THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 280, NO. 37, pp. 32245–32253, September 16, 2005

Discovery, Structural Determination, and Putative Processing of the Precursor Protein That Produces the Cyclic Trypsin Inhibitor Sunflower Trypsin Inhibitor 1*

Received for publication, June 3, 2005, and in revised form, July 19, 2005 Published, JBC Papers in Press, July 21, 2005, DOI 10.1074/jbc.M506060200

Jason P. Mulvenna, Fiona M. Foley, and David J. Craik[1]

From the Institute for Molecular Bioscience, Australian Research Council Center for Functional and Applied Genomics, University of Queensland, Brisbane, Queensland 4072, Australia

Backbone-cyclized proteins are becoming increasingly well known, although the mechanism by which they are processed from linear precursors is poorly understood. In this report the sequence and structure of the linear precursor of a cyclic trypsin inhibitor, sunflower trypsin inhibitor 1 (SFTI-1) from sunflower seeds, is described. The structure indicates that the major elements of the reactive site loop of SFTI-1 are present before processing. This may have importance for a protease-mediated cyclizing reaction as the rigidity of SFTI-1 may drive the equilibrium of the reaction catalyzed by proteolytic enzymes toward the formation of a peptide bond rather than the normal cleavage reaction. The occurrence of residues in the SFTI-1 precursor susceptible to cleavage by asparaginyl proteases strengthens theories that involve this enzyme in the processing of SFTI-1 and further implicates it in the processing of another family of plant cyclic proteins, the cyclotides. The precursor reported here also indicates that despite strong active site sequence homology, SFTI-1 has no other similarities with the Bowman-Birk trypsin inhibitors, presenting interesting evolutionary questions.

Naturally occurring circular proteins are becoming increasingly well known, with examples in bacteria, plants, and animals discovered over recent years (1). These topologically interesting proteins have a continuous cycle of peptide bonds in their backbone and, accordingly, are devoid of N or C termini. Such proteins were unknown a decade ago, and they differ from previously known cyclic peptides, such as the immunosuppressant cyclosporin and other cyclic peptides found in micro-organisms, in that they are conventional gene products rather than the output of nonribosomal synthetic processes. This new class of protein has excited interest as cyclization of a protein backbone has the potential to provide stabilization relative to conventional proteins, both in a thermodynamic sense (2) and biologically (3). Because circular proteins have no termini, they have no target sites for amino- or carboxypeptidases. Furthermore, because the termini of conventional proteins are often flexible, and the degree of flexibility can be reduced by cyclization, entropic factors can lead to improved receptor binding affinities of circular proteins over corresponding acyclic proteins. Thus, circular proteins potentially have a range of advantages over conventional proteins and may find interesting applications in protein engineering, agriculture, and drug design (4).

SFTI-1[2] is a 14-amino-acid, backbone-cyclized protein found in the seeds of *Helianthus annus* (sunflower), which possesses potent trypsin inhibitory activity as well as weaker inhibitory activity against proteases such as thrombin (5, 6). Based on sequence homology, SFTI-1 appears to be related to the Cys-rich trypsin inhibitors called the Bowman-Birk inhibitors (BBIs) that are widely distributed within plants of the Fabaceae (leguminous) and Poaceae (graminaceous) families (7, 8). Two classes of BBIs have been identified as follows: the first class, found in the Poaceae, possessing a single active site, and the second class, found in both plant families, possessing two active sites that are capable of inhibiting both chymotrypsin and trypsin simultaneously in a 1:1:1 stoichiometry (9, 10). SFTI-1 appears to be a natural peptide mimetic of the BBIs as it shares strong sequence and structural homology with the trypsin inhibitory loop of the BBI family (Fig. 1) (6, 11). Remarkably, despite strong sequence and structural conservation, SFTI-1 has only been found in the seeds of the sunflower, which, as a member of the Compositae, is phylogenetically removed from other plant families containing BBIs. Furthermore, no other proteins sharing similarity with the BBIs have been identified in the Compositae (12), suggesting that SFTI-1 is restricted to a single, phylogenetically isolated species.

Two other classes of cyclic proteins have been discovered in the plant kingdom, namely the cyclotides (13) and the cyclic knottins (3, 14), but only the cyclotides have been characterized at a genetic level to reveal their linear precursor (15, 16). The conservation of key residues, particularly Asn and Asp residues, in the cyclotide precursor has led to the suggestion that proteases such as asparaginyl endoproteases are involved in the cyclization reaction (13, 15). The involvement of proteases in cyclization has been further suggested by the report of *in vitro* cyclization of linearized SFTI-1 with trypsin (17) and the characterization of the biosynthetic pathway of the cyclic sex pilin subunit TrbC. In the pilin system a serine protease-like enzyme, TraF, mediates both the cleavage of the TrbC precursor and the subsequent ligation of the N and C termini via aminolysis of an acyl intermediate (18–20). Both of these reactions rely on the presence of certain residues, and if the *in vivo* cyclization of plant proteins proceeds via a protease-catalyzed path, then the determination of precursor sequences for all classes of plant-cyclized proteins is of paramount importance for understanding the cyclization mechanism.

To date, nothing is known about the biosynthetic pathway of SFTI-1,

* This work was supported by a grant (to D. J. C.) from the Australian Research Council. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 2AB9) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

[1] Australian Research Council Professional Fellow. To whom correspondence should be addressed: Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia. Tel.: 61-7-3346-2019; Fax: 61-7-3346-2029; E-mail: d.craik@imb.uq.edu.au.

[2] The abbreviations used are: SFTI-1, sunflower trypsin inhibitor 1; BBI, Bowman-Birk inhibitor; RE, regular expression; EST, expressed sequence tag; RP-HPLC, reverse phase-high pressure liquid chromatography; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; TOCSY, total correlation spectroscopy; PDB, protein data bank; r.m.s.d., root mean square deviation.

