 % is the best structure ranked by MolProbity.

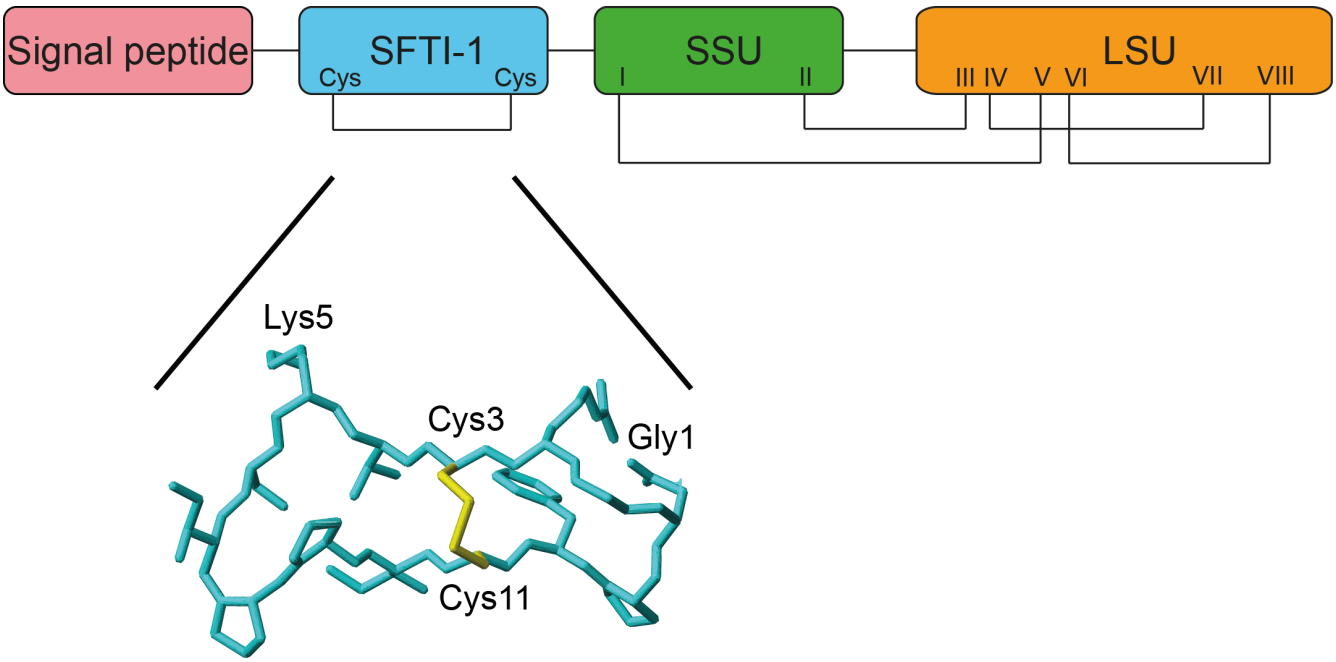

Figure 1: Cartoon schematic of the domains of the preproalbumin PawS1 from *Helianthus annuus*. The N-terminal signal peptide domain is highlighted in pink, the SFTI-1 peptide domain in cyan, the small subunit (SSU) of the albumin domain in green and the large subunit (LSU) in orange. During processing the SFTI-1 sequence is liberated and cyclized into a 14-residue peptide with one disulfide bond and a cyclic backbone. The cysteine connectivity highlighted by connecting lines (I-V, II-III, IV-VII and VI-VIII within the albumin sequence) is conserved amongst plant 2S albumins.

