Cloning, Expression, and Crystallization of Jack Bean (Canavalia ensiformis) Canavalin

Joseph D. Ng, Tzu-Ping Ko, and Alexander McPherson*

Department of Biochemistry, University of California, Riverside, California 92521

Canavalin is the major storage protein of the jack bean (Canavalia ensiformis) and belongs to the classical vicilin fraction. A full-length cDNA for canavalin was generated by the polymerase chain reaction. The nucleotide sequence coding for canavalin and the corresponding amino acid sequence were determined and shown to be homologous with those of other seed storage proteins. The amino acid sequence contained an internal sequence duplication corresponding to the structural redundancy in the monomer demonstrated by crystallographic analysis. The coding region of the canavalin cDNA was inserted into a T7 RNA polymerase expression vector and used to transform Escherichia coli. A recombinant protein with a molecular mass of 47 kilodaltons was expressed and purified to 95% homogeneity. The protein exhibited the same physical, immunological, and biochemical properties as native jack bean canavalin. Recombinant canavalin, following treatment with trypsin, was crystallized in two forms. Crystals of a rhombohedral habit grew to 1 mm in the longest dimension and diffracted to beyond 3-Å resolution. Three-dimensional diffraction data demonstrated crystals of the recombinant protein to be isomorphous with crystals of the natural plant protein, thereby confirming the identity of their structures.

The name canavalin was given in 1919 by the renowned biochemist J.B. Sumner to an amorphous protein precipitate that he had isolated from defatted meal of the jack bean (Canavalia ensiformis). In the same paper (Sumner, 1919), he also described two other proteins that he had not only purified but crystallized; these were concanavalin A and concanavalin B. Both concanavalin A and B were subsequently characterized by a number of techniques and the three-dimensional structures of both of these proteins are now known by x-ray diffraction (Reeke et al., 1975; Morrison et al., 1984; McPherson, unpublished).

In spite of repeated attempts, Sumner himself was unable to crystallize canavalin and it was otherwise given little attention. In 1936, however, S. Howell, a student working in Sumner's laboratory, left a solution of canavalin for an extended period of time on the benchtop without regard for sterility. He subsequently discovered masses of large rhombohedral crystals on the bottom of the flask. He had crystallized canavalin (Sumner and Howell, 1936).

Suspecting that the crystallization had somehow been affected by the action of degradative enzymes produced by microbes, Sumner and Howell subsequently succeeded in reproducing the crystals from sterile solutions by the direct addition of trypsin, chymotrypsin, and several other common proteases. Sumner conjectured that the crystalline protein resulted from the hydrolysis of contaminating proteins that obstructed canavalin crystallization, and he also suggested that the crystals might be a proteolytic product of the native protein. At the time, however, he could not discriminate between the two possibilities. No biological function was ascribed to canavalin, and its biochemical features remained unknown.

Canavalin next appeared in the literature in 1974, when researchers at the Massachusetts Institute of Technology reproduced Sumner and Howell's earlier work and obtained large crystalline specimens suitable for x-ray diffraction study. In addition, they determined some of the biochemical properties of canavalin (McPherson and Rich, 1973). Later work (Smith et al., 1982) demonstrated that the native protein isolated from the plant had a monomer mol wt of about 47,000 and that the monomer was cleaved roughly in half by proteases. They further showed that three cleaved but intact monomers were organized about a perfect 3-fold axis of symmetry in a native molecule having a mol wt of about 142,000.

Not readily explained was x-ray evidence requiring that the two fragments composing a canavalin monomer be virtually identical in a structural sense to produce the exceptional pseudo-32 symmetry of the rhombohedral crystals (McPherson and Rich, 1973; McPherson and Spencer, 1975). This was perplexing because the two fragments composing the canavalin monomer were clearly derived from the amino and carboxyl terminal halves of an intact subunit. The implication, realized early on, was that the canavalin subunit of Mx = 47,000 consisted of two nearly identical structural domains related to one another by a quasi-dyad axis of symmetry.

During the course of studies on other plant seed proteins (Johnson et al., 1982), we became aware that the biochemical properties of canavalin were very similar to those of phaseolin. Exceptions were that phaseolin was glycosylated and composed of mixtures of three different but highly similar subunits, whereas canavalin had no carbohydrate and appeared to be composed of only a single kind of polypeptide (Smith et al., 1982). From this point it became clear that

Abbreviations: CD, circular dichroism; FPLC, fast protein liquid chromatography; IPTG, isopropyl-β-D-thiogalactopyranoside; LB, Luria-Bertani; PCR, polymerase chain reaction; pI, isoelectric point; QELS, quasi-elastic light scattering; ssDNA, single-stranded DNA.

1 This research was supported by grants from the National Science Foundation, the National Institutes of Health, and the National Aeronautics and Space Administration.

* Corresponding author; fax 1-909-787-3790.
canavalin was a member of the vicilin family of seed storage proteins and that its properties were shared by that general class.

Given the biological importance of canavalin and its impending structure determination, it was clear that the protein provided an ideal system for coupling X-ray techniques with the method of site-directed mutagenesis. With these tools, alterations in the polypeptide sequence could be introduced and then amended based on direct visualization of attendant structural changes. In so doing, we could achieve nutritional enhancement of the seed storage protein while minimizing structural perturbations. Thus, canavalin appears to be an excellent system for the realistic application of protein engineering. Given that legume storage proteins constitute the third largest source of dietary protein on earth (Cremer et al., 1988), we felt it to be an important system as well.

In this paper, we describe an essential component of our approach to the engineering of this protein, the cloning of the cDNA that codes for its polypeptide chain, the sequencing of the cDNA and the polypeptide, and the efficient expression and assembly of canavalin in bacteria. Finally, we describe the crystallization of recombinant canavalin, and the demonstration that its structure is essentially isomorphous with that of the native protein produced by jack bean plants.

MATERIALS AND METHODS

Growth of Plants

Jack bean (Canavalia ensiformis) seeds were purchased from Sigma Chemical Co. (Poole, Dorset, UK). The seeds were planted individually in University of California, Riverside soil mix II (Chandler, 1957). Upon germination, the plants were transferred to large pots containing the same soil type and grown to maturity in an air-filtered greenhouse at a controlled temperature of 25 to 30°C. Seeds were harvested 40 DAF (after approximately 5 months of growth) and removed from their pods, and the coatings were stripped. The beans were washed with sterile water and immediately frozen in liquid nitrogen and stored at −70°C.

Synthesis of Single-Stranded cDNA from Total RNA

Frozen beans were ground to a fine powder in the presence of liquid nitrogen in a Waring blender. Total RNA was isolated by methods described by Kirby (1968) and Chirgwin et al. (1979). ssDNA was generated from total RNA by reverse transcription (Verma, 1981) and the ssDNA was quantitated according to directions supplied by the manufacturer but with the following modification. The annealing reaction mixture was boiled for 3 min and immediately frozen in an ethanol dry ice bath for 5 min. The mixture was then thawed at room temperature prior to the sequencing labeling reaction. A total of 25 oligonucleotide primers were used to sequence both strands of the PCR-amplified cDNA with overlapping sequence runs.