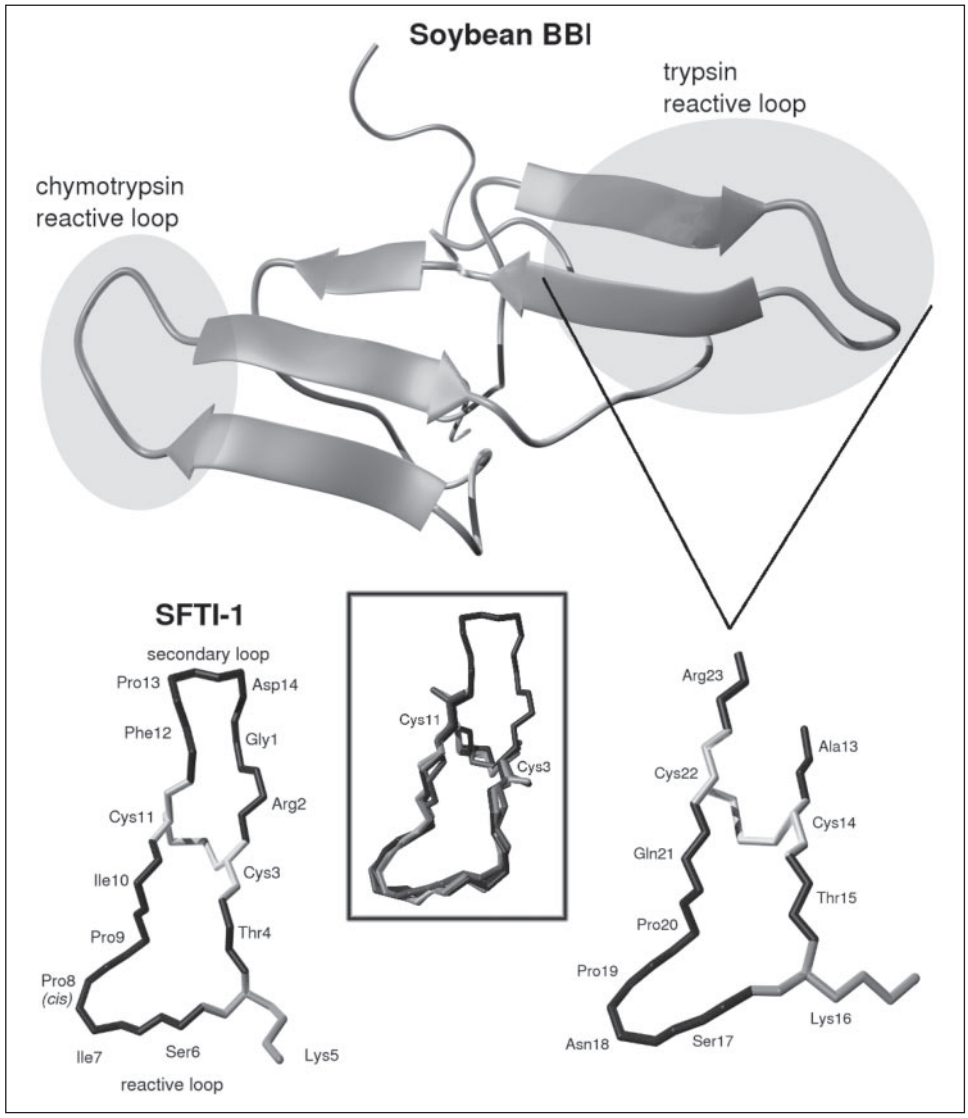

FIGURE 1. **A ribbon depiction of the structure of a BBI from *Glycine max* (soybean, PDB code 1BBI) with the chymotrypsin and trypsin-reactive loops highlighted is shown.** The detailed structure of the trypsin inhibitory loop is shown *below* the ribbon diagram and alongside the solution structure of SFTI-1 (PDB code 1JBL). In both cases the Cys residues and the lysine residue in the P1 position (45) are highlighted with the side chains depicted in *stick* form. The *inset* shows SFTI-1 (PDB code 1JBL) superimposed with the reactive loops of BBIs from mung bean (PDB code 1SMF), adzuki bean (PDB code 1TAB), and soybean (PDB code 1BBI). For orientation purposes, the Cys residues have been labeled on SFTI-1.

and a number of possibilities has been suggested (21), namely that it is nonribosomally produced in a fashion analogous to nonribosomally produced proteins in bacteria, that it is expressed as a larger protein, which may or may not bear resemblance to the BBIs, or that it is expressed as a 14-amino acid protein that is consequently cyclized. A resolution of this problem has been hampered by the small size of SFTI-1, which restricts the usefulness of PCR approaches due to the limited sequence available for degenerate primer design and the diminutive product produced. Furthermore, the cyclic nature of SFTI-1 complicates primer design because the order in which residues appear in the linear precursor is unknown. Because SFTI-1 is expressed only in seeds, isolation of cDNA containing the SFTI-1 transcript is also problematic due to the unknown timing of its expression. Accordingly, in this work, an approach utilizing bioinformatics and data base screening was perceived to be the most appropriate, and most likely to succeed, method of determining the sequence of the SFTI-1 precursor.

In this paper the sequence and structure of the linear precursor of SFTI-1 is reported. The mature peptide was first reported in 1999, but until now it was not known whether it was a gene product or nonribosomally produced. The presence of a linear precursor in the sunflower suggests that SFTI-1 is gene-expressed, and the sequence of the 52-amino acid precursor that is processed into mature cyclic SFTI-1 shares interesting parallels with the cyclotides, the largest known family of naturally occurring circular proteins. The structure reported here shows that the major structural elements of the reactive site loop of SFTI-1 are present before processing, whereas the precursor region adopts a much greater degree of conformational freedom. The sequence of the precursor further indicates that, apart from the reactive site loop, SFTI-1 does not bear any further resemblance to the BBIs.

EXPERIMENTAL PROCEDURES

Regular Expression Searches—For regular expression searching of the sunflower data base, a regular expression (RE) coding for the active site residues (CTKSIPPIC) was generated. For searching of the *Arabidopsis* genomic data base, a regular expression was generated coding for the active site residues of known dicot BBIs (Fig. 2). The scripting language PERL was used to search flat file data bases of DNA sequences in FASTA format for protein patterns corresponding to the regular expressions. The script used the BioPERL modules to translate each sequence in six reading frames and these translations were searched for the particular regular expression. If a match was made then the script returned the pattern found as well as 20 residues on either side of the pattern. If the data base being searched contained EST sequences, then the results were filtered by examining the preceding 20 amino acids for stop codons and discarding hits if these elements were present. The remaining hits were then analyzed manually using GENSCAN and SignalP. If a

genomic data base was being examined, then the contig containing the hit was retrieved and analyzed using GENSCAN; if the hit was recognized as an open reading frame, it was further analyzed with SignalP. By using this procedure, searches were made by utilizing the SFTI-1 active site RE and the BBI trypsin-reactive loop RE on a sunflower EST data base and the complete genome sequence data bases of *Arabidopsis*.

The sun-QH-All-Phred sunflower data base was downloaded from the Compositae Genome Project Data base (cgpdb.ucdavis.edu/), and the genome sequence data base for *Arabidopsis* was downloaded from the Tair website (www.arabidopsis.org/index.jsp).

Genomic PCR Analysis—DNA was isolated from sunflower seedlings using the Nucleospin Plant® kit from Machery-Nagel. This DNA was used in the PCR using primers (Proligo Australia Pty. Ltd., Lismore, New South Wales, Australia) designed from the data base hit. The forward primer coded for the sequence MATTMAK (5′-ATGGCAACCACAATGGCA-3′), and the reverse primer coded for the sequence KSIPPIC (5′-AAATCGGGGGAATCGACTTA-3′). PCR products were excised from agarose gel and cloned into pCR®2.1-TOPO vector using Invitrogen TOPO® cloning kit for sequencing.

High Pressure Liquid Chromatography—Preparative reverse phase-high pressure liquid chromatography (RP-HPLC) was performed on a Waters 600-MS controller system equipped with a Waters 484 tunable absorbance detector. Samples were manually loaded onto a Vydac C18 column (22 × 250 mm, 300-Å pore size, 10-μm particle size) and eluted at a flow rate of 8 ml/min with a linear gradient of 0–80% Buffer B (90% HPLC grade acetonitrile in H_2O, 0.05% trifluoroacetic acid) for 80 min with UV detection at 230 nm. Semipreparative RP-HPLC was performed using the same conditions and equipment with a Vydac C18 column (10 × 250 mm, 300-Å pore size, 10-μm particle size) and a flow rate of 3 ml/min. Analytical RP-HPLC was performed on a Waters 600 controller system equipped with a Waters 486 tunable absorbance detector. Samples were auto-injected using a Waters 717plus onto a Vydac C18 analytical column (4.6 × 250 mm, 300-Å pore size, 5-μm particle size). Samples were eluted at a flow rate of 1 ml/min with a linear gradient of 0–80% Buffer B for 40 min with UV detection at 215 nm.

Mass Spectrometry—All mass data were obtained using electrospray/time of flight mass spectrometry on a Mariner instrument (PerSeptive Biosystems). A 25-μl sample was injected into an Applied Biosystems 140B solvent delivery system and passed through the mass spectrometer at a flow rate of 80 μl/min with a constant organic solvent ratio of 80% Buffer B. Protein mass peaks were recorded in the linear mode between 1,000 and 25,000 m/z^+ with a step size of 0.1 or 0.2 Da and a delay time of 0.3 s. The nozzle potential was set between 60 and 100 V, and data were acquired as $[M + 2H]^{2+}$ and processed using Biospec Data Explorer 3.0.0.0 (PerSeptive Biosystems) software.