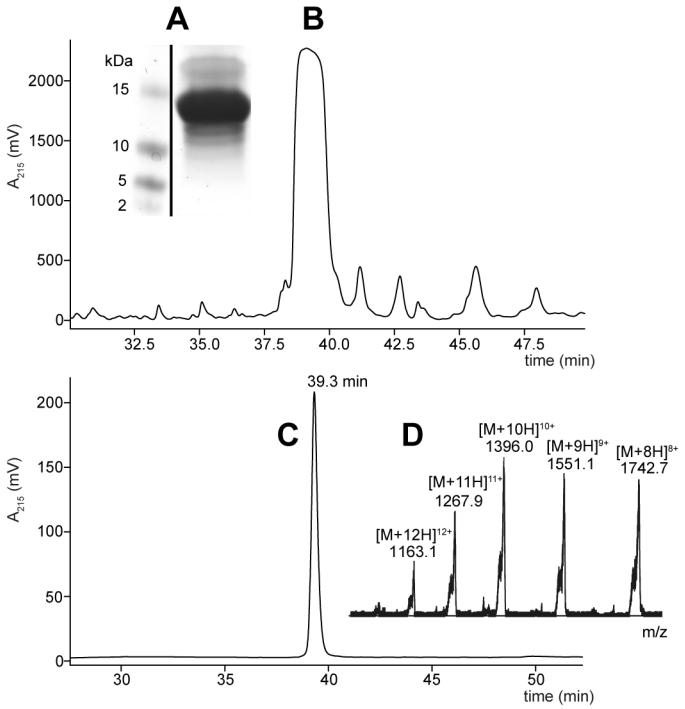

Figure 2: Expression and purification of PawS1. (A) 1D gel showing His_6-TEV- PawS1 fusion protein before Ni Sepharose HisTrap purification. Lanes are from the same gel but moved next to each other for clarity. **(B)** RP-HPLC chromatogram showing purification of PawS1 after removal of N-terminal His_6-fusion tag by TEV protease. **(C)** Analytical RP-HPLC trace showing ~95% pure protein judged by the single uniform peak. **(D)** LC-MS of ^{13}C and ^{15}N labeled PawS1 purified by RP-HPLC.