Preparation of Oligonucleotide Primers for PCR and Direct Sequencing

Oligonucleotide primers were synthesized using phosphoramidite chemistry on an Applied Biosystems DNA synthesizer (model 380), or they were prepared at the Biotechnology Instrumentation Facility at the University of California, Riverside, using an Applied Biosystems DNA synthesizer (model 394). The oligonucleotide products were further purified by Applied Biosystems oligonucleotide purification cartridges as described by the manufacturer.

Synthesis of a Full-Length cDNA for Canavalin by PCR

About 150 ng of single-stranded template cDNA synthesized from total jack bean RNA by reverse transcriptase was used with primer pairs selected from the 5’ and 3’ sequence of the cDNA coding for the vicilin protein from C. gladiata (Yamauchi et al., 1988). The primers were the 5’ primer 5’-ATCATCCCTCACACTGCAATACCA-3’, and the 3’ primer, 5’-AGAGAGAAGAAAACCATGAGGTG-3’.

The reaction conditions included the standard PCR buffer (Mullis et al., 1986) and 25 pmol of each primer in 100 μL layered with an equal volume of mineral oil. PCR was carried out using a DNA Thermal Cycler (Ericomp) for 1 min and 30 s at 96°C initial denaturation, followed by 30 cycles for 30 s at 55°C, 2 min at 72°C, and 15 s at 96°C. A final extension step was performed for 2 min at 55°C and 7 min at 72°C.

Direct Sequencing of the Canavalin cDNA

The PCR product containing the amplified canavalin cDNA was extracted with an equal volume of phenol and chloroform. The DNA was precipitated with 0.3 M sodium acetate at pH 5.0 and 2 volumes of absolute ethanol. The precipitate was resuspended in water and passed through a Sepharose spin-tube column (Select 5L mini-columns from 5-Prime, 3-Prime, Inc., Boulder, CO). The DNA was again ethanol precipitated and resuspended in water. About 0.5 μg of template was used for each annealing reaction with a sequencing primer.

The purified PCR product was directly sequenced using Sanger’s dyeoxy method (Sanger et al., 1977) with the Sequenase kit version 2.0 from United States Biochemical according to directions supplied by the manufacturer but with the following modification. The annealing reaction mixture was boiled for 3 min and immediately frozen in an ethanol dry ice bath for 5 min. The mixture was then thawed at room temperature prior to the sequencing labeling reaction. A total of 25 oligonucleotide primers were used to sequence both strands of the PCR-amplified cDNA with overlapping sequence runs.

Construction of a Canavalin Expression Plasmid, pETCan

The cloning strategy employed to generate the expression system is illustrated in Figure 1. For the expression of jack bean canavalin in Escherichia coli, the restriction sites upstream (NcoI) and downstream (BamHI) of the canavalin coding sequence (initial translation at residue 26 with an additional N-terminal Met) were generated using the corresponding oligonucleotide sequences as primers for the PCR. These were the 5’ primer 5’-GGGCGCATGGCGACTCT-GGGACACAGTGG-3’ and the 3’ primer 5’-CCCGA-TCCCTCAGTTGATGTTACATAGCC-3’. The conditions for the reaction were the same as those described above.

The amplified PCR NcoI-BamHI restriction fragment of canavalin cDNA was further purified using the Sepharose spin-tube column. The DNA ends of the purified PCR product were recessed with Klenow fragment (Sambrook, 1989). The final products were extracted with phenol and chloroform and precipitated with ethanol as previously described.
was the primary host cell for the pETCan expression vector. The cells were rendered competent by methods described by Mandel and Higa (1970) and transformed with the recombinant pETCan vector (Dagert and Erhlich, 1979). The transformation mixture was spread at various dilutions on LB agar plates (Sambrook, 1989) mixed with 100 \( \mu \)g/mL of ampicillin. The plates were incubated overnight at 37°C and the resultant colonies were individually screened for positive transformants.

Bacterial colonies containing the pET-3d expression vector with the canavalin cDNA insert were identified by inoculating 3 mL of LB media containing 100 \( \mu \)g/mL of ampicillin with a single colony of bacteria grown on the LB-ampicillin plates. About 20 colonies were selected and grown in LB-ampicillin medium overnight before plasmid DNA was isolated by the rapid boiling method (Holmes and Quigley, 1981). Approximately 500 ng of plasmid isolated from each colony was loaded onto a 1.2% agarose electrophoresis gel in 100 mm Tris-borate, 2 mm EDTA and stained with ethidium bromide. The relative migration of the supercoiled DNA plasmid band was compared to 500 ng of supercoiled pET-3d without any insert. Plasmid bands from each colony that displayed migration retardations in the gel when compared with the wild-type plasmid were assumed to be positive transformants containing the pET-3d vector inserted with the canavalin coding fragment cDNA.

**Double-Stranded Sequencing of Recombinant Plasmids**

Single colonies of bacteria supporting the different plasmids were grown in 3 mL of LB media containing 50 \( \mu \)g/mL of ampicillin and incubated overnight at 37°C. Bacterial cells were collected in a 1.5-mL microcentrifuge tube and the recombinant plasmid was isolated by the boiling method. Plasmid DNAs were alkaline denatured (Limb and Pene, 1988) and then sequenced using the Sequenase kit.

**The Expression of Canavalin and Detection in *E. coli***

The pETCan recombinant plasmid was isolated and purified by the boiling method and transfected into *E. coli* strain BL21(DE3)pLysE (Studier and Moffatt, 1986) by the transformation method described above. The host cell used in these expression studies was a \( \lambda \) lysogen of BL21 in which the prophage contains a copy of the gene for T7 RNA polymerase, which is in turn controlled by the lacUV5 promoter. As a result, the induction of the recombinant protein was modulated by the concentration of IPTG. In addition, this bacteria strain also carries the pLysE plasmid containing the T7 lysozyme gene that is directed by the tet promoter from pACYC184 (Studier and Moffatt, 1986; Studier et al., 1990). Transformed BL21(DE3)pLysE cells were spread on ampicillin- and chloramphenicol-treated LBagar plates and incubated overnight at 37°C.

Three milliliters of LB medium containing 100 \( \mu \)g/mL of ampicillin and 50 \( \mu \)g/mL of chloramphenicol were inoculated with bacterial colonies containing the canavalin expression plasmid and incubated overnight. About 100 \( \mu \)L of the liquid culture of recombinant cells were used to inoculate 20 mL of LB medium treated with ampicillin and chloramphenicol, and

---

**Figure 1.** Construction of a recombinant expression vector. The restriction sites upstream (Ncol) and downstream (BamHI) of the canavalin coding sequence were generated using the primers described in “Materials and Methods” for PCR. The PCR product was purified and subcloned into the Ncol/BamHI site of the pET-3d expression vector as outlined in this figure. The PCR pETCan recombinant plasmid was isolated and purified by electrophoresis in low-melting-point agarose. Approximately 500 ng of purified PCR canavalin cDNA fragment with digested 5’ terminal Ncol and 3’ terminal BamHI sites were isolated from five initial 100-\( \mu \)L PCR reactions.

