Plant Physiol. (1993) 101: 729–744

The Three-Dimensional Structure of Canavalin from Jack Bean (*Canavalia ensiformis*)[1]

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

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

The three-dimensional structure of the vicilin storage protein canavalin, from *Canavalia ensiformis*, has been determined in a hexagonal crystal by x-ray diffraction methods. The model has been refined at 2.6 Å resolution to an *R* factor of 0.197 with acceptable geometry. Because of proteolysis, 58 of 419 amino acids of the canavalin polypeptide are not visible in the electron density map. The canavalin subunit is composed of two extremely similar structural domains that reflect the tandem duplication observed in the cDNA and in the amino acid sequence. Each domain consists of two elements, a compact, eight-stranded β-barrel having the "Swiss roll" topology and an extended loop containing several short α-helices. The root mean square deviation between 84 pairs of corresponding C_α atoms making up the strands of the two β-barrels in a subunit is 0.78 Å, and for 112 pairs of structurally equivalent C_α atoms of the two domains the deviation is 1.37 Å. The interface between domains arises from the apposition of broad hydrophobic surfaces formed by side chains originating from one side of the β-barrels, supplemented by at least four salt bridges. The interfaces between subunits in the trimer are supplied by the extended loop elements. These interfaces are also composed primarily of hydrophobic residues supplemented by six salt bridges. The canavalin subunits have dimensions about 40 × 40 × 86 Å, and the oligomer is a disk-shaped molecule about 88 Å in diameter with a thickness of about 40 Å. The distribution of domains lends a high degree of pseudo-32-point group symmetry to the molecule. There is a large channel of 18 Å diameter, lined predominantly by hydrophilic and charged amino acids, running through the molecule along the 3-fold axis. The majority of residues conserved between domains and among vicilins occur at the interface between subunits but appear otherwise arbitrarily distributed within the subunit, although predominantly on its exterior.

There are two major types of storage proteins in legume seeds, legumin and vicilin (Derbyshire et al., 1975). Legumins, or 11S globulins, are hexamers with molecular masses of 300 to 400 kD. Vicilins, or 7S globulins, are trimers with molecular masses of 150 to 200 kD. Another class of smaller (2S) globulins are single polypeptides of molecular mass 30 to 35 kD. The three-dimensional structures of a vicilin, phaseolin (Lawrence et al., 1990), and a 2S globulin narbonin (Hennig et al., 1992) have been determined by x-ray crystallography. The monomer of phaseolin consists of two similar domains, each having a β-barrel and several α-helices. Narbonin assumes a structure of the frequently reported α/β barrel (TIM barrel), does not have any known enzyme activity, and has been assumed to be a reserve protein as well. These reserve proteins are synthesized during seed development and may provide a nitrogen source during seed germination.

Canavalin, Con A, Con B, and urease are four major soluble globulins in the seed extract of jack bean (*Canavalia ensiformis*) (Sumner, 1919). All of them have been crystallized (Sumner, 1919, 1926; Sumner and Howell, 1936) and analyzed by x-ray diffraction. Con A is a tetrameric lectin that has long been a subject of research and whose structure has been determined (Reeke et al., 1975). Con B is a 2S globulin and, like narbonin, consists of a TIM β-barrel structure; however, evidence suggests that it may have enzymic activity (Morrison et al., 1984; McPherson, unpublished data). The crystal structure of urease has also been investigated (Jabri et al., unpublished data). In this paper, we present the three-dimensional structure of canavalin, the vicilin-class storage protein of jack bean, which is, in many respects, very similar to that of phaseolin.

Crystals of the proteolytic products of canavalin were first obtained by Sumner and Howell (1936) and first characterized by McPherson and Rich (1973). Other crystal forms of the protein were subsequently grown and their preliminary analyses carried out by McPherson and Spencer (1975). Crystals of canavalin have been obtained only from native protein that has been cleaved, almost in half, by proteases. Crystals of native canavalin have not yet been produced. In this paper, reference to canavalin implies the proteolytically modified molecule consisting of subunits comprising polypeptide chains derived from the amino- and carboxyl-terminal halves of the native protein.