Expression of Pro-SFTI-1—The PCR was used to amplify pro-SFTI-1 from sunflower DNA using the same procedure as above. The forward primer in this reaction coded for the sequence GYKTSIS with a SapI site included at the 5′ end (5′-GCTCTTCCAACGGTTACAAAACCTCTATCTC-3′) and the reverse primer coded for the sequence PDGRP* and included a PstI site at the 5′ end (5′-CTGCAGTTATGGCCTGCCATCGGGA-3′). The PCR product was excised from agarose gel and cloned into pTWIN1 to produce a plasmid encoding pro-SFTI-1 N-terminally fused to the intein. This plasmid was transformed into *Escherichia coli* (strain ER2566) and was then grown in LB medium containing 100 μg/ml ampicillin to an A_{600} of 0.5. Protein expression was induced by the addition of isopropyl 1-thio-β-D-galactopyranoside to a final concentration of 0.5 mM, and the culture was incubated overnight at 15 °C.

After incubation the cells were harvested by spinning at 5000 × *g* for 15 min at 4 °C. The cell pellet was resuspended in Buffer B1 (20 mM Tris-HCl, pH 8.5, containing 500 mM NaCl, 1 mM EDTA, and 20 μM phenylmethylsulfonyl fluoride) and lysed by sonication. This mixture was centrifuged at 19,000 × *g* for 30 min, and 40 μl of the supernatant was analyzed by using SDS-PAGE to confirm the presence of the fusion protein in the soluble fraction. A column was packed with 10 ml of chitin beads, equilibrated with 10 bed volumes of Buffer B1 at 4 °C, and the supernatant loaded at ~1 ml/min. The column was washed with 10 bed volumes of Buffer B1. On column cleavage of the intein fusion protein was induced by overnight incubation of the column in Buffer B2 (20 mM Tris-HCl, pH 7.0, containing 500 mM NaCl, 1 mM EDTA, and 20 μM phenylmethylsulfonyl fluoride) at room temperature, and pro-SFTI-1 was eluted with Buffer B1. The eluant was further purified using preparative HPLC, and the presence of pro-SFTI-1 was confirmed by using mass spectrometry. The peak containing pro-SFTI-1 was lyophilized yielding dry peptide.

Synthesis of Pro-SFTI-1—The SFTI precursor was assembled using manual solid phase peptide synthesis with *t*-butoxycarbonyl/*O*-benzotriazole-*N*,*N*,*N*′,*N*′-tetramethyluronium hexafluorophosphate chemistry on a 0.5 mM scale. The C-terminal proline was attached to the resin via a 4-(oxymethyl)-phenylacetimido-methyl linker (*t*-butoxycarbonyl-Pro-4-(oxymethyl)-henylacetimido-methyl resin 100–200 mesh, Novabiochem), and amino acids were added to the resin by using *O*-benzotriazole-*N*,*N*,*N*′,*N*′-tetramethyluronium hexafluorophosphate with *in situ* neutralization. Cleavage of the peptide from the resin was achieved using hydrogen fluoride with *p*-cresol and *p*-thiocresol as scavengers (9:0.5:0.5 (v/v) hydrogen fluoride/cresol/thiocresol). The reaction was allowed to proceed at −5 to 0 °C for 90 min. The hydrogen fluoride was removed, and the peptide was precipitated with diethyl ether. After filtration, the peptide was solubilized in 50% Buffer A (0.05% trifluoroacetic acid/water)/Buffer B (90% acetonitrile/0.05% trifluoroacetic acid/water) and lyophilized. The crude peptide was initially purified using preparative HPLC followed by semi-preparative HPLC. Mass spectrometry was used to identify peaks containing pro-SFTI-1.

Purified pro-SFTI-1 was oxidized in ammonium carbonate with and without glutathione. In the former reaction, pro-SFTI-1 was incubated overnight in 0.1 M ammonium carbonate at room temperature. In the latter, pro-SFTI-1 was incubated overnight at room temperature in 0.1 M ammonium carbonate with 10 mM reduced glutathione and 1 mM oxidized glutathione. The products were purified with semi-preparative HPLC, and the correct peak was identified by using mass spectrometry.

NMR Spectrometry and Structure Calculations—Approximately 2 mg of dry pro-SFTI-1 was dissolved in 550 μl of 90% H_2O, 10% D_2O (99.9%, Cambridge Isotope Laboratories, Woburn, MA). Spectra were recorded at 295, 298, and 305 K on a Bruker Avance-600 NMR spectrometer. All two-dimensional spectra were recorded in phase-sensitive mode using time-proportional phase incrementation for quadrature detection in the t_1 dimension (22). Two-dimensional experiments performed included a DQF-COSY (23), a TOCSY (24) with a mixing time of 80 ms, an ECOSY (25) and NOESY spectra (26) with mixing times of 100, 150, and 200 ms. Solvent suppression for NOESY and TOCSY experiments was achieved using a modified WATERGATE sequence (27). For the DQF-COSY experiment, the water signal was suppressed by lower power irradiation during the relaxation delay (1.8 s). Spectra were routinely acquired over 8802 Hz with 4096 complex data points in F_2 and 512 increments in the F_1 dimension, with 40 scans per increment (96 for NOESY). A series of one-dimensional and TOCSY spectra were run immediately after dissolving the fully protonated sample in D_2O for identification of slowly exchanging amide protons.

Spectra were processed on a Silicon Graphics Indigo work station using XWIN-NMR (Bruker) software. The t_1 dimension was zero-filled

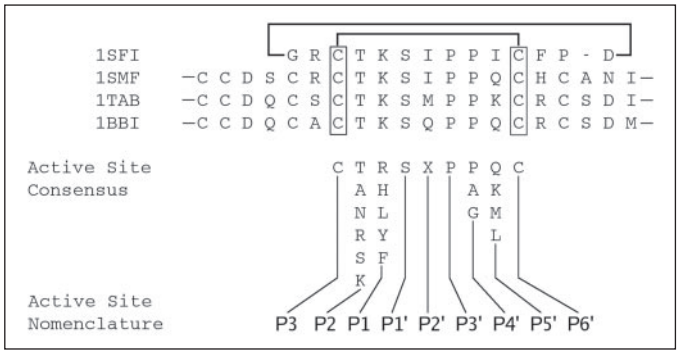

FIGURE 2. **An alignment of the active site residues from SFTI-1 and various BBIs; mung bean trypsin inhibitor (PDB code 1SMF), adzuki bean trypsin inhibitor (PDB code 1TAB), and soybean trypsin inhibitor (PDB code 1BBI).** The consensus sequence for dicot BBIs is given *below* the alignment and was derived from data in Ref. 8. The active site nomenclature of Schechter and Berger (45) is displayed at the bottom of the alignment.