The purified PCR Ncol-BamHI restriction fragment of canavalin cDNA was ligated into the corresponding sites of the T7 RNA polymerase expression vector pET-3d (Novagen, Inc., Madison, WI) to generate a recombinant canavalin expression plasmid designated pETCan. *E. coli* strain HB101
the culture was allowed to grow to $A_{595} = 0.3$. At this point, IPTG was added to a final concentration of 1 mM to induce synthesis of the protein, and the cells were incubated for 24 h at 37°C in a shaking water bath.

Samples of the cells at 1, 2, 3, 4, 5, 6, 8, and 24 h after induction were harvested by centrifugation and disrupted by alkaline lysis as described by Birnboim (1983). Cell debris and supernatant were separated by centrifugation at 10,000g. Total cell lysates were dissolved in SDS-sample buffer (62.5 mM Tris-HCl, pH 6.8, 2% SDS, 20% glycerol, 0.2% β-mercaptoethanol, and 0.001% bromophenol blue) and analyzed by SDS-PAGE to provide an initial determination of the expression of the recombinant protein in *E. coli*.

**Isolation and Purification of Recombinant Canavalin**

Recombinant canavalin was isolated and purified from pETCan/DE2lpLysE *E. coli* cells prepared as described above. The cells were grown at 37°C in 12 L of LB broth with 100 µg/mL of ampicillin and 50 µg/mL of chloramphenicol. The cells were grown to an $A_{595}$ of about 0.2 and mixed with IPTG to a final concentration of 1 mM. After incubation for 6 h at 37°C, the induced cells were harvested by centrifugation and the bacterial pellet stored at -70°C. For isolation of recombinant canavalin, the pellet was resuspended in 500 mL of lysis buffer (0.25 M Tris-HCl, pH 7.5, and 25% [w/v] Suc) and thawed to a slurry; all subsequent steps were carried out at 4°C. The bacterial cells were disrupted by three cycles of sonication using a Heat Systems Ultrasonic cell disrupter, model W-375. The resulting lysate was centrifuged for 60 min at 25,000 rpm in a Beckman SW28 rotor. The turbid supernatant (fraction I) was retained. Fraction I was heated to 70°C for 10 min and centrifuged for 20 min at 10,000 rpm in a Beckman SS34 rotor to remove denatured protein. The supernatant (fraction II) was adjusted to pH 5.1 and incubated overnight at 4°C. The resulting solution was centrifuged for 20 min at 10,000 rpm in a Sorvall SS34 rotor and the pellet resuspended in 40 mL of buffer (50 mM Tris-Cl, pH 7.5, 0.1 M NaCl, and 1 mM EDTA), yielding fraction III.

**Protein Determination**

All protein concentrations were determined by the method of Bradford (1976) with BSA as a standard.

**Gel Electrophoresis**

SDS-PAGE was carried out according to the method of Laemml (1970) using 11% polyacrylamide gels in the presence of 0.1% SDS. Gels were stained with a solution containing 25% (v/v) methanol, 10% (v/v) acetic acid, and 300 µg/mL of Coomassie brilliant blue for 2 h at room temperature. Destaining was performed in 10% (v/v) methanol and 10% (v/v) acetic acid for 12 h under the same conditions.

Non-denaturing gels were prepared by the same methods but with SDS omitted. In addition, β-mercaptoethanol was not included in the loading buffer and the samples were not boiled prior to loading. Staining and destaining of the gels were otherwise performed in the same manner.

**Proteolytic Cleavage of Recombinant Canavalin**

Intact canavalin purified from bacteria was combined with trypsin (GIBCO-BRL), at a molar ratio of 100:1 in 50 mM Tris, pH 7.5, and 150 mM NaCl in a 20-µL reaction volume at 37°C. Recombinant canavalin was exposed to trypsin for periods ranging from 15 min to 48 h. The reactions were terminated by rapidly combining the digestion mixture with sample buffer containing 10% SDS and boiling for 3 min. Digestion products were sampled at expanding time intervals over the course of the experiment and analyzed by SDS-PAGE.

**Immunoblot Analysis**

Antisera against plant canavalin were prepared by Bethyl Labs, Inc., (Montgomery, TX) and purified by the methods of Mohanty and Elazhary (1989). The IgG was further purified by removing anti-*E. coli* antibodies by methods described by Sambrook et al. (1989). After electrophoresis, proteins from SDS-PAGE were transferred to nitrocellulose paper with an electrobloater. The primary antibodies were allowed to bind to the filters and were visualized using alkaline phosphatase-conjugated antibody as described by Granger-Schnarr et al. (1988).

**CD Spectroscopy**

CD spectra were obtained with a Jasco model J500C spectropolarimeter interfaced to a DP 500N data processor. Measurements were made at a protein concentration of 1 mg/mL in 50 mM Tris at pH 7.5. The spectrum described was the result of averaging nine scans with subtraction of a solvent blank.

**FPLC**

Samples of canavalin isolated from plants and *E. coli* were analyzed by FPLC chromatography. A ready-packed, size-exclusion column (Pharmacia Suprose 12) equilibrated with 100 mM Tris, pH 7.5, 150 mM NaCl, 1 mM PMSF, and 0.01% sodium azide was used on a Pharmacia FPLC system. Approximately 200 µL of a 10 mg/mL protein solution was used for each analysis with a pump pressure of 1.5 MPa to generate a flow rate of 0.3 mL/min. The proteins were detected by a Pharmacia Single Path UV-monitor and the chromatographic profile was plotted using a Spectra-Physics 4270 integrator.

**MS**

MS was performed by the UCR Biotechnology Instrumentation Facility using a Finnigan MAT Lasermat laser desorp-
tion, sinipinic acid, matrix-assisted, time-of-flight mass spectrometer. Analyses were made with 2.5 µg of protein in 0.5 µL of 50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM EDTA.

Characterization of Canavalin Oligomers by Inelastic Light Scattering

QELS in photon correlation mode was used to compare the particle size distribution of purified canavalin from plants with that from E. coli under non-dissociation conditions. The theory and techniques underlying these analyses are discussed in detail elsewhere (Chu, 1974; Berne and Pecora, 1976; Pecora 1985) and will not be reviewed here. QELS analysis enables the translational diffusion coefficient of particles undergoing Brownian motion to be calculated. If canavalin is assumed to have a roughly spherical shape, its hydrodynamic radius, \( r_h \), can be calculated using the Stokes-Einstein equation

\[
r_h = \frac{kT}{6\pi\eta D}
\]

with the diffusion coefficient, \( D \); temperature, \( T \); viscosity \( \eta \) of the solvent, and the Boltzmann constant, \( k \).

QELS measurements were carried out using a Malvern 4700C submicro particle analyzer (Malvern Instruments, Inc., Southborough, MA). The light source was a 1-W Innova 70-3 argon laser (Coherent, Palo Alto, CA). The photomultiplier was positioned 90° to the incident laser beam during data collection. The digital signal from the photomultiplier was processed through a 256-channel correlator and further analyzed by the exponential sampling method (Cummins and Staples, 1987) on an IBM PC computer.

The samples investigated contained 10 mg of protein per mL of buffer (100 mM Tris, pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM PMSF). Total volumes of 60 µL in square microcuvettes (Starna Cells, Inc., Atascadero, CA) were used. The protein samples were centrifuged for 15 min at 12,000g and prefiltered through 0.22-µm syringe filters (Millipore Co., Bedford, MA) prior to each analysis.