The first, and most common, crystal form of canavalin was shown to be of space group R3 with $a = b = c = 83.0$ Å and $\gamma = 111.1$ Å, with equivalent triply centered hexagonal unit cell dimensions of $a = b = 136.8$ Å and $c = 75.7$ Å. There was one canavalin subunit of 47,860 D as the asymmetric unit of those crystals, which demonstrated an extraordinary degree of pseudo-R32 symmetry (McPherson and Rich, 1973; McPherson and Spencer, 1975). This could be explained only by assuming that the amino- and carboxyl-terminal halves

[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.

Abbreviations: *B*, temperature factor in Å^2; Con A, concanavalin A; Con B, concanavalin B; DPBS, Dulbecco's phosphate buffered saline; *I*, intensity; MIR, multiple isomorphous replacement; OCMP, *o*-chloromercuriphenol; *R*, crystallographic residual Σ_{hkl} ($[F_o] - k[F_c]$) / Σ_{hkl} (F_o); R_{sym}, Σ_{hkl} ($[F_{hkl}] - [F_{-h-k-l}]$) / Σ_{hkl} ($[F_{hkl}] + [F_{-h-k-l}]$); rms, root mean square; σ, estimated standard error; TIM, triose phosphate isomerase.

of the subunit were extremely similar in structure and that they were related by a quasi-2-fold axis of symmetry. This proposal has now been confirmed by the nucleotide and implied amino acid sequence of the cDNA coding for the protein (Doyle et al., 1986; Gibbs et al., 1989; Ng et al., 1993) and by the crystallographic analysis we report here.

From an earlier attempt to determine the structure of canavalin using the R3 crystal form, a model for the folding of a polyalanine chain was proposed (McPherson, 1980). Although some features of this earlier model were correct, there was no amino acid sequence available. That model was topologically incorrect, and it could not subsequently be refined. Ultimately, we abandoned an MIR approach to the R3 crystals.

Among the four known crystal forms of canavalin, the hexagonal crystal of space group $P6_3$, having cell dimensions of $a = b = 126.4$ Å and $c = 51.6$ Å, was ultimately used to effect an MIR phased structure solution of the protein. This crystal form yielded superior diffraction data for a number of technical reasons and permitted improved formation of isomorphous derivatives. Coupled with substantially improved data-collection techniques, this allowed us to construct a model that incorporates the now known amino acid sequence.

We report here our determination by x-ray crystallography of the three-dimensional structure of canavalin at 2.6 Å resolution and its refinement by crystallographic least-squares and difference Fourier methods combined with simulated annealing. We have also shown, by molecular replacement techniques, that this molecular structure is stringently maintained in other crystal forms as well. The other three crystal forms are of space groups R3, $C222_1$, and $P2_13$. Their structures have been determined and refined with little alteration of the $P6_3$ model described in this paper (T.P. Ko, J.D. Ng, A. McPherson, unpublished data).

Canavalin is of interest to us from the standpoint of protein architecture and function, and particularly for its internal structural redundancy. The protein is also relevant in a practical and economic sense as well. Canavalin is an ideal representative of the highly homologous class of storage proteins, the vicilins, of the legume seeds (Derbyshire et al., 1976; Bewley and Black, 1978; Doyle et al., 1986; Gibbs et al., 1989). These constitute the third largest source of dietary protein on earth, a source of particular importance in developing countries that lack extensive supplies of animal protein (Creamer et al., 1988).

Because the foundation is now established for the application of molecular genetics (Ng et al., 1993), the value of an accurate three-dimensional structure is increased. By coupling the analysis of a precise image of the molecule with a genetic approach to its alteration and synthesis, a systematic and intelligent approach to the nutritional enhancement of the protein through site-directed mutagenesis becomes possible (Oxender and Fox, 1987).