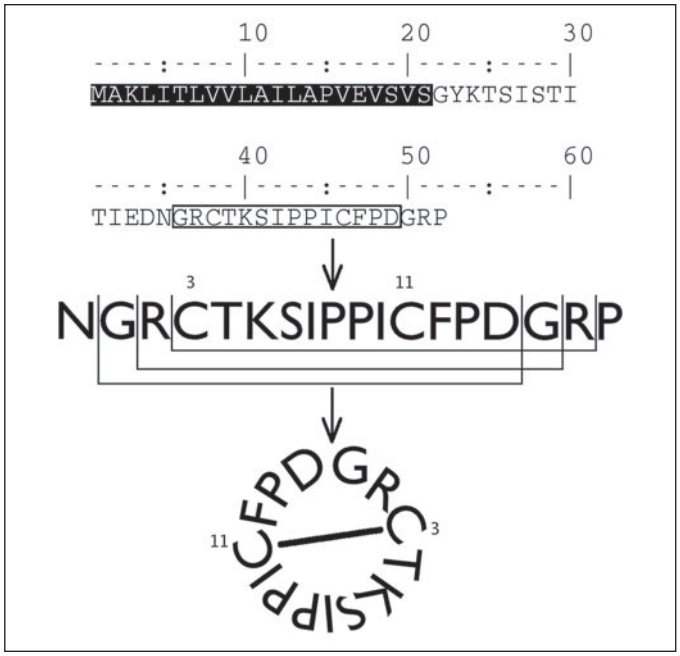

FIGURE 3. **The sequence of the 52 amino acid prepro-SFTI-1 as derived from the data base hit (*top*).** A predicted endoplasmic reticulum signal region is shown in the *black box* and the putative mature peptide in the *white box*. In order to produce cyclic SFTI-1 the precursor must be cleaved at the N-terminal and the C-terminal, but the repeat of a Gly-Arg motif at either end of the mature sequence provides three different cleavage combinations (*middle*) to produce the same cyclic peptide (*bottom*). This is similar to the situation encountered in the *Oak* clones from *O. affinis* (15).

to 2048 data points, and shifted sinebell window functions were applied prior to Fourier transformation. Chemical shifts were referenced to 2,2-dimethyl-2-silapentane-5-sulfonate sodium salt at 0.00 ppm. Two-dimensional spectra were assigned using TOCSY cross-peaks for intra-amino acid assignments and NOESY cross-peaks for sequential assignment (28). Spectra were analyzed with the program SPARKY (29).

The intensities of cross-peaks in NOESY spectra with a mixing time of 100 ms were calibrated, and distance constraints were derived using CYANA (30). Corrections for pseudoatoms were added to distance constraints where needed (31). Backbone dihedral angle restraints were derived from $^3J_{HNH\alpha}$ coupling constants measured from line shape analysis of antiphase cross-peak splitting in the DQF-COSY spectrum. Angles were restrained to $-120° \pm 40°$ for $^3J_{HNHa} > 9$ Hz. Positions of slowly exchanging amides were analyzed within preliminary structures, and in two cases hydrogen bonds were unambiguously assigned and added to the restraints. After initial structure calculations using CYANA, sets of 50 structures were calculated using a torsion angle simulated annealing protocol within CNS (32). The resulting structures were subjected to further molecular dynamics and energy minimization in a water shell (33). The final ensemble of 20 structures was deposited in the PDB as 2AB9.

Platelet Aggregation Assays—Platelet aggregation assays were performed using human platelet-rich plasma by drawing blood from a human volunteer who had not had aspirin or related products for at least 14 days. The aggregation assay was performed by adding 50 μl of a 0.6 mM solution of cyclic SFTI-1 to 400 μl of platelet-rich plasma at 37 °C. After incubation for 5 min 25 μl of 0.2 mM adenosine diphosphate was added and the light transmittance recorded.

RESULTS

Regular Expression Searching—SFTI-1 contains only 14 amino acids, of which only seven share sequence identity with the BBIs. Accordingly, BLAST searches are ineffective when searching for homologous sequences due to the short query sequence. Hence, to locate proteins that contain sequence identity to the active site of SFTI-1 and the Bowman-Birk inhibitors, the RE capabilities of the Perl scripting language were used to examine various plant data bases. Using the sequence of SFTI-1 and the active site residues of known Bowman-Birk inhibitors, an RE was generated that coded for the seven-residue loop that constitutes the active site in SFTI-1 and the BBIs (Fig. 2). This RE was used in a Perlscript that translated each data base entry in six reading frames and searched each translation for the pattern specified in the RE. By utilizing this approach, it was hoped that both the precursor of SFTI-1 would be identified in sunflower sequence data bases and that peptides similar to SFTI-1 would be identified in the model plant *Arabidopsis*. The search of the *Arabidopsis* genome yielded no significant hits using the RE, but the same search of a sunflower EST data base produced a significant hit for a gene that encoded the full sequence of mature SFTI-1 embedded in a 52-amino acid precursor protein.

Description of Identified Genes—Fig. 3 sets out the putative precursor of SFTI-1, denoted prepro-SFTI-1, identified in RE searches of the sunflower EST data base. The program SignalP (34) was used to identify signal sequence, and output from this program suggested that the precursor contains a 21-amino acid N-terminal signal sequence. Following the signal sequence, a 14-amino acid pro-peptide precedes the 14 amino acids that putatively constitute the mature SFTI-1 sequence. The mature sequence is followed by a three-residue tail. Most interestingly, the first two residues of this short tail (Gly-Arg) are the same as the first two N-terminal residues of the mature sequence. Consequently, the exact processing points of SFTI-1 cannot be identified. As processing of SFTI-1 from this precursor must include two proteolytic cleavages at the N and C termini prior to cyclization, these cleavages can occur at three separate positions to produce the same cyclic peptide (see Fig. 3). As an aside, this highlights the "seamlessness" of backbone cyclization as a post-translational modification and mirrors the situation in the cyclotides, in which the linear precursors coded in the Oak1, Oak2, and Oak3 clones could in principle be processed in four different positions to produce the same cyclic peptide (15).

To confirm the presence of the data base hit in the sunflower, genome primers were designed for the upstream region of the data base hit and for a region within the mature SFTI-1 sequence. The PCR was then used to amplify the gene from sunflower DNA. The PCR produced two bands of ~150 and 1000 bp. Both of these bands were excised from the agarose gel and cloned for sequencing. The smaller band contained the same sequence predicted by the data base mining, but the larger band contained sequence coding for the beginning of the gene in one reading frame, a large insert, and the end of the gene in a different reading frame. Accordingly, it would appear that the sunflower genome contains a