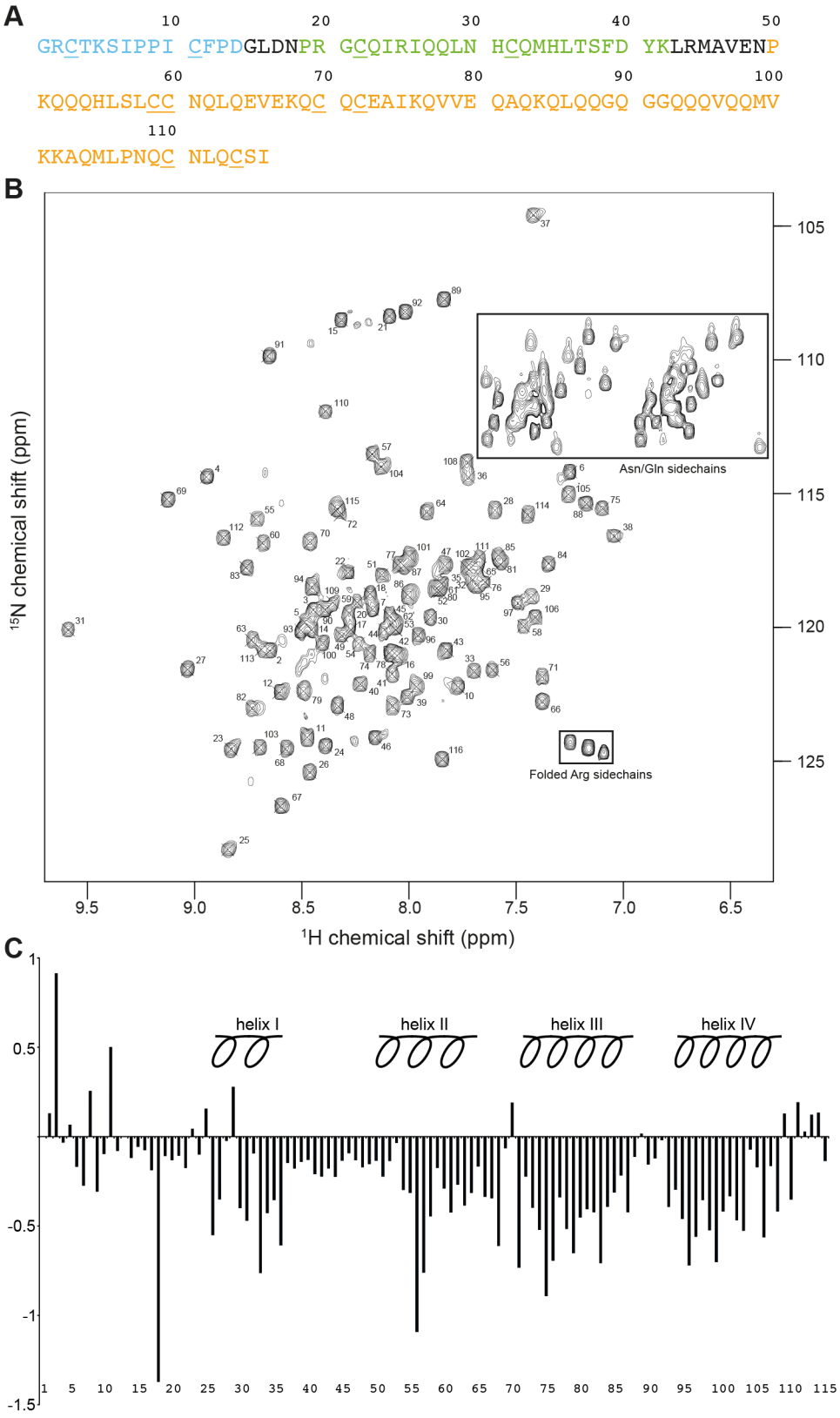

Figure 3: Sequence, 2D ^{1}H-^{15}N HSQC spectrum acquired at 25°C and 600 MHz and secondary H*α* chemical shifts for PawS1. (A) Sequence and numbering of the studied PawS1 protein showing the N-terminal SFTI-1 domain highlighted in cyan, the SSU in green and the LSU in orange. Linker peptides shown in black connect SFTI-1 and the SSU as well as the SSU and the LSU of the heterodimeric albumin. **(B)** Peaks are labeled with the residue numbers of their corresponding

backbone amides. Resonances originating from Asn/Gln and folded Arg side chains are highlighted by boxes. **(C)** Secondary shifts are observed chemical shifts subtracted by the chemical shift expected for a random coil conformation, and are indicative of secondary structure. The secondary Hα chemical shifts of the N-terminal SFTI-1 domain show the same trend as cyclic SFTI-1. Positive secondary shifts are indicative of β-sheet. Both the small and large subunits of the albumin domain show extensive stretches of negative secondary shifts, indicating a helical structure.