Crystalline canavalin has also been a focus of attention for its value as a defined, reproducible system for the study of the protein crystallization process. In particular, it has been the subject of crystallization experiments during 10 missions of the U.S. Space Shuttle over the past 6 years. These were designed to identify and evaluate the effects of microgravity on protein crystal growth (DeLucas et al., 1986; McPherson et al., 1991; Day and McPherson, 1992). In the course of those experiments, canavalin crystals growing in space have provided a number of useful and intriguing observations.

MATERIALS AND METHODS

Native canavalin was purified from defatted jack bean (*Canavalia ensiformis*) meal (Sigma) and treated with trypsin (Gibco) as described previously (Sumner, 1919; Sumner and Howell, 1936; Smith et al., 1982). The final product was a mass of rhombohedral crystals of canavalin. The protein was recrystallized three times by successive dissolutions with trace amounts of NH_4OH in minimal amounts of H_2O followed by dialysis against DPBS (Gibco) at room temperature. The four-times recrystallized canavalin was dissolved in distilled water with the addition of trace amounts of NH_4OH to a protein concentration of 30 to 40 mg/mL. Crystallization for x-ray diffraction analysis was achieved using the vapor diffusion method on nine-well glass depression plates sealed in plastic sandwich boxes (McPherson, 1982, 1990). The reservoirs were 2.0% NaCl buffered with 0.05 M phosphate at pH 6.8. Crystallization was carried out at between 4 and 8°C, an essential condition for growth of the hexagonal crystal form.

Large hexagonal prisms were obtained after 3 to 6 d, and these were generally stable for months, provided the ambient conditions were not significantly altered. Crystals were mounted in quartz capillaries by conventional means and data collection was carried out at 17°C.

Hexagonal prisms used for collection of x-ray diffraction data were at least 1 mm in length and about 0.3 mm across the prismatic face. The crystals are of space group $P6_3$ and have cell dimensions of $a = b = 126.35$ Å and $c = 51.64$ Å. The limit of resolution is about 2.6 Å but the diffraction pattern, as with almost all of the crystal forms of canavalin, begins to decline rapidly in average intensity beyond 3.0 Å Bragg spacings.

For formation of isomorphous heavy atom derivatives, 5 μL of 5-mM solutions of K_2HgI_4, $Hg(C_2H_3O_2)_2$, K_2PtCl_4, or $K_2UO_2F_5$ in DPBS were added to canavalin crystals in the mother liquor from which they were grown. One to 2 weeks were generally allowed before data collection was undertaken. In the case of OCMP, 5 μL of an OCMP-saturated DPBS solution was added, and for $PtBr_2(NH_3)_2$, which is rather insoluble, a small grain of the platinum compound was added directly to the 20-μL droplet of mother liquor. Two weeks, and often longer, were allowed to elapse before data were collected from these latter crystals.

Data for native and heavy-atom derivatives were collected using an automated Enraf-Nonius CAD 4 diffractometer with an omega-two theta scan. A 722-mm helium path was inserted between crystal and counter, and reflections were collected as Friedel pairs. Correction for absorption and geometrical factors as well as merging of the data were as described previously (Brayer and McPherson, 1983). In general, observations were collected from several different native crystals but only one crystal was used to measure the Friedel pairs for each heavy-atom derivative. The maximum resolution for each derivative was 3.2 Å and R_{sym} values were

consistently between 0.048 and 0.062 for all heavy-atom derivative data sets. Once an initial model of the structure had been obtained, diffraction data of high redundancy, having a minimum of 10 observations per reflection with an R_{sym} of 0.045, were collected for a single, large, native crystal using a two-panel Xuong-Hamlin (Xuong et al., 1985) area detector system (San Diego Multiwire Systems, San Diego, CA). The x-ray source for those measurements was a Rigaku RU-200 rotating anode generator fitted with a Supper graphite crystal monochromator and operated at 175 kV and 45 mA. A 535-mm helium path was installed between crystal and detectors, and the data were processed using the programs of Howard and Nielson supplied by San Diego Multiwire Systems (Hamlin et al., 1981).