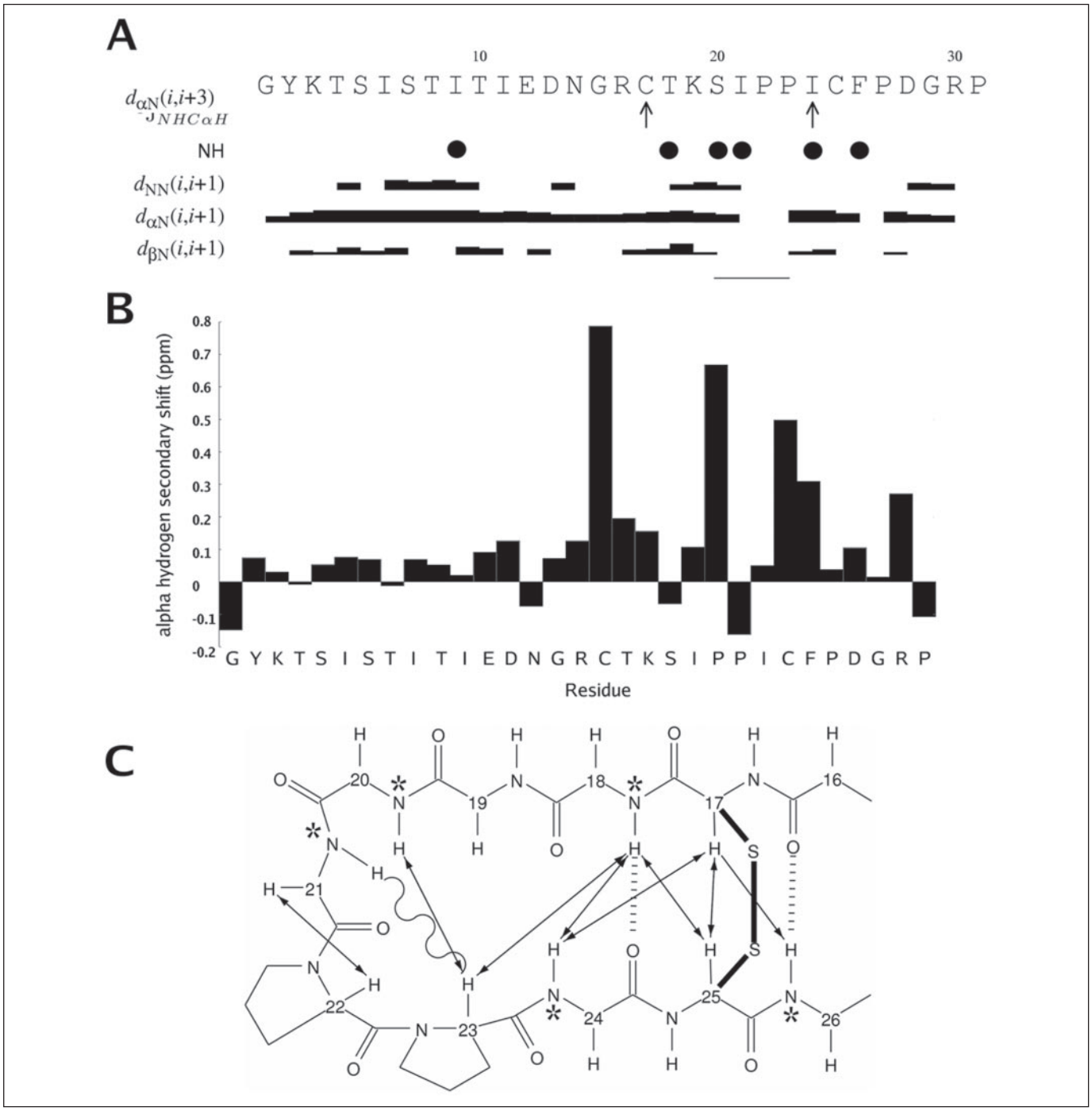

FIGURE 4. *A,* summary of the short and medium range NOEs, coupling constants, and slowly exchanging amide protons for pro-SFTI-1. Slowly exchanging NH protons (observed in TOCSY spectra recorded in ^2H_2O) are indicated by *filled circles*. Residues with a $^3J_{HNH\alpha}$ coupling constant greater than 8.5 Hz are marked with an ↑. Below the coupling constant row is a summary of sequential and medium range NOE connectivities. *Filled bars* indicate sequential connections, and the height of the bar indicates the strength of the NOE. *B,* deviation of the Hα chemical shifts from random coil values for each residue in pro-SFTI-1. *C,* schematic representation of the β-sheet region of pro-SFTI-1. Interstrand NOEs are shown as *arrows*; slowly exchanging amide protons (after dissolution in ^2H_2O) are indicated with an *asterisk*, and potential hydrogen bonds are depicted with *broken dashes.* Sequential NOEs are not shown for clarity, apart from an αH-αH NOE that is consistent with a *cis* peptide bond between residues 21 and 22. An expected NOE between Ile^{21}HN and Pro^{23}Hα that was obscured by overlap is shown with a *wavy line.*

complete SFTI-1 precursor gene, as well as one that has been destroyed some time in the past by an insertion. The presence of a short repeat (CTAAG) at either end of the insertion suggests that the insertion was the result of transposon activity.

Expression and Synthesis of Pro-SFTI-1—By using the output of SignalP, the signal sequence of the identified prepro-SFTI-1 was deemed to end immediately prior to Gly^{22}. To obtain a correctly folded sample of the pro-peptide, two approaches were utilized. First, pro-SFTI-1 was synthesized using solid phase chemistry and oxidized in ammonium carbonate in two reactions with and without glutathione. For both reactions, RP-HPLC and electrospray/time of flight mass spectroscopy was used to separate the major isoform from minor peaks. Second, the same sequence was cloned into pTWIN1 and expressed in *E. coli* as an N-terminal mini-intein fusion protein incorporating a chitin binding domain. Expressed pro-SFTI-1 was then loaded onto a chitin column and autocatalytically cleaved from the mini-intein tag using a combination of pH and temperature. Once again RP-HPLC and electrospray/time of flight mass spectroscopy was used to purify pro-SFTI-1. Analysis of all sam-

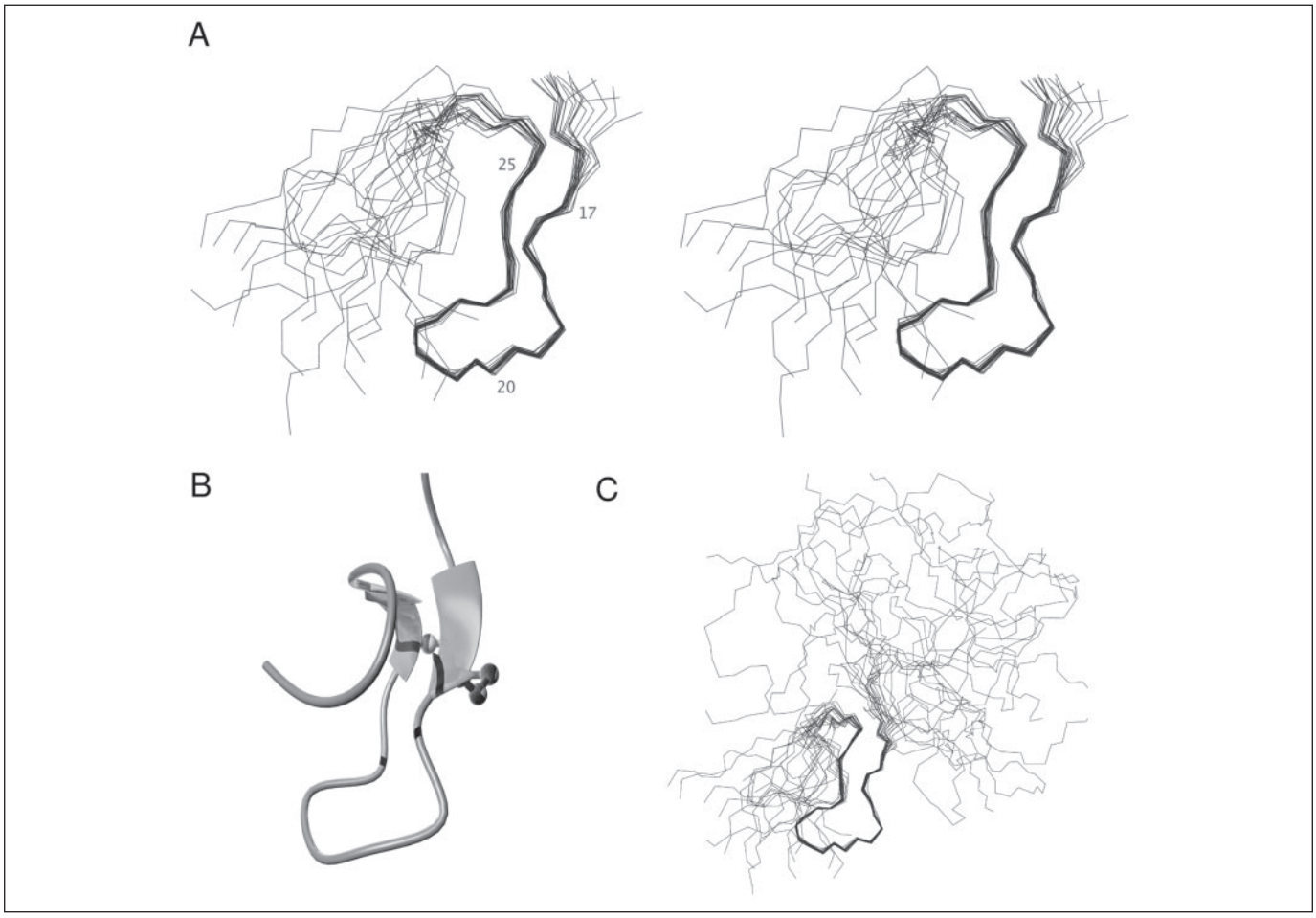

FIGURE 5. *A,* stereoview of the well defined reactive site loop of pro-SFTI-1 (PDB ID 2AB9). Residue numbers are indicated on the *left-hand view*. *B,* ribbon diagram of pro-SFTI-1 showing the two-stranded β-sheet and short tail. The disulfide bond between Cys^{17} and Cys^{25} is depicted in *ball* and *stick*. *C,* the 20 lowest energy structures of pro-SFTI-1 aligned over the reactive site loop. The disorder in the precursor region is clearly visible.

ples with analytical RP-HPLC and mass spectrometry showed that all samples possessed the correct mass and the same elution time under similar conditions.