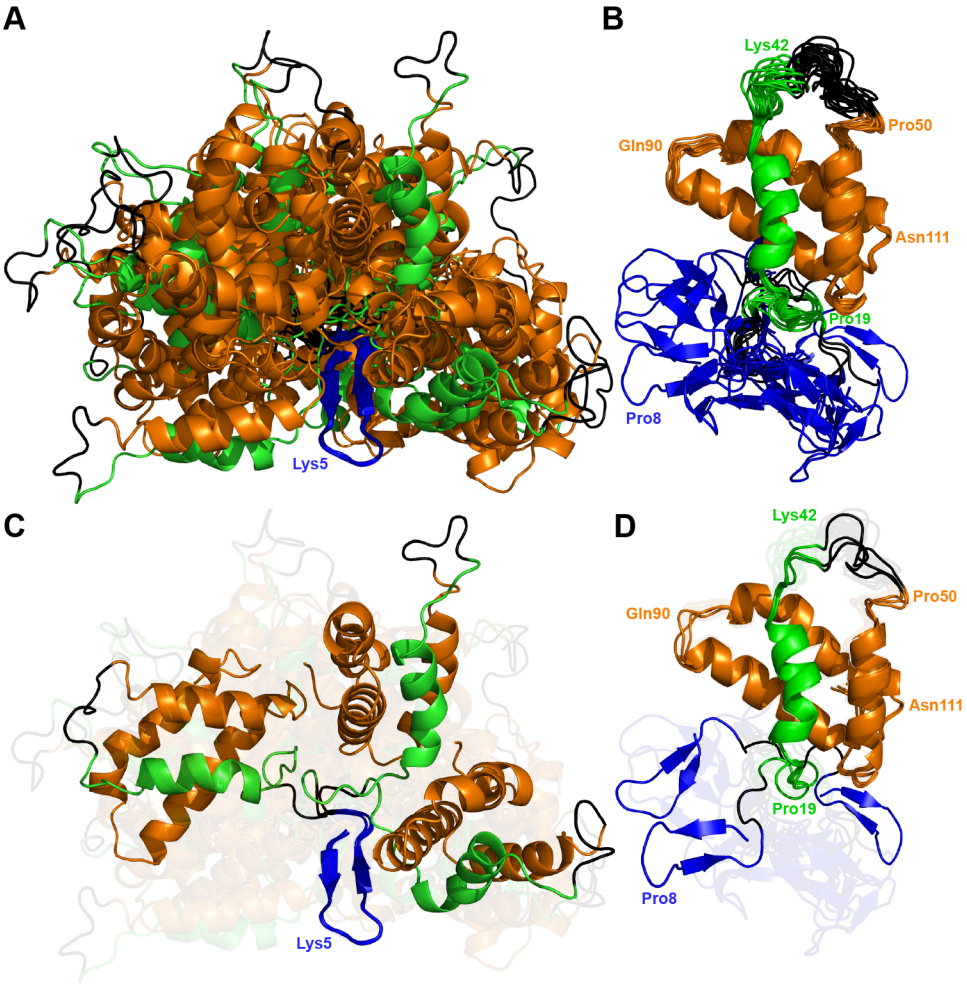

Figure 4: Three-dimensional solution NMR structure of PawS1. (A) Ensemble of the 20 structures representing PawS1 superimposed over the SFTI-1 domain. **(B)** Ensemble of the 20 structures representing PawS1 superimposed over the albumin domain. The SFTI-1 domain is shown in blue, the small subunit in green, the large subunit in orange and the linker peptides GLDN and LRMAVEN are shown in black. **(C, D)** Three structures from the ensemble are highlighted to clarify the range of conformations observed, with the remainder of the structures shown in transparent. Movie files highlighting the conformational rearrangements within the ensemble are provided as supplementary information.