Potential heavy-atom derivatives were screened by collecting 5.0 Å resolution data by diffractometry, processing the data, scaling the derivative data to native, and computing a difference Patterson map. If the map could be successfully interpreted, then data collection on the same crystal was simply continued to the maximum resolution employed in the analysis.

Scaling of derivative to native data employed a Fourier-Bessel procedure (Terwilliger et al., 1982), and all Patterson and Fourier calculations were carried out using the program FFT (Ten Eyck et al., 1976). Heavy-atom parameter refinement and MIR phase calculation were carried out using the program HEAVY written by Terwilliger and Eisenberg (1984). Initially, heavy-atom derivatives were refined using the origin-removed Patterson method with final cycles employing traditional Blow-Crick (Blow and Crick, 1959) approaches.

The only complication in establishing the heavy-atom positions was the correlation of the z coordinates of the heavy-atom sites of different derivatives. This was done initially using a difference Patterson method proposed by Rossmann (1960), which has as coefficients the squared differences between pairs of derivatives. Later difference Fourier maps of alternate derivatives confirmed the assignments. Relative z coordinates were ultimately refined using all derivatives combined through the Blow-Crick (Blow and Crick, 1959) procedure.

The heavy-atom compounds utilized in the MIR phasing (Watenpaugh, 1985) of the canavalin structure and the refined parameters describing their individual sites of substitution are shown in Table I. The refinement statistics for each compound are shown in Table II. Several of the derivatives have sites in common with one another, but there were enough unique sites to yield adequate phase information.

The overall figure of merit of the MIR phases for 20- to 3.2-Å resolution data was 0.66, and that for the centric reflections alone was 0.80. The distribution of the figure of merit is presented in Figures 1 and 2. Electron-density maps were calculated at 0.8-Å intervals along all axes and produced in "minimap" format on plexiglass sheets for inspection and interpretation. Density modification of the MIR maps was carried out using the programs and procedures of Wang (1985). The modified phases were always tethered to the original MIR phases and a solvent volume of 40% was assumed. At later stages, phases were calculated from the model structure and combined with both MIR phases and density-modified MIR phases as well. Model building to electron-density maps and most other graphics analysis was carried out on an Evans and Sutherland PS 390 system using the program FRODO written by T.A. Jones (1982, 1985). Illustrations of the structure were produced with FRODO and RIBBONS, written by Carson and Bugg (1986).

The canavalin structure was refined by alternating various types of crystallographic least-squares procedures with inspection of difference Fourier maps, model rebuilding, and simulated annealing (Brunger et al., 1987; Brunger, 1988, 1991). Initially, the structure was refined using the constrained-restrained least-squares approach of CORELS (Sussman et al., 1977; Sussman, 1985) followed by restrained

Table I. *Heavy-atom sites of the $P6_3$ canavalin crystal*

Compound	Site	Occupancy	x	y	z	B	Nearest Amino Acid
K_2HgI_4	1	1.75	0.8414	0.2709	0.6013	38	His^{297}
	2	0.49	0.4267	0.3836	0.3343	38	Met^{222}
	3	0.84	0.3985	0.0260	0.2064	35	—[a]
$PtBr_2(NH_3)_2$	1	1.40	0.9732	0.3745	0.1955	40	Met^{222}
	2	0.78	0.9845	0.3690	0.2235	35	—
	3	0.25	0.3435	0.4861	0.0173	28	—
K_2PtCl_4	1	1.38	0.9762	0.3745	0.1984	56	Met^{222}
	2	0.54	0.3852	0.0065	0.2441	29	Met^{331}
	3	0.44	0.4835	0.3333	0.3955	13	—
	4	0.16	0.3634	0.0939	0.1083	08	—
OCMP	1	0.92	0.3670	0.8717	0.0000	41	His^{102}, Phe^{175}
	2	0.79	0.8222	0.2888	0.6510	45	Cys^{285}, Tyr^{266}
$Hg(C_2H_3O_2)_2$	1	1.02	0.8766	0.5103	0.5060	34	His^{102}
	2	1.02	0.8367	0.2672	0.5956	37	His^{297}
$K_2UO_2F_5$	1	0.36	0.2894	0.7973	0.8823	17	—
	2	0.11	0.4633	0.3603	0.2537	08	—
	3	0.08	0.0806	0.4777	0.4616	07	—
	4	0.15	0.4972	0.2784	0.4671	20	—

[a] Residue involved in binding could not be determined with certainty.