Proteolytic Modification of Canavalin for Crystallization

Purified canavalin expressed by E. coli was proteolytically modified by trypsin. A 10-ml reaction containing 1.0 mg/mL of canavalin from bacteria was mixed with trypsin to a final concentration of 0.063 mg/mL in 50 mM Tris-HCl, pH 7.5, 0.1 M NaCl, 1 mM EDTA. The reaction mixture was incubated at 37°C for 48 h. The final product was applied to a column of Sephadex G-100 and the fractions containing canavalin were pooled as judged by absorbance at 280 nm.

Cloning of Canavalin

Total RNA isolated from jack beans was sufficient for the synthesis of the single-stranded cDNA template subsequently used for the selected amplification of a full-length canavalin cDNA by PCR. Approximately 30 cycles of PCR were generally adequate to produce microgram amounts of canavalin cDNA that could then be used for sequencing and recombinant DNA manipulation.

The PCR reaction yielded an amplified DNA product that, when examined by agarose gel electrophoresis, gave the results seen in Figure 2. The DNA fragment was about 1.4 kb and appeared as a homogeneous band comprising more

![Figure 2](https://www.plantphysiol.org/)

Figure 2. The synthesis of a full-length cDNA for canavalin by PCR. Amplification of the canavalin cDNA was described in "Materials and Methods." Five-microliter and 10-µL samples of the PCR reaction were analyzed on a 1.2% agarose gel stained with ethidium bromide (lanes B and C, respectively) and compared with 0.5 µg of ΦX174 digested with HaeIII endonuclease (lane A) and 0.5 µg of λ DNA digested with HindIII endonuclease (lane D). The resulting amplified canavalin cDNA fragment was about 1.4 kb in size and was shown to be highly homogeneous.
than 90% of the total product. Approximately 5 µg of purified canavalin cDNA could be amplified in a single PCR reaction.

Sequencing of the Canavalin cDNA

High-resolution nucleotide sequences were obtained by direct sequencing of the PCR-amplified fragment. A typical sequencing reaction generally resolved 50 to 200 bp per primer. A total of 25 primers were used to sequence the entire cDNA PCR product. The sequences of the oligonucleotide primers and the directions in which they were used to sequence are listed in Table I and shown in Figure 3.

The complete nucleotide sequence of the canavalin cDNA contains a 1335-bp open reading frame coding for a polypeptide having 445 amino acid residues as well as 22-bp 5’ and 48-bp 3’ untranslated regions. The nucleotide sequence contains no glycosylation recognition sites of the type Asn-X-Thr/Ser. This is consistent with biochemical evidence that canavalin has no covalently bound carbohydrate residues, as indicated by the phenol-sulfuric acid test and the absence of amino sugars (Smith et al., 1982).

AATAAA has been shown to be a highly conserved sequence that usually appears 10 to 30 nucleotides upstream from the polyadenylation site in mRNAs (Croy et al., 1982). The hexanucleotide sequence AATAAA located 17 nucleotides upstream from the 3’ end was assumed to be that signal sequence.

The canavalin cDNA shows nucleotide and deduced amino acid sequences very similar to those of the vicilin protein from C. gladiata (Yamauchi et al., 1988). Only five nucleotides differ: (a) C to G at base 505, (b) C to T at base 889, (c) C to A at base 1058, (d) A to C at base 1166, and (e) A to T at base 1189.

Oligonucleotide primers used for cDNA sequencing

Table I. Oligonucleotide primers used for cDNA sequencing

<table>
<thead>
<tr>
<th>Lowercase a, t, g, and c correspond to bases that are noncomplementary to the template strand. F and B oligonucleotides were primers used to obtain sequences of the coding strand and the complementary strand, respectively.</th>
<th>Oligonucleotide</th>
<th>Sequence (5’-3’)</th>
<th>Location (Base Position)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ATCATCCCTCACACTGCAATACC</td>
<td>23–46</td>
<td></td>
</tr>
<tr>
<td>F1&lt;sup&gt;b&lt;/sup&gt;</td>
<td>ggggccATGCGTTTTCGCGCCGATTCCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>ggggccatgCCGACCTCGGGACACAGTGG</td>
<td>98–117</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>GAACCTGTTGACCTAAAGATGTTGATTTG</td>
<td>155–190</td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>AAGACCAACACGTTTCT</td>
<td>203–220</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>TTGAAGACTCCTGCTAGTC</td>
<td>275–292</td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td>TTGCTGAACCCTGACGGGC</td>
<td>383–400</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>GCCGTCAGGGTCACCAA</td>
<td>383–400</td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td>AGTCCTAGATAGGAAAGAA</td>
<td>545–562</td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>AAAAGACTGGCCATCCTAC</td>
<td>563–580</td>
<td></td>
</tr>
<tr>
<td>B5</td>
<td>CAGAGCTCTCTGCTCTCCT</td>
<td>635–652</td>
<td></td>
</tr>
<tr>
<td>B6&lt;sup&gt;a&lt;/sup&gt;</td>
<td>gggggatcctcaGG AGCTAG ATTGGGCATGT</td>
<td>718–736</td>
<td></td>
</tr>
<tr>
<td>F6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>ggggccatgGCCCAATCTAGCTCCAG</td>
<td>88–117</td>
<td></td>
</tr>
<tr>
<td>B7</td>
<td>TCTGGGCTAGATGGGGCG</td>
<td>722–739</td>
<td></td>
</tr>
<tr>
<td>F7</td>
<td>ACACCTTCTCCTCCAAGAT</td>
<td>743–760</td>
<td></td>
</tr>
<tr>
<td>B8</td>
<td>TGCTGTGCTCTCATCATAA</td>
<td>815–832</td>
<td></td>
</tr>
<tr>
<td>F8</td>
<td>TTGCTGGCACGCAGAAGCAAC</td>
<td>971–995</td>
<td></td>
</tr>
<tr>
<td>B9</td>
<td>GTGGCAGCGTACCTCCGCAGTTGCATACTTTCTAATCC</td>
<td>1001–1038</td>
<td></td>
</tr>
<tr>
<td>B10</td>
<td>GAAGACTGCTGTGGAAGCCA</td>
<td>1084–1101</td>
<td></td>
</tr>
<tr>
<td>F9</td>
<td>CATCTAAATAATTGTTGGG</td>
<td>1103–1120</td>
<td></td>
</tr>
<tr>
<td>B11</td>
<td>CTGCCATTACGCTTTCCT</td>
<td>1172–1189</td>
<td></td>
</tr>
<tr>
<td>B12</td>
<td>TTCACAAAGTTAGTCTC</td>
<td>1264–1281</td>
<td></td>
</tr>
<tr>
<td>F10</td>
<td>GGTACGCTTACGCTATT</td>
<td>1283–1300</td>
<td></td>
</tr>
<tr>
<td>B13</td>
<td>CTAGCTCCTTCCACCAGGGC</td>
<td>1303–1320</td>
<td></td>
</tr>
<tr>
<td>RP&lt;sup*a&lt;/sup&gt;</td>
<td>AGAGAGAAAAACCAATGATGGT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Oligomers also used as PCR primers.
The nucleotide and deduced amino acid sequences of the cloned canavalin are shown in Figure 4 along with a schematic illustration of its structure, the determination of which is described in Ko et al. (1993). The regions in the primary structure of canavalin that correspond to the specific areas in the three-dimensional structure are indicated by color. Yanamachi et al. (1988) have determined the 14-amino acid terminal sequence of canavalin purified from C. gladiata and it is identical to the sequence between the 27th and the 40th residue of the canavalin precursor. Because the canavalin cDNA sequence from C. ensiformis is almost identical to the amino acid sequence found in the mature vicilin cDNA sequence from C. gladiata, the protein sequence of the is almost identical to the sequence between the 27th and the 40th terminal sequence of canavalin purified from C. gladiata seeds. The hydrophobic residue-rich region from the first Met to Ala must, therefore, be the signal peptide, as shown in Figure 4.