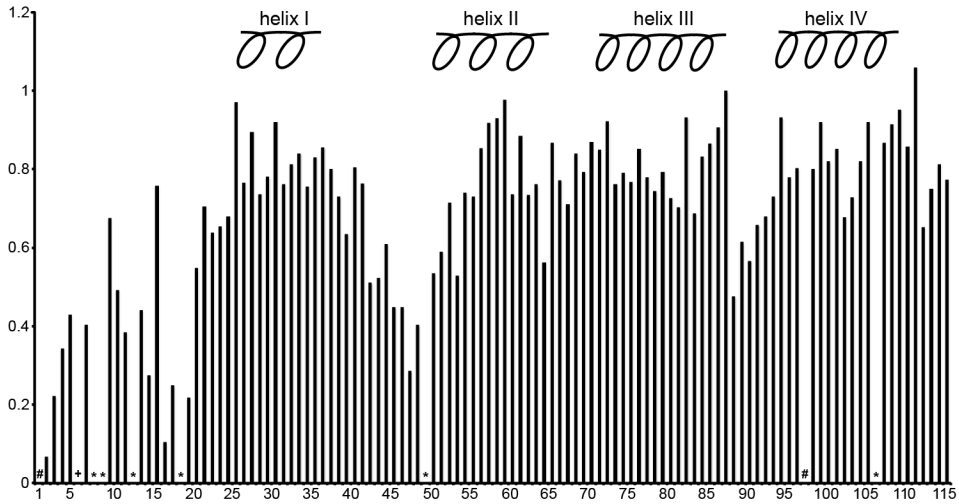

Figure 5: Heteronuclear ^{1}H-^{15}N steady state NOE data for ^{15}N labeled PawS1. *Proline residues Pro8, Pro9, Pro13, Pro19, Pro50 and Pro107 lack amide protons. #The amide proton for Gln98 was not identified. +The peak for Ser6 was not detected in the NOE on experiment and thus a ratio could not be calculated.

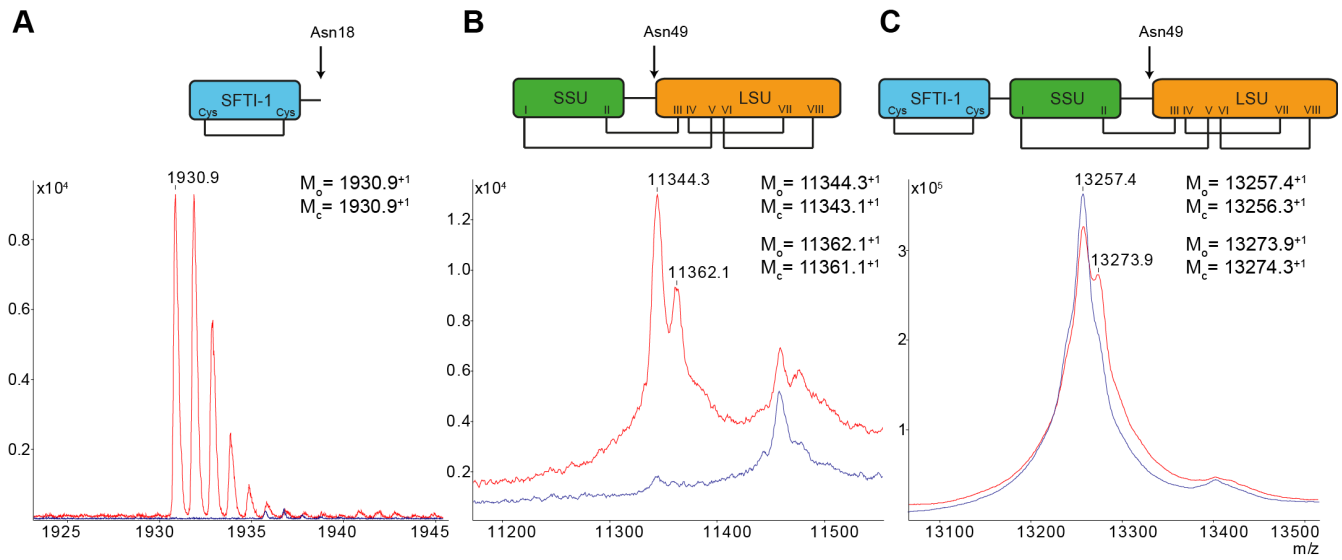

Figure 6: *In vitro* digest of PawS1 with sunflower HaAEP1. MALDI MS spectra showing cleavage products after incubation of PawS1 with HaAEP1 for 92 h at 37°C. **(A)** Sunflower HaAEP1 cleaves the PawS1 after Asn18 and produces the linear SFTI-GLDN peptide. **(B)** HaAEP1 also cleaves between the two subunits of the albumin PawS1 after Asn49, which results in a mass shift of +18 Da. **(C)** This cleavage can also be seen in full length PawS1, resulting in a mass shift of +18 Da. Observed (M_o) and calculated (M_c) monoisotopic masses (Da; $[M+H]^+$) are listed for SFTI-GLDN and observed and calculated average masses (Da; $[M+H]^+$) are listed for the albumin PawS1 and PawS1 respectively. The red line represents MALDI MS data of PawS1 incubated with HaAEP1 in AEP activity buffer. The blue line represents the control, MALDI MS data of PawS1 incubated without HaAEP1 in AEP activity buffer.