Table II. *Refinement statistics for heavy-atom derivatives of the $P6_3$ canavalin crystal*

Compound	Statistic[a]	Resolution Range (Å)								
		Overall	9.36	6.64	5.41	4.68	4.18	3.81	3.53	3.30
K_2HgI_4	Centric: #Refs	458	63	54	56	62	61	58	50	54
	R	0.57	0.47	0.43	0.51	0.63	0.60	0.78	0.88	0.80
	rms(FH)/rms(E)	1.17	1.17	1.40	1.44	1.17	1.01	1.02	1.00	0.92
	Acentric: #Refs	5411	365	501	648	751	804	317	857	668
	rms(FH)/rms(E)	1.54	1.57	1.86	1.83	1.51	1.35	1.32	1.50	1.30
	R(Deriv to Nat)	0.19								
$PtBr_2(NH_3)_2$	Centric: #Refs	492	67	59	64	67	65	69	48	53
	R	0.60	0.52	0.62	0.48	0.63	0.61	0.61	0.79	0.79
	rms(FH)/rms(E)	1.03	1.45	1.54	1.21	0.78	0.62	0.63	0.68	0.69
	Acentric: #Refs	5621	369	511	674	799	856	928	857	627
	rms(FH)/rms(E)	1.32	1.99	2.04	1.69	1.24	0.94	0.99	0.89	0.89
	R(Deriv to Nat)	0.14								
K_2PtCl_4	Centric: #Refs	407	67	60	64	60	58	63	35	
	R	0.66	0.61	0.81	0.64	0.66	0.60	0.66	0.69	
	rms(FH)/rms(E)	0.81	0.89	0.82	0.82	0.84	0.60	0.73	0.70	
	Acentric: #Refs	4189	369	497	661	727	725	744	466	
	rms(FH)/rms(E)	1.06	1.16	1.33	1.01	1.14	0.87	0.92	0.78	
	R(Deriv to Nat)	0.18								
OCMP	Centric: #Refs	423	65	58	67	61	62	72	38	
	R	0.57	0.36	0.47	0.50	0.72	0.73	0.81	0.91	
	rms(FH)/rms(E)	1.22	1.96	1.48	1.53	0.93	0.76	1.00	0.93	
	Acentric: #Refs	4666	364	499	677	753	781	948	644	
	rms(FH)/rms(E)	1.50	2.12	1.97	1.95	1.29	1.03	1.31	1.39	
	R(Deriv to Nat)	0.14								
$Hg(C_2H_3O_2)_2$	Centric: #Refs	477	67	60	59	59	60	58	52	62
	R	0.53	0.33	0.39	0.49	0.50	0.58	0.80	0.89	0.98
	rms(FH)/rms(E)	1.26	2.05	1.99	1.12	1.18	0.84	0.88	1.13	1.16
	Acentric: #Refs	5513	367	503	672	765	767	700	903	836
	rms(FH)/rms(E)	1.59	2.58	2.32	1.63	1.60	1.24	1.09	1.51	1.63
	R(Deriv to Nat)	0.12								
$K_2UO_2F_5$	Centric: #Refs	382	65	60	59	58	59	55	26	
	R	0.67	0.59	0.83	0.65	0.73	0.63	0.68	0.70	
	rms(FH)/rms(E)	0.57	0.61	0.69	0.54	0.64	0.43	0.53	0.41	
	Acentric: #Refs	4292	367	499	656	740	726	817	487	
	rms(FH)/rms(E)	0.69	0.95	0.88	0.77	0.68	0.63	0.60	0.54	
	R(Deriv to Nat)	0.11								