^{1}H NMR Resonance Assignments—Pro-SFTI-1 was dissolved in water, and NMR spectra were recorded. The amino spin systems were identified using TOCSY and DQF-COSY spectra recorded at 298 K. Despite some overlap in the precursor region between residues 1 and 14, complete sequential assignments were made using TOCSY and NOESY spectra recorded at 298, 295, and 307 K. The difference between the secondary shifts of the backbone Hα protons of the individual spins systems and the Hα chemical shift of the same residue in a random coil peptide is a guide to the secondary structure present in a particular protein (35). When combined with short and medium range NOEs, coupling constants, and slowly exchanging amide protons, an indication of the major forms of secondary structure can be derived. A summary of this information for pro-SFTI-1 is presented in Fig. 4.

The major form of secondary structure in pro-SFTI-1 is a short well defined antiparallel β-sheet between residues 17 and 26 with a turn centered on Ile^{21}. A series of positive secondary shifts is generally indicative of extended structure, and positive shifts between Arg^{16}–Lys^{19} and, to a less extent, Cys^{25}–Phe^{26} suggest that these residues comprise two parts of the antiparallel sheet. This is supported by the pattern of inter-strand NOEs and large coupling constants for residues Cys^{17} and Ile^{24}. Analysis of slow exchanging amide protons indicates that main chain hydrogen bonds predicted for this conformation are present between Thr^{18}HN and Ile^{24}O as well as Phe^{26}HN and Arg^{16}O. A bifurcated hydrogen bond between the side chain carboxyl of Thr^{18} and the backbone amides of Ser^{20} and Ile^{21} has been identified in cyclic SFTI-1 (6, 11), and the presence of slowly exchanging amide protons in these residues is consistent with the presence of the same arrangement in pro-SFTI-1. Furthermore, the presence of a strong NOE between the Hα protons of Ile^{21} and Pro^{22} indicates that this peptide bond is in the *cis* conformation which is typical of SFTI-1 (6, 11) and dicot BBIs (21). Overall the secondary structure predicted here is consistent with the secondary structure of the mature cyclic SFTI-1 and indicates that the active site loop of pro-SFTI-1 adopts a similar structure to cyclic SFTI-1.

As mentioned above, assignment of the spectra for pro-SFTI-1 was complicated by overlap in residues comprising the precursor region of the protein. Furthermore, only sequential NOEs were identified in this region with no long or medium range interactions seen in the spectra, suggesting that this part of the protein is essentially unstructured. Confirmation of the unstructured nature of the precursor region was achieved by analysis of secondary Hα chemical shifts for residues 1–13. Fig. 4 shows that, with the exception of Gly^1 and Asp^{13}, the shifts do not depart from random coil values by more than 0.1 ppm, indicating that the precursor region of pro-SFTI-1 contains no defined secondary structure and is essentially random coil.

Three-dimensional Structure—A set of 50 structures was calculated by simulated annealing using 115 inter-residue NOE distance restraints and 10 dihedral angle restraints. As predicted by analysis of the secondary Hα shifts, NOE connectivities, and slowly exchanging amides, the structure of pro-SFTI-1 has a poorly defined region between residues 1

TABLE ONE

Geometric and energetic statistics for pro-SFTI-1

The values are given as mean ± S.D. for the ensemble of the 20 final solution structures. Experimental distance restraints include only interresidual NOEs.

Energies (kcal mol^{-1})	
Overall	−961.05 ± 35.28
Bonds	6.79 ± 0.95
Angles	35.09 ± 7.85
Improper	4.51 ± 1.31
van der Waals	−21.44 ± 13.78
NOE	7.23 ± 2.40
cDih	0.06 ± 0.06
Dihedral	98.11 ± 8.20
Electrostatic	−1091.41 ± 43.20
r.m.s.d.	
Bond (Å)	0.004 ± 0.0002
Angle (°)	0.51 ± 0.05
Improper (°)	0.35 ± 0.05
NOE	0.03 ± 0.006
cDih	0.26 ± 0.20
Pairwise r.m.s.d. (Å)[a]	
Backbone/heavy	0.28 ± 0.11
	1.18 ± 0.31
Experimental data	
Distance restraints	115
Dihedral restraints	10
NOE violations exceeding 0.20 Å	5 (largest 0.25)
cDih violations exceeding 2.0	0
Ramachandran[a]	
Most favored	76.00%
Additionally allowed	24.00%
Disallowed	0.00%

[a] Statistics quoted apply only to the defined reactive site loop between residues 17 and 25.

and 16, a well defined domain between residues 17 and 25, and a flexible tail region. Fig. 5 shows the three-dimensional structure of pro-SFTI-1, including a superimposition of the 20 lowest energy structures after energy minimization in a water box. Residues 17–25 correspond to the reactive loop of mature SFTI-1, and over this region the structures show excellent convergence with an r.m.s.d. of 0.28 ± 0.11 Å over backbone atoms and 1.18 ± 0.31 Å for all heavy atoms. There was no divergence from ideal covalent geometry in this region. Over the whole molecule the structures fit the experimental restraints with minimal violations. There were five NOE violations greater than 0.2 Å (largest 0.25 Å) and no dihedral angle restraint violations greater than 2° present in the set of lowest energy structures. A summary of the structural statistics is given in TABLE ONE.

Analysis of the 20 lowest energy structures with PROMOTIF indicates that the major structural elements are two β-strands joined by a type IV β-turn in the region corresponding to the reactive site loop. The absence of any NOE connectivities between the precursor region and the well defined domain of pro-SFTI-1 indicates that this region is essentially extended away from the β-sheet, and the local r.m.s.d. for residues 1–16 is increased significantly, reflecting the lack of short and long range NOE restraints. The tail of the molecule, between residues 26 and 31, also exhibits an increase in local r.m.s.d., although the presence of NOE interactions between the side chain of Arg^{30} and Pro^{22} suggests that it may have a tendency to fold back over the β-sheet region.

Platelet Aggregation Assays—The mature sequence of SFTI-1 contains a reverse RGD sequence. To investigate the possibility that cyclic SFTI-1 interacts with integrin receptors, platelet aggregation assays were conducted on mature cyclic SFTI-1. No inhibitory activity was shown, implying that cyclic SFTI-1 does not interact with integrins.

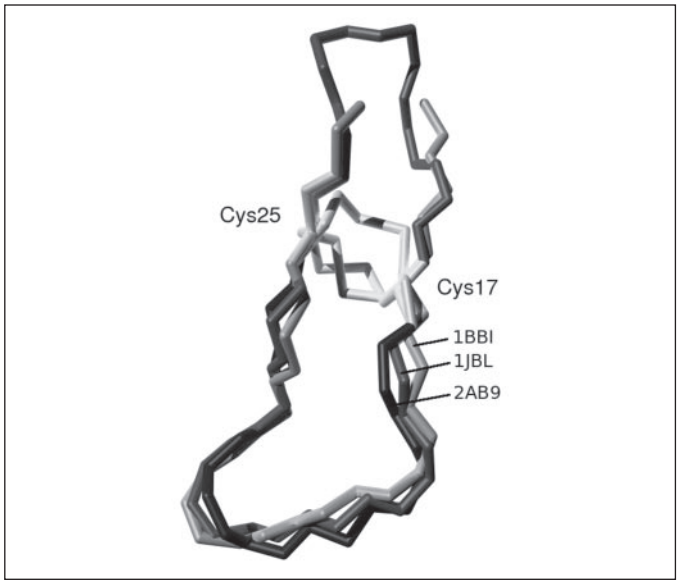

FIGURE 6. **Superimposition of the reactive site loops of pro-SFTI-1 (PDB ID 2AB9), cyclic SFTI-1 (PDB ID 1JBL), and the trypsin inhibitory loop of the soybean BBI (PDB ID 1BBI).** Cys residues are highlighted, and the Cys numbering of pro-SFTI-1 is indicated for reference.