There is a relatively high occurrence of G followed by A base 1213. Two base changes are silent, whereas the other three correspond to amino acid substitutions: Asn to Lys at residue 161 (AAC to AAG), Leu to Ile at residue 348 (CTC to ATC), and Asn to His at residue 382 (AAC to CAC).

The nucleotide and deduced amino acid sequences of the cloned canavalin are shown in Figure 4 along with a schematic illustration of its structure, the determination of which is described in Ko et al. (1993). The regions in the primary structure of canavalin that correspond to the specific areas in the three-dimensional structure are indicated by color. Yanamachi et al. (1988) have determined the 14-amino acid terminal sequence of canavalin purified from C. gladiata and it is identical to the sequence between the 27th and the 40th residue of the canavalin precursor. Because the canavalin cDNA sequence from C. ensiformis is almost identical to the amino acid sequence found in the mature vicilin cDNA sequence from C. gladiata, the protein sequence of the is almost identical to the sequence between the 27th and the 40th terminal sequence of canavalin purified from C. gladiata seeds. The hydrophobic residue-rich region from the first Met to Ala must, therefore, be the signal peptide, as shown in Figure 4.

There is a relatively high occurrence of G followed by A base 1213. Two base changes are silent, whereas the other three correspond to amino acid substitutions: Asn to Lys at residue 161 (AAC to AAG), Leu to Ile at residue 348 (CTC to ATC), and Asn to His at residue 382 (AAC to CAC).

The nucleotide and deduced amino acid sequences of the cloned canavalin are shown in Figure 4 along with a schematic illustration of its structure, the determination of which is described in Ko et al. (1993). The regions in the primary structure of canavalin that correspond to the specific areas in the three-dimensional structure are indicated by color. Yanamachi et al. (1988) have determined the 14-amino acid terminal sequence of canavalin purified from C. gladiata and it is identical to the sequence between the 27th and the 40th residue of the canavalin precursor. Because the canavalin cDNA sequence from C. ensiformis is almost identical to the amino acid sequence found in the mature vicilin cDNA sequence from C. gladiata, the protein sequence of the is almost identical to the sequence between the 27th and the 40th terminal sequence of canavalin purified from C. gladiata seeds. The hydrophobic residue-rich region from the first Met to Ala must, therefore, be the signal peptide, as shown in Figure 4.

There is a relatively high occurrence of G followed by A base 1213. Two base changes are silent, whereas the other three correspond to amino acid substitutions: Asn to Lys at residue 161 (AAC to AAG), Leu to Ile at residue 348 (CTC to ATC), and Asn to His at residue 382 (AAC to CAC).

The nucleotide and deduced amino acid sequences of the cloned canavalin are shown in Figure 4 along with a schematic illustration of its structure, the determination of which is described in Ko et al. (1993). The regions in the primary structure of canavalin that correspond to the specific areas in the three-dimensional structure are indicated by color.
ucts of canavalin produced by endogenous proteases, or they may be the consequence of premature translation events that give rise to truncated proteins.

**Expression of the N-Terminal and C-Terminal Domains of Canavalin**

Using the same methods that were used to produce full-length canavalin cDNA, equivalent constructs were also made, incorporating in one case (designated [N]) only the amino-terminal half of the polypeptide (amino acid residues 26–238) and in a second case (designated [C]) the carboxyl terminal half of the polypeptide (amino acid residues 234–445) (Fig. 6). As for the full-length cDNA, bacteria were transformed and grown that contained the abbreviated constructs. Because of the domain structure of canavalin, described in detail in Ko et al. (1993), we anticipated that each half of the polypeptide might fold independently to produce the correct three-dimensional structure.

SDS-PAGE and western blot analysis using antibodies conjugated with alkaline phosphatase showed that both the amino- and carboxyl-terminal domains of the protein were indeed produced by the bacteria, and in significant amounts (Fig. 6).

![SDS-PAGE analysis of total cell protein containing the expressed canavalin. Allquots from bacterial cultures at different IPTG induction times were diluted to the same absorbance (A_{600} = 0.8), mixed with an equal volume of SDS-PAGE loading buffer, and run on an 11% polyacrylamide gel. Lane M contains the Promega medium molecular mass protein standards. Lane NR contains total cell protein from a nonrecombinant bacterium. The next nine lanes contain total protein from bacteria expressing canavalin with no leader peptide sequence at 0, 1, 2, 3, 4, 5, 6, 8, and 24 h, respectively, after induction with IPTG. Each lane contains total protein from 20 μL of cells at A_{600} = 0.8 and is stained with Coomassie brilliant blue.](image)

**Figure 6.** Constructs for the expression of the N-terminal and C-terminal canavalin domains and the expression of the resultant proteins as visualized by SDS-PAGE. The N-terminal and C-terminal coding domains were each amplified by PCR with primers carrying 5' Ncol and 3' BamHI restriction sites and containing the DNA sequence of canavalin from nucleotides 98 to 736 (coding for Ala^{26} to Ser^{238}) and nucleotides 722 to 1368 (coding for Ala^{234} to Tyr^{445}), respectively. A Met residue was incorporated in the N terminus of each domain due to the created initiation codon. The 3' primer used to amplify the N-terminal domain also contained an additional TGA codon to terminate translation after Ser^{238}. The PCR products were purified and subcloned into the Ncol/BamHI site of the pET-3d expression vector. The resulting expression plasmids were transfected into E. coli and the recombinant proteins induced by IPTG. Total protein from recombinant cells induced for 3 h was analyzed on SDS-PAGE (panel 1) and immunoblotted with alkaline phosphatase-conjugated rabbit polyclonal antibodies (panel 2). Lanes N and C contain the recombinant N-terminal and C-terminal domains, respectively.

We have not yet succeeded in purifying large amounts of the two domains that would allow us to characterize the domain polypeptides in detail and perform crystallization studies. Nonetheless, using rather crude preparations and meager amounts, we have obtained microcrystalline needles of the amino-terminal domain of the protein made by recombinant bacteria. With determined effort, we believe that...
ultimately both domains can be isolated, characterized, and crystallized and their structures determined by x-ray diffraction analysis.