[a]All quantities defined in Terwilliger and Eisenberg (1983).

least-squares refinement with TNT (Ten Eyck et al., 1976; Tronrud et al., 1987) coupled to difference Fourier reviews. Later, when we were confident of the model, simulated annealing with XPLOR (Brunger et al., 1987; Brunger, 1988, 1991) was employed, and this was followed by additional cycles of refinement with TNT to bring the geometry of the structure nearer ideal values. We also used Ramachandran plots (Ramachandran and Sassiekharan, 1968) of the ϕ,ψ angle distribution as a guide to regions of the polypeptide requiring attention. These were appropriately refitted and refined further with TNT. The program XPLOR was executed on an SGI 320 computer, and all other calculations were made on a VAX 8820.

All possibly observable native data to 2.6-Å resolution were recorded, but only data with $I \geq 3\sigma$ were employed in the refinement. After elimination of marginal data, a set was obtained that was 90% complete to 3.0 Å but diminished to 60% complete at 2.6 Å resolution. The R_{sym} for all native data to 2.6 Å was 0.044. Although low resolution data ($\infty \rightarrow$ 8 Å) were included in the initial stages of refinement, they were eliminated subsequently as has commonly been done in refinements of most other protein structures (Tronrud et al., 1987).

RESULTS

In the initial MIR electron-density map, the envelopes of the individual subunits of the trimer were apparent. Within the monomer, two pseudo-dyad related domains were, with a few exceptions, also well defined. Molecular boundaries were further enhanced by the solvent-leveling procedure (Wang, 1985), but at the expense of side chain density for many surface residues. A representative 8-Å thick section through the MIR map at 3.2-Å resolution is shown in Figure 3. This section is particularly interesting because it is along the 3-fold axis of the crystal and contains the pseudo-dyad axis relating two amino and carboxyl terminal domains. On the top and bottom of this axis can be seen two similar, but

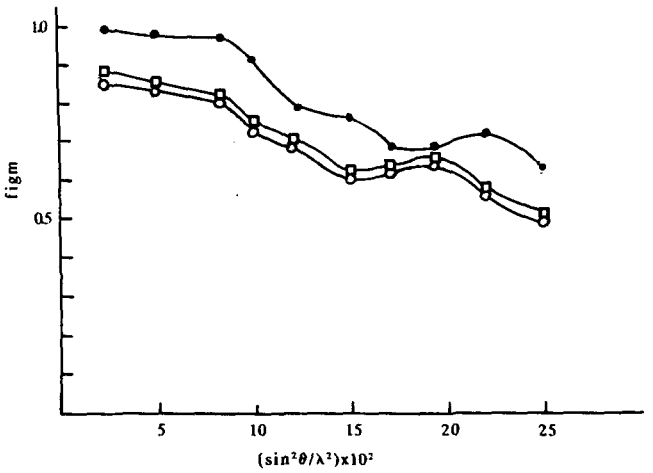

Figure 1. The distribution of the figure of merit (figm) as a function of resolution from 15 to 3.2 Å resolution (□). Also shown is the same distribution for centric reflections (●) and acentric reflections (○).

crystallographically unique, horseshoe-shaped densities. These represent sections through amino and carboxyl Swiss roll β-barrels (Richardson, 1981; Branden and Tooze, 1991), and the ellipsoidal voids of the horseshoes are the lumina of the 2-fold related barrels.

Strands of the β-barrels were relatively straightforward to trace once the overall topology of the molecule emerged, and most of the extended chain alternating with helical segments could also be delineated. Some connecting loops and portions of the extended chain, however, remained ambiguous. These were later defined through inspection of $2F_o$-F_c and F_o-F_c difference Fourier maps based on phases calculated from partial models that were combined with MIR phases. From this series of gradually improved maps a complete model of the canavalin subunit was constructed, and this provided the starting point for refinement.