DISCUSSION

The mechanism, or mechanisms, by which linear precursors are processed into cyclized proteins in the plant kingdom is still poorly understood. The linear precursors of only one class of plant cyclic peptide, namely the cyclotides, have been identified (15, 16), but the mechanism of cyclization remains unknown. The work presented here demonstrates that SFTI-1 is derived from a 52-amino acid, gene-expressed precursor protein and that the major structural elements of mature SFTI-1 are present in the precursor. Additionally, certain sequence similarities between this precursor and the linear precursor of the cyclotides suggest that a similar mechanism of cyclization operates in both plant species. The phylogenetic isolation of SFTI-1 from its BBI cousins is further confirmed in this study and raises interesting questions regarding the evolution of this fascinating circular peptide.

Structural Comparison of Pro-SFTI-1 with SFTI-1 and BBI Active Site Loops—The family of structures of pro-SFTI-1 shows high convergence in the region of the molecule corresponding to the reactive site loop of mature cyclic SFTI-1. As is the case with cyclic SFTI-1, the major structural element is a short two-stranded β-sheet, although in pro-SFTI-1 this β-sheet is truncated to the region about the disulfide bonds. Outside of this region, particularly in the precursor region, the molecule is essentially unstructured. The structure of the reactive site loop in pro-SFTI-1 adopts a similar conformation to the corresponding region in cyclic SFTI-1 and the trypsin inhibitory loop of the BBIs. This structural similarity can clearly be seen in Fig. 6, which depicts a superimposition of the reactive loops from pro-SFTI-1, cyclic SFTI-1, and the trypsin inhibitory loop of the soybean BBI. The backbone r.m.s.d. between the three molecules over this region of 0.65 Å underscores the structural resemblance.

Given the similarity in the reactive site loop between pro-SFTI-1 and cyclic SFTI-1, it is likely that this region also exhibits the conformational rigidity shown by cyclic SFTI-1. The existence of slowly exchanging amides in residues 18, 20, 21, 24, and 26 and the pattern of NOE con-

nectivities across the β-sheet suggests that the network of cross-bracing hydrogen bonds that stabilizes the reactive site loop of SFTI-1 is also present in pro-SFTI-1. The structure reported here shows that the basic structure of the reactive site loop of cyclic SFTI-1 has already taken shape before the processing of the pro-peptide.

The sequence of prepro-SFTI-1 illustrates that the processing point where the N and C termini of linear SFTI-1 are ligated is located in the secondary loop. As would be expected, the structure of this loop is greatly altered to accommodate the precursor region and the short tail region. Although the presence of a GR motif at either end of the mature sequence makes the precise location of the processing points ambiguous, it is clear that cleavage of pro-SFTI-1 will occur in unstructured regions. A recent report suggested that protein digestion was enhanced by the presence of unstructured regions (36), and the lack of structure in the areas of pro-SFTI-1, in which cleavage must occur may, likewise assists in the processing of the precursor.

Processing of Pro-SFTI-1—The only other linear precursor of cyclized plant proteins that has been characterized is that of the cyclotides. In this case the presence of a conserved Asn/Asp at the C-terminal processing point has led to the suggestion that asparaginyl proteases are involved in the cyclization reaction (15). The members of this class of protease, collectively known as legumains, are reported to cleave various proteins in the secretory pathway by proteolytic cleavage on the C-terminal side of Asn residues (37–39). Most interestingly, an asparaginyl endopeptidase is also suspected of mediating a transpeptidation reaction in the processing of the concanavalin A precursor in *Canavalia ensiformis* (jack bean) (40). In this case the enzyme mediates the Asn-specific excision and transpeptidation of the precursor to produce an inversion of peptide fragments in the mature peptide.

As discussed above, there are three possible processing points that could produce mature SFTI-1. One of these processing points would require cleavage after Asn to form the N terminus and after Asp to form the C terminus. It is possible that an asparaginyl protease could mediate both the cleavage of SFTI-1 from the precursor and the subsequent ligation of the free termini at this position in the precursor. The two other possible processing points involve cleavages after Gly or Arg residues at both the N and C termini. Cleavage after Arg residues could be mediated by trypsin or a trypsin-like enzyme. Although in that case involving the scissile bond (Lys^5–Ser^6), the recent report (17) of the *in vitro* cyclization of linearized SFTI-1 by trypsin suggests the involvement of a similar mechanism in the biosynthesis of SFTI-1. It is thought that the rigid structure of SFTI-1 prevents conformational change upon linearization, and consequently the equilibrium of the catalyzed reaction shifts toward the reformation of the peptide bond (17). Given that the reactive site loop of pro-SFTI-1 is well defined prior to processing, it is possible that after cleavage the linear form of SFTI-1 produced is sufficiently rigid to allow a similar process to occur, in this case involving an asparaginyl protease or a trypsin-like protease depending on the exact location of the processing points.

Although the involvement of a protease is a compelling idea, it is also possible that after cleavage of the precursor cyclization could proceed via a succinimide or cyclic anhydride, to yield a cyclic peptide without the need for a processing enzyme. Although a reaction such as this has not been observed *in vitro*, the presence of an iso-Asp in SFTI-1 linearized between Gly^1 and Asp^{14} highlights the increased reactivity of this location (17). If pro-SFTI-1 is cleaved after Asn^{14} N-terminally and after Asp^{28} C-terminally then the reactivity of this site could mediate the autocatalytic ligation of the free termini, thus obviating the need for a cyclizing enzyme.

SFTI-1 and the BBIs—The sequence reported here also demonstrates that, apart from the reactive loop residues, SFTI-1 does not appear to share any other similarities with the BBIs. This rules out the possibility that SFTI-1 is processed from a larger BBI-like protein but does not rule out the possibility of an evolutionary relationship between the two classes of protein. However, if SFTI-1 is evolutionarily related to its larger cousins, then it would be reasonable to expect that other members of the Compositae would contain SFTI-1 or BBI-like proteins. A study of the protease content of seeds from the Compositae found no BBI-like proteases in this family (12). Accordingly, if SFTI-1 and the BBIs are related, then the gene has consequently been lost in all other species from the Compositae. Despite this, it is intriguing that the study of trypsin inhibitors in the Compositae did partially characterize an unidentified trypsin inhibitor in *Zinnia elegans* with an M_r of 11,350 that shared limited homology with the BBIs (17). Likewise, a number of trypsin inhibitory molecules with molecular weights of ~7600 were also identified in sunflower seeds but were not further characterized. In principle these masses could represent BBI-like proteins, or descendants of BBIs, within sunflower seeds. One possible hypothesis could be the separate evolution of a single-headed BBI within the Compositae from an ancestral single-headed monocot BBI. The single-headed variant is thought to be ancestral to modern BBIs (41), and derivation of SFTI-1 from a single-headed descendant in the Compositae is certainly possible.