**Purification of Recombinant Canavalin**

Canavalin expressed in *E. coli* was rapidly purified in three steps without resort to chromatography. Table II summarizes the purification procedure. Canavalin was found exclusively in the soluble portion of the cell extract and the use of denaturants was not required for recovery at any stage. An attractive property of recombinant canavalin is that it is heat tolerant. Crude bacterial cell extract can be exposed to elevated temperature, thus denaturing contaminant proteins. These precipitate while native canavalin remains in solution. Figure 7 shows two heat-fractionated samples of cell lysate treated at 60 and 70°C (A, lanes 1 and 2) and the corresponding immunoblot prepared using polyclonal antibodies directed against canavalin (B, lanes 1 and 2). Canavalin is indicated by arrows and has a molecular mass of about 47 kD.

The optimal temperature for heat fractionation of canavalin was 70°C. Following this step, the protein was nearly 60% homogeneous according to densitometric analysis on Coomassie blue-stained SDS-polyacrylamide gels. The pl of canavalin was calculated from the amino acid sequence to be 5.4. The heat-treated fraction of the recombinant cell lysate was gradually adjusted to pH 5.1, and this selectively precipitated canavalin. The pH fractionation step yielded recombinant canavalin of about 95% homogeneity, as measured by SDS-PAGE. The fraction at each of the steps during purification, as seen on SDS-PAGE, is shown in Figure 8.

The purification procedure used in isolating the intact recombinant canavalin cannot, unfortunately, be used to purify the independent canavalin domains from recombinant bacteria. The heat resistance of the full-length protein is apparently not maintained at the domain level.

**Characterization of Recombinant Canavalin**

**IEF and Nondenaturing PAGE**

Purified canavalin from *E. coli* was compared to canavalin isolated from jack bean by IEF (Fig. 9). The pl of *E. coli* canavalin was determined to be 5.5, which agrees with the pl calculated from the sequence. Canavalin expressed in bacteria had the same pl as redissolved crystals of plant canavalin. Recombinant canavalin was compared with canavalin purified from the jack bean on nondenaturing PAGE in Figure 10. Canavalin expressed in bacteria migrated the same as plant canavalin under nondenaturing conditions.

**Proteolytic Digest Pattern**

Recombinant canavalin exhibits almost the same proteolytic digestion pattern as native plant canavalin. Figure 11 shows the time course digestion of bacterial canavalin cleaved with trypsin and visualized on SDS-PAGE. The gel shows that canavalin is converted first into two polypeptides of mol...
Figure 10. Nondenaturing polyacrylamide gel of plant and recombinant canavalin. Purified canavalin from E. coli (lane 1) was compared with purified canavalin from jack bean (lane 2).

wts 23,000 and 24,000. With the reaction conditions applied here, the cleavage of recombinant canavalin by trypsin occurs within 15 min. Prolonged exposure to trypsin results in a further breakdown of the protein within 48 h into two stable polypeptides having mol wts of about 21,600 and 22,000.

The proteolytic digestion pattern illustrated here is almost the same as the time course digestion pattern of native jack bean canavalin when treated with \( \alpha \)-chymotrypsin (Smith et al., 1982); however, the conversion of the single canavalin subunit from bacteria into two stable polypeptides occurs with the appearance initially of two transient intermediates.

Figure 8. SDS-PAGE analysis of E. coli-expressed canavalin at different stages of purification. Lane 1, Crude extract of induced recombinant bacteria (fraction I); lane 2, heat-fractionated cell lysate (fraction II); lane 3, pH fractionated cell lysate. The arrow indicates the position of canavalin.

Figure 9. IEF agarose gel analysis of canavalin from jack bean and E. coli. Native canavalin isolated from jack beans (lane 2) is compared with purified jack bean canavalin crystals (lane 3) and to purified canavalin from E. coli (lane 4) according to their pi values. Lanes 1 and 5 contained pi markers that include amylglucosidase from Aspergillus niger (pi 3.6), trypsin inhibitor for soybean (pi 4.6), \( \beta \)-lactoglobulin A from bovine milk (pi 5.1), carbonic anhydrase II from bovine erythrocytes (pi 5.9), carbonic anhydrase I from human erythrocytes (pi 6.6), and myoglobin from horse heart (pi 7.2). Both the crystalline plant canavalin and the bacterial recombinant canavalin focused at about pi 5.5. Each lane contained 5 to 10 \( \mu \)g of protein.

Figure 11. Time course digestion by trypsin of intact recombinant canavalin visualized on an SDS-11% polyacrylamide gel. Lane 0 contains purified intact canavalin from E. coli. The other lanes correspond, from left to right, to incubation times of 0.25, 0.5, 1.0, 1.5, 2.0, 4.0, 6.0, 8.0, 24, and 48 h, as indicated above each lane. This time sequence shows the conversion of recombinant monomers to polypeptide fragments having mol wts of about 21,600 and 22,000. The digestion products were compared with molecular mass markers in the lanes marked Mr.
These are present but less pronounced in the digestion of native plant canavalin.

**CD Spectroscopy**

Conformational differences between recombinant canavalin and native plant canavalin were investigated by comparing the far-UV CD spectra of the two molecules. The results are presented in Figure 12. The two proteins were almost indistinguishable in their secondary structure according to the Hennessey and Johnson algorithm (Hennessey and Johnson, 1981). The proportions of secondary structure derived from this analysis were 18% α-helix and 35% β-sheet for *E. coli* canavalin and 10% α-helix and 34% β-sheet for the canavalin purified from jack beans. The plant canavalin used in these measurements was the product of redissolved crystalline protein that was proteolytically modified. Portions of the polypeptide comprising helical structure may have been removed, thereby accounting for the 8% difference in α-helix content.

**MS**

The molecular mass of the recombinant protein was further quantitated by MS (data not shown). The molecular mass of the monomeric storage protein expressed in bacteria was found to be 47,633 D, which is comparable to the native plant canavalin reported in other studies (Smith et al., 1982).

**FPLC and Light-Scattering Analysis**

The quaternary state of canavalin expressed by *E. coli* was evaluated by FPLC size-exclusion analysis under nondissociating conditions. Recombinant canavalin displays a broad peak in the chromatographic profile in the 47-kD molecular mass range (Fig. 13A). A small fragment of about 12 kD is also seen. This is most likely a contaminating peptide. A protein fragment of similar size can occasionally be seen when canavalin is analyzed on SDS-PAGE (Fig. 8, lane 3).

Another possibility is that intact canavalin may have been acted upon by contaminating proteases that produced the fragment. Canavalin from *E. coli* that has been trypsin treated, crystallized, and redissolved shows several peaks that have retention times corresponding to approximately 24, 47, 145, and >300 kD molecular masses (Fig. 13B). As for intact canavalin, the 47-kD product is the predominant species. The 145-kD product would correspond to the trimer. We do not know if the very large peak represents specific oligomers or simply nonspecific aggregation.

Aggregates that are greater than 300 kD are not observed for the intact canavalin. This suggests that proteolytic proc-
cessing may be required for aggregation to occur. The peak corresponding to the 24-kD species almost certainly represents the major fragments of the cleaved monomer.

Crystalline plant canavalin appears to exist in only two multimeric states. The FPLC profile of redissolved crystals of plant canavalin shows two peaks of molecular mass of about 145 and 300 kD (Fig. 13C). The 145-kD species, which is about 85% of the sample, undoubtedly is the trimer. The minor species is apparently a dimer of the trimeric protein.