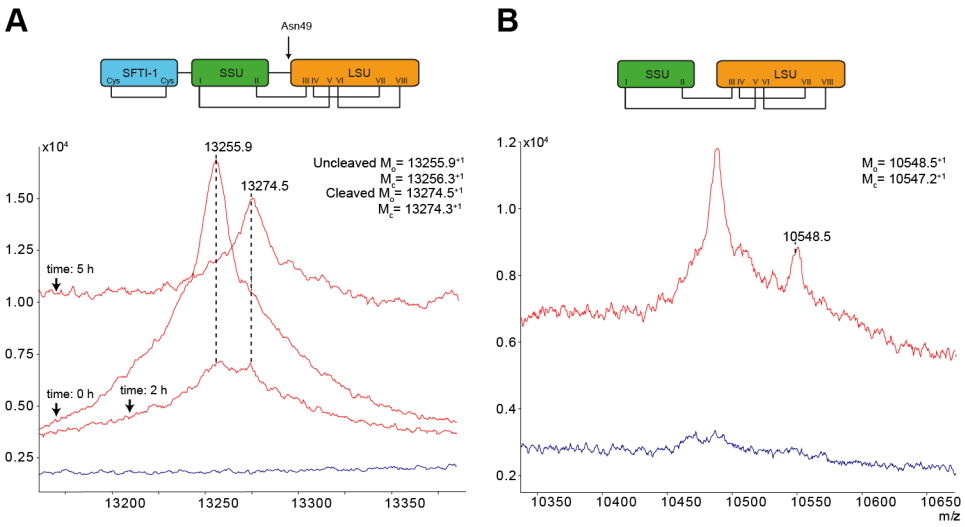

Figure 7: *In situ* digest of PawS1 with sunflower seed extract. (A) MALDI MS spectra showing full length PawS1 during incubation with sunflower seed extract at 37°C. After 2 h cleavage has occurred after the linker peptide LRMAVEN, between the two subunits of PawS1. After 5 h, most of the PawS1 linker region was cleaved**. (B)** After 5 h fully matured PawS1 heterodimeric albumin with SFTI-GLDN and the LRMAVEN linker removed was also observed. The red line represents MALDI MS data of PawS1 incubated with sunflower seed extract in AEP activity buffer. The blue line represents the control, MALDI MS data of the sunflower seed extract incubated in AEP activity buffer for 5 h at 37°C without PawS1 protein.