A number of other explanations are possible to explain the phylogenetic isolation of SFTI-1. The reactive site loop of SFTI-1 may represent an example of convergent evolution. Another possibility is a domain swap placing the reactive loop of a BBI into a novel precursor protein (43) sometime prior to the divergence of the Compositae from the Fabaceae. Although such events are not common (42), the repetition of reactive domains in monocot and dicot BBIs shows that gene duplication is an active force in BBI evolution (8, 44), and displacement of the reactive loop may have occurred through serendipity or by an as yet unidentified mechanism. Finally, the reactive site loop may be an example of gene transfer; one conceivable mechanism by which such a process could occur is via the accidental incorporation of the reactive site loop into a viral DNA.

In conclusion, the field of backbone cyclized proteins has expanded rapidly since the first discovery a decade ago, and the discovery of the linear precursor of SFTI-1 offers an opportunity to better understand the process by which these fascinating proteins are produced.

Acknowledgment—We thank Dr. Paul Masci for assistance and advice in the platelet aggregation assay.

REFERENCES

1. Trabi, M., and Craik, D. J. (2002) *Trends Biochem. Sci.* **27,** 132–138
2. Zhou, H. X. (2004) *Acc. Chem. Res.* **37,** 123–130
3. Felizmenio-Quimio, M. E., Daly, N. L., and Craik, D. J. (2001) *J. Biol. Chem.* **276,** 22875–22882
4. Craik, D. J., Simonsen, S., and Daly, N. L. (2002) *Curr. Opin. Drug Discovery Dev.* **5,** 251–260
5. Konarev, A. V., Anisimova, I. N., Gavrilova, V. A., and Shewry, P. R. (1999) in *Genetics and Breeding for Crop Quality and Resistance* (Mugnozza, G. T. S., Porceddu, E., and Pagnotta, M. A., eds) pp. 135–144, Kluwer Academic Publishers Group, Dordrecht, Netherlands
6. Luckett, S., Garcia, R. S., Barker, J. J., Konarev, A. V., Shewry, P. R., Clarke, A. R., and Brady, R. L. (1999) *J. Mol. Biol.* **290,** 525–533
7. Birk, Y. (1985) *Int. J. Pept. Protein Res.* **25,** 113–131
8. Prakash, B., Selvaraj, S., Murthy, M. R., Sreerama, Y. N., Rao, D. R., and Gowda, L. R. (1996) *J. Mol. Evol.* **42,** 560–569
9. McBride, J. D., Watson, E. M., Brauer, A. B., Jaulent, A. M., and Leatherbarrow, R. J. (2002) *Biopolymers* **66,** 79–92
10. Song, H. K., Kim, Y. S., Yang, J. K., Moon, J., Lee, J. Y., and Suh, S. W. (1999) *J. Mol. Biol.* **293,** 1133–1144
11. Korsinczky, M. L., Schirra, H. J., Rosengren, K. J., West, J., Condie, B. A., Otvos, L., Anderson, M. A., and Craik, D. J. (2001) *J. Mol. Biol.* **311,** 579–591

12. Konarev, A. V., Anisimova, I. N., Gavrilova, V. A., Vachrusheva, T. E., Konechnaya, G. Y., Lewis, M., and Shewry, P. R. (2002) *Phytochemistry* **59,** 279–291
13. Craik, D. J., Daly, N. L., Mulvenna, J., Plan, M. R., and Trabi, M. (2004) *Curr. Protein. Pept. Sci.* **5,** 297–315
14. Hernandez, J. F., Gagnon, J., Chiche, L., Nguyen, T. M., Andrieu, J. P., Heitz, A., Hong, T., Pham, T. T., and Nguyen, D. L. (2000). *Biochemistry* **39,** 5722–5730
15. Jennings, C., West, J., Waine, C., Craik, D., and Anderson, M. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10614–10619
16. Dutton, J. L., Renda, R. F., Waine, C., Clark, R. J., Daly, N. L., Jennings, C. V., Anderson, M. A., and Craik, D. J. (2004) *J. Biol. Chem.* **279,** 46858–46867
17. Marx, U. C., Korsinczky, M. L., Schirra, H. J., Jones, A., Condie, B., Otvos, L., and Craik, D. J. (2003) *J. Biol. Chem.* **278,** 21782–21789
18. Kalkum, M., Eisenbrandt, R., Lurz, R., and Lanka, E. (2002) *Trends Microbiol.* **10,** 382–387
19. Eisenbrandt, R., Kalkum, M., Lurz, R., and Lanka, E. (2000) *J. Bacteriol.* **182,** 6751–6761
20. Haase, J., and Lanka, E. (1997) *J. Bacteriol.* **179,** 5728–5735
21. Korsinczky, M. L. J., Schirra, H. J., and Craik, D. J. (2004) *Curr. Protein Pept. Sci.* **5,** 351–364
22. Marion, D., and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.* **113,** 967–974
23. Rance, M., Sorensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R., and Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun.* **117,** 479–485
24. Braunschweiler, L., and Ernst, R. R. (1983) *J. Magn. Reson.* **53,** 521–528
25. Griesinger, C., Sorensen, O. W., and Ernst, R. R. (1987) *J. Magn. Reson.* **75,** 474–492
26. Jeener, J., Meier, B. H., Bachmann, P., and Ernst, R. R. (1979) *J. Chem. Phys.* **71,** 4546–4553
27. Piotto, M., Saudek, V., and Sklen, V. (1992) *J. Biomol. NMR* **2,** 661–665
28. Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids,* pp. 130–161, Wiley-Interscience New York
29. Goddard, T. D., and Kneller, D. G. (2004) *SPARKY 3 Manual,* University of California, San Francisco
30. Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997) *J. Mol. Biol.* **273,** 283–298
31. Wüthrich, K., Billeter, M., and Braun, W. (1983) *J. Mol. Biol.* **169,** 949–961
32. Brunger, A. T., Adams, P. D., and Rice, L. M. (1997) *Structure* (*Lond.*) **5,** 325–336
33. Linge, J. P., and Nilges, M. (1999) *J. Biomol. NMR* **13,** 51–59
34. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). *Protein Eng.* **10,** 1–6
35. Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) *J. Biomol. NMR* **5,** 67–81
36. Prakash, S., Tian, L., Ratliff, K. S., Lehotzky, R. E., and Matouschek, A. (2004) *Nat. Struct. Biol.* **11,** 830–837
37. Scott, M. P., Jung, R., Muntz, K., and Nielsen, N. C. (1992) *Proc. Natl. Acad. Sci. U. S. A.* **89,** 658–662
38. Hara-Hishimura, I., Takeuchi, Y., Inoue, K., and Nishimura, M. (1993) *Plant J.* **4,** 793–800
39. Takeda, O., Miura, Y., Mitta, M., Matsushita, H., Kato, I., Abe, Y., Yokosawa, H., and Ishii, S. (1994) *J. Biochem.* (*Tokyo*) **116,** 541–546
40. Ishii, S. (1994) *Methods Enzymol.* **244,** 604–615
41. Mello, M. O., Tanaka, A. S., and Silva-Filho, M. C. (2003) *Mol. Phylogenet. Evol.* **27,** 103–112
42. Fliess, A., Motro, B., and Unger, R. (2002) *Proteins* **48,** 377–387
43. Zhu, S., Darbon, H., Dyason, K., Verdonck, F., and Tytgat, J. (2003) *FASEB J.* **17,** 1765–1767
44. Odani, S., Koide, T., and Ono, T. (1986) *J. Biochem.* (*Tokyo*) **100,** 975–983
45. Schechter, I., and Berger, A. (1967) *Biochem. Biophys. Res. Commun.* **27,** 157–162

Discovery, Structural Determination, and Putative Processing of the Precursor Protein That Produces the Cyclic Trypsin Inhibitor Sunflower Trypsin Inhibitor 1

Jason P. Mulvenna, Fiona M. Foley and David J. Craik

J. Biol. Chem. 2005, 280:32245-32253.
doi: 10.1074/jbc.M506060200 originally published online July 21, 2005