Size distributions under nondissociating conditions were estimated using QELS in photon correlation spectroscopy mode. The results showed that solutions containing redissolved crystals of plant canavalin exhibited aggregates with an average diameter of 93 ± 5 Å with a polydispersity of 0.073. According to the structure of jack bean canavalin derived by x-ray diffraction analysis (see Ko et al., 1993), the diameter of the trimeric form of the globulin is about 90 ± 5 Å. This is in good agreement with the molecular mass estimation of 145 kD for the trimer by FPLC analysis. The higher aggregate class seen in the FPLC profile (300-kD fragment) is not observed by light scattering.

Trypsin-treated canavalin from E. coli, on the other hand, exhibits a very large aggregate of about 3000 Å mean diameter with a polydispersity of 0.397. The large particle size observed here likely corresponds to the large aggregates found by FPLC (Fig. 13B). Under the conditions investigated, some part of the canavalin isolated from bacteria appears to form aggregates of a higher order than that purified from plants.

Crystallization of Recombinant Canavalin

A characteristic of plant canavalin is that once it has been cleaved with trypsin it readily crystallizes upon exposure to pH 6.8 in the presence of 0.7 to 2.0% NaCl. This procedure was applied to recombinant canavalin purified from bacteria. The results are illustrated in Figure 14a. As with the plant protein, large rhombohedral crystals grew within a matter of a few hours to a day at 22°C. These crystals appear identical in almost all regards to those obtained from the plant protein. In addition, a second crystal form of recombinant canavalin was also grown under similar conditions (Fig. 14b). These crystals are long needles as large as 1.5 mm in length. This particular crystal form has not been observed for plant canavalin.

Preliminary X-Ray Analysis of Recombinant Canavalin Crystals

A preliminary x-ray diffraction analysis was carried out on the rhombohedral crystals shown in Figure 14a. The space group of the crystals was shown to be R3 with very high pseudo-R32 symmetry. The unit cell dimensions according to the equivalent triply centered hexagonal cell are a = b = 137 Å and c = 76 Å. This unit cell is identical to that for the native plant protein (McPherson and Spencer, 1975), and the distribution of intensities in the diffraction pattern indicates the crystals of recombinant and native protein to be isomorphous.

The rhombohedral crystals described here produced different intensities on precession photographs that extended to a resolution of at least 3 Å (Fig. 15). The crystals also showed little decay with x-ray exposure until about 48 h had elapsed. Although recombinant canavalin crystals withstand extended x-ray exposure once mounted inside a capillary tube, bacterial canavalin crystals often degrade within a few hours after maturity if left in the mother liquor.

A three-dimensional x-ray diffraction data set was collected to 3.2-Å resolution from crystals of the recombinant protein using a multiwire area detector system (Xuong-Hamlin, San Diego Multiwire Systems, San Diego, CA). Although a detailed comparison of the structures of the recombinant and natural protein is currently in progress, it is sufficient here to note that the R factor resulting from scaling of corresponding recombinant crystal and native plant protein crystal x-ray intensity data sets was 0.075 for 5323 reflections within 3.2 Å resolution. Thus, it can be concluded that for all practical
The expression of canavalin mediated by basal levels of T7 RNA polymerase appeared toxic to the bacteria host. Canavalin have virtually identical three-dimensional structures.

**DISCUSSION**

The Recombinant Protein

The cDNA for canavalin, the vicilin reserve protein from the jack bean, *C. ensiformis*, was amplified from total plant mRNA by PCR techniques and cloned into *E. coli*. This was facilitated by knowledge of the sequence of the closely related protein from *C. gladiata* (Yamauchi et al., 1988), which permitted the construction of appropriate primers. The cDNA was directly sequenced to generate a consensus at each nucleotide. The cloning and sequencing procedures we employed provided an unequivocal nucleotide and deduced amino acid sequence for recombinant jack bean canavalin.

A number of DNA vectors with strong transcription promoters and ribosome binding sites were investigated for the expression of canavalin in bacteria. These included those with common promoters from λ bacteriophage, λP_C, *E. coli*, and *lac* and the bacteriophage T7 RNA polymerase promoter. We found that the pET expression vector, which contains the T7 RNA polymerase promoter, gave the greatest level and ease of regulation of canavalin expression in *E. coli*.

The expression of canavalin mediated by basal levels of T7 polymerase appeared toxic to the bacteria host. Canavalin itself may be directly responsible for the harmful effects. Alternatively, the high level of transcription of the coding gene may inhibit the ability of the recombinant plasmid to replicate, leading to its eventual loss from the culture. The presence of the pLysE in *E. coli* strain BL21(DE3) gave rise to lysozyme expression. This was essential in the reduction of basal expression, thereby allowing high-density cell growth prior to induction of canavalin.

Recombinant canavalin was found exclusively in the soluble portion of the cell extract. Roughly 3% of the total cellular protein of *E. coli* was expressed as canavalin, as judged by SDS-PAGE, and the recombinant protein appeared principally as monomers. This may be attributed to the relatively low expression level of canavalin in bacteria compared with the high-level expression in plants, where canavalin appears as trimers. Thus, the aggregate state of canavalin in vivo may be dependent upon the level of expression and concentration of the protein.

Recombinant canavalin was readily identified in the bacterial extract, at high levels, by SDS-PAGE and by western blot analysis using polyclonal antibodies conjugated with appropriate marker enzymes. As described above, confirmation was provided as well by the similarity of the recombinant protein to plant canavalin in a number of chemical and physical properties and, ultimately, by the demonstration that the recombinant protein could be crystallized in a form isomorphous with the common crystal form of native plant canavalin.

Using the property of unusual heat stability, characteristic of both the native plant protein and the bacterially expressed canavalin, an efficient purification scheme was devised. This procedure permits the ready purification of recombinant canavalin in 50- to 100-mg amounts, from a few liters of cells, in a matter of days, and to nearly complete homogeneity. The large quantities of the recombinant protein available to us permitted extensive characterization, both biochemical and physical, using a number of techniques.

Comparison with Other Vicilin Proteins

We suspected when we began this work that the sequence of canavalin from *C. ensiformis* was extremely similar to that from its sister species *C. gladiata*, but we were uncertain as to how similar. In addition, to refine the structure of canavalin obtained by x-ray diffraction analysis (Ko et al., 1993), it was essential that we know the amino acid sequence of canavalin precisely.

We showed in previous work (Gibbs et al., 1989) that the vicilin protein from *C. gladiata* contained a prominent internal sequence repeat and that the amino- and carboxyl-terminal halves of the molecule probably had arisen from duplication and joining of a single primordial gene. This genetic, tandem duplicate in turn gave rise to a double-domain, structurally repetitive protein (Ko et al., 1993). We also showed that this internal sequence repetition was true for all of the five vicilin proteins as well. These observations agree satisfactorily with similar sequence comparisons of vicilin proteins made by other investigators (Doyle et al., 1986).