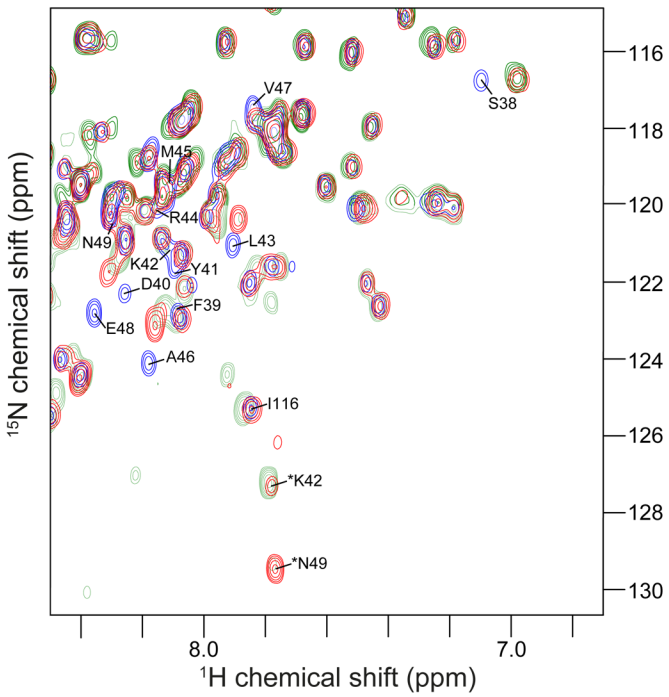

Figure 8: 2D ^{1}H-^{15}N HSQC spectra of PawS1 incubated in *in situ* sunflower seed extract acquired at 37°C and 600 MHz after 0 h, 7 h and 10 days. Blue peaks represent ^{1}H-^{15}N HSQC data of PawS1 incubated with *in situ* sunflower seed extract at 0 h. Red peaks represent ^{1}H-^{15}N HSQC data of PawS1 incubated with *in situ* sunflower seed extract at 7 h. Green peaks represent ^{1}H-^{15}N HSQC data of PawS1 incubated with *in situ* sunflower seed extract at 10 days. * indicate C-terminal residues of the intermediate and fully processed forms..

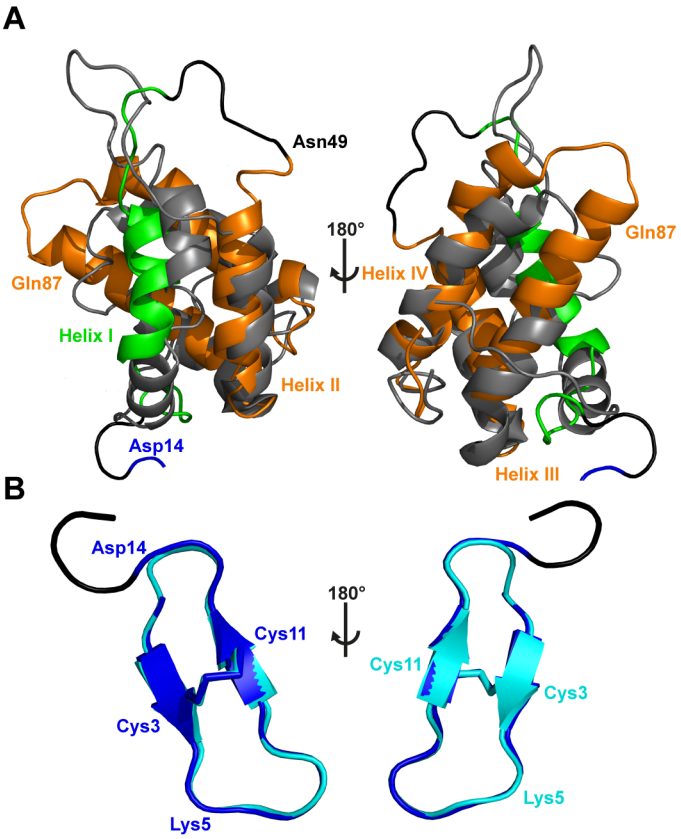

Figure 9: Comparison of the NMR structure PawS1 with the sunflower albumin SESA3 and SFTI-1. (A) The albumin domain of PawS1, showing SSU in green and LSU in orange, is superimposed on the sunflower albumin SESA3 shown in grey (PDP code of SESA3: 1S6D) (28). Helices are numbered with Roman numbers and selected residues are labeled for clarity. **(B)** The SFTI-1 domain of PawS1 shown in blue is superimposed on the solution NMR structure of SFTI-1 shown in cyan (PDP code of SFTI-1: 1JBL) (18).

Two Proteins for the Price of One: Structural Studies of the Dual Destiny Preproalbumin with Sunflower Trypsin Inhibitor-1

Bastian Franke, Amy M. James, Mehdi Mobli, Michelle L. Colgrave, Joshua S. Mylne and K. Johan Rosengren

J. Biol. Chem. published online May 23, 2017