Not only have all vicilin proteins apparently evolved from the same precursor, the tandem duplicate of a single primordial gene, but it seems nearly certain that all vicilin proteins...
The two tertiary structures are extraordinarily similar. The canavalin, are shown optimally aligned. The two proteins are about 60% identical and otherwise are highly homologous. As has already been noted (but warrants emphasis), proteins whose three-dimensional structures are precisely conserved are easily tolerated by the structural requirements of the amino acid sequence, homology is very high. In terms of identical amino acids in equivalent positions, there are 38 common to both domains. These amino acids, listed in Table III, are of particular importance to either the structure, processing, or function of canavalin because they have resisted mutation since the gene duplication event. A comparison of the amino acid sequence of the two individual domains of canavalin with those of five vicilins from other plants (Gibbs et al., 1989) shows that there are, respectively, 42 and 46 totally conserved amino acids among the six proteins. Of the amino acids conserved both between canavalin domains and among five vicilin proteins,

Another pathway that seems equally acceptable is that legumin preceded vicilin and that the primordial vicilin gene split off from legumin, duplicated, and proceeded to evolve. Indeed, there is some evidence from other studies (Borroto and Dure, 1987) that this latter course may be valid.

If one compares the amino- and carboxyl-terminal domains of canavalin according to amino acid sequence, homology is very high. In terms of identical amino acids in equivalent positions, there are 38 common to both domains. These amino acids, listed in Table III, are of particular importance to either the structure, processing, or function of canavalin because they have resisted mutation since the gene duplication event. A comparison of the amino acid sequence of the two individual domains of canavalin with those of five vicilins from other plants (Gibbs et al., 1989) shows that there are, respectively, 42 and 46 totally conserved amino acids among the six proteins. Of the amino acids conserved both between canavalin domains and among five vicilin proteins.

will have the same underlying secondary, tertiary, and quaternary structure as well. The similarity in structure between canavalin, which we present in Ko et al. (1993), and phaseolin (PHA) strongly confirms that relationship.

**Conserved Amino Acid Residues**

Canavalin and phaseolin are currently the only two vicilin proteins whose three-dimensional structures are precisely known. As has already been noted (but warrants emphasis), the two tertiary structures are extraordinarily similar. The two sequences are seen aligned in Figure 16. The identity of amino acids is about 65%, and in spite of numerous short insertions and deletions, the homology overall is very high. We wish here only to point out that a 35% change in sequence is easily tolerated by the structural requirements of the protein.

We would also like to call attention to our earlier observation (Gibbs et al., 1989) that vestiges of the same primordial gene that gave rise to the two domains of canavalin are present as well in the amino-terminal portion of the legumin proteins. We proposed earlier that the legumins may have originated from the concatenation of the primordial vicilin domain gene with that of a second polypeptide, thereby producing a chimeric legumin protein. This is a plausible explanation.
there are 12 in the amino-terminal domain and 18 in the carboxyl-terminal half. Thus, 30 amino acids in canavalin survived from the progenitor gene not only through the internal evolutionary divergence of the two domains making up the protein, but through divergent evolution of the plant species as well.

Sebastiani et al. (1991) noted three highly conserved regions among vicilin proteins with sequence motifs of LQR/KF, KPN/HTX6P, and PHY/FNSR/K that they postulated to produce a common secondary structure that could interact with other macromolecules. These motifs are found in the canavalin molecule at residues 70–73, 94–101, and 296–302 and do contain a number of the highly conserved amino acids. The first region (residues 70–73), located in the amino-terminal domain, is similar to the yeast carboxypeptidase Y vacuolar targeting signal and is present in both storage proteins and lectins (Valls et al., 1990). The second region also resides in the amino-terminal domain, 20 residues from the first, and the last conserved region is within the carboxyl-terminal domain. These conserved amino acid residues reside in the β-strands of the molecule on the surface regions, where they could be important for interactions with cellular components involved in targeting transport vesicles, oligomerization, and presentation of sites for hydrolysis.

The Assembly of Canavalin in the Bacteria

It is easy to overlook the significance of successful expression of canavalin in E. coli, but the appearance of the bacterially synthesized protein has some important implications. All of our evidence suggests that, with few exceptions, the polypeptide synthesized in the bacteria has the correct sequence, that it folds successfully into identically the same secondary and tertiary structure as the canavalin protein synthesized in plant cells, and that the monomeric 47-kD units aggregate correctly to form the characteristic trimer structure. We note that each domain of canavalin possesses a single, free Cys in the native protein, but we see no evidence of disulfide formation in the recombinant product. One must conclude that little posttranslational assistance is required from the plant cellular machinery once synthesis is completed.

Indeed, the physical-chemical results are almost totally consistent in showing the native and recombinant canavalins to be the same. These include cross-reaction with antibodies, identical mol wts of monomer and oligomer, the same pi, the same secondary and tertiary structure by CD methods, and the same tertiary structure by x-ray crystallography. The only exceptions, and their sources are not clear, are that (a) recombinant canavalin shows a slightly altered trypsin digest pattern containing an additional intermediate, and (b) on FPLC the recombinant protein shows some tendency to form a hexamer of Mr, 300,000, whereas the native plant protein apparently does not.

One question that we were interested in examining was whether canavalin, following synthesis in E. coli, was subsequently cleaved in the cell to yield its characteristic fragment pattern or whether enzymes to carry this out were absent in the bacterium. In the germinating seed, cleavage of virtually all canavalin is complete within a day or so of imbition (Smith et al., 1982); thus, the seed must contain or express highly active proteolytic enzymes. SDS-PAGE and western blot analysis of the crude cell lysate suggest that although some cleavage occurs, it is confined to only a small portion of the total recombinant canavalin present and most expressed protein is fully intact.

Because the successful crystallization of a recombinant protein depends on proper synthesis, folding, and aggregation, and all of these might be easily misdirected in a foreign expression system, we were pleased and gratified to obtain crystals of the recombinant protein. The appearance of the rhombohedral crystal form, so consistently characteristic of the native plant protein, not only proved by its isomorphism that the three-dimensional structures of the natural and recombinant canavalins were identical, it now provides us with a powerful tool for assessing structural changes in the protein introduced by chemical or genetic alteration.

The Genetic Engineering of Canavalin

We are interested in canavalin not only for its important natural role as a storage protein in the plant and its equally interesting structural properties, but because it supplies us with an ideal system for directed, purposeful protein engineering. With this protein we can couple the two powerful techniques of recombinant DNA manipulation and x-ray crystallography. We can now genetically alter canavalin, crystallize the recombinant forms, and visualize the structural changes that occur using difference Fourier x-ray diffraction analysis. Because the structure of the naturally occurring protein is known, and because site-directed mutagenesis is rapid and available, this entire process can be carried out in a very expeditious manner.

If we are able to reintroduce an intelligently engineered gene for canavalin into the plant, then the possibility exists that we could substantially enhance the nutritional properties of the seed storage protein. The importance of such a genetic engineering effort is increased by the demonstration, by several techniques, that canavalin is genetically, biochemically, and structurally very representative of all of the vicilin class proteins. Thus, although our work focuses on canavalin, our results bear upon similar efforts using more economically important crops such as soybeans and peas. C. ensiformis provides a model system for the alteration and nutritional improvement of all of the leguminous plants. Because leguminous seeds are a major source of the world’s dietary protein, the consequences for world health would be appreciable.

ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of Alex Malkin, Frank Huang, and James Kwong.

Received August 7, 1992; accepted November 23, 1992.
Copyright Clearance Center: 0032-0889/93/101/0713/16.
The EMBL/GenBank accession number for the sequence reported in this article is 59467.

LITERATURE CITED