Portions of the N-terminal segments of both the amino- and carboxyl-terminal domains were clearly visible in the electron-density maps. In both cases, the terminal strands assumed an extended conformation and appeared to be associated with a flanking strand of the β-barrel of an adjacent domain. The distances from the C terminus of the first domain to the two N termini of adjacent domains were 35 and 68 Å. The 35-Å distance could reasonably be accommodated by a connecting polypeptide of about 15 residues; the longer distance could not. Therefore, in our model, the two domains of a subunit were chosen as they appear in the schematic diagram of Figure 4. This assignment is the same as that made by Lawrence et al. (1990) for phaseolin based on other considerations that are generally applicable to canavalin as well.

The letter codes used in Figures 4 and 5 to specify the strands and helices are according to the convention established for viral coat proteins (Rossmann et al., 1983) and phaseolin (Lawrence et al., 1990). In Figure 5, the amino acid sequence of the homologous regions of the two domains of canavalin have been placed one above the other, and the secondary structural elements are noted by letters. In three-dimensional space, the strands and helices appear as illustrated in Figure 4.

The molecule comprising the asymmetric unit of the crystals, it should again be emphasized, was a proteolytically modified form of the naturally occurring seed storage protein and it had experienced extended exposure to trypsin. In interpreting the electron-density maps, therefore, we could not be certain precisely where the primary cleavage sites in the polypeptide occurred. Furthermore, we did not know if there might be some excision of amino acids at either the natural termini or at secondary termini created by cleavages.

The electron-density map, which could contain some disordered regions at the termini that obscure polypeptide present in the crystal, is consistent with a protein molecule lacking 58 of the 419 amino acid residues constituting the native polypeptide. Of these, 20 are at the amino terminus. These are probably lost by a trypsin cleavage somewhere near Arg^{43} that releases a peptide too small to be seen on SDS-PAGE (Smith et al., 1982), followed by some limited digestion. We also failed to identify in the electron-density map residues 228 through 244 and residues 427 through 445 at the carboxyl terminus. Again, the loss of the carboxyl-terminal peptide is likely due to an additional cleavage releasing a short polypeptide. Residues 228 through 244, however, encompass the primary cleavage site that we believe is probably at Arg^{239}–Lys^{240}. This portion of the protein, which connects the two domains of canavalin, was also not visible in the structure of phaseolin even though that protein was intact in its crystals (Lawrence et al., 1990). We suspect that some of the amino acid residues from 228 to 244 may be present in the canavalin crystals, but that cleavage between the two domains converted this inherently disordered polypeptide into two free ends.

There is a second prominent cleavage that occurs in canavalin upon prolonged exposure to trypsin and splits one of the two domains roughly in half (Smith et al., 1982). This

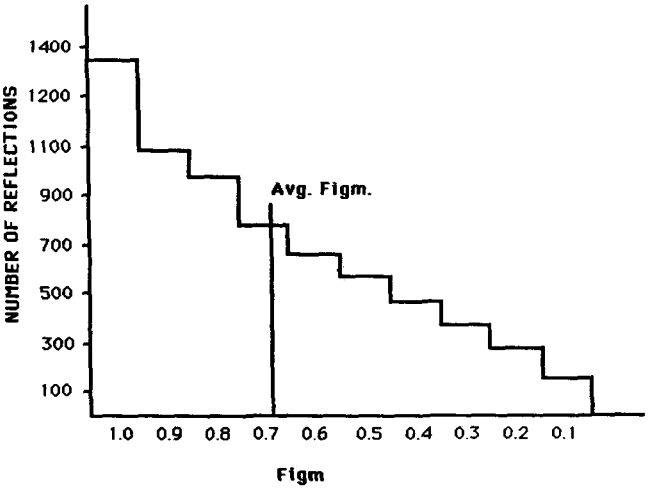

Figure 2. Histogram of the figure of merit (Figm) for all reflections used in computing the MIR-phased electron-density map of canavalin.

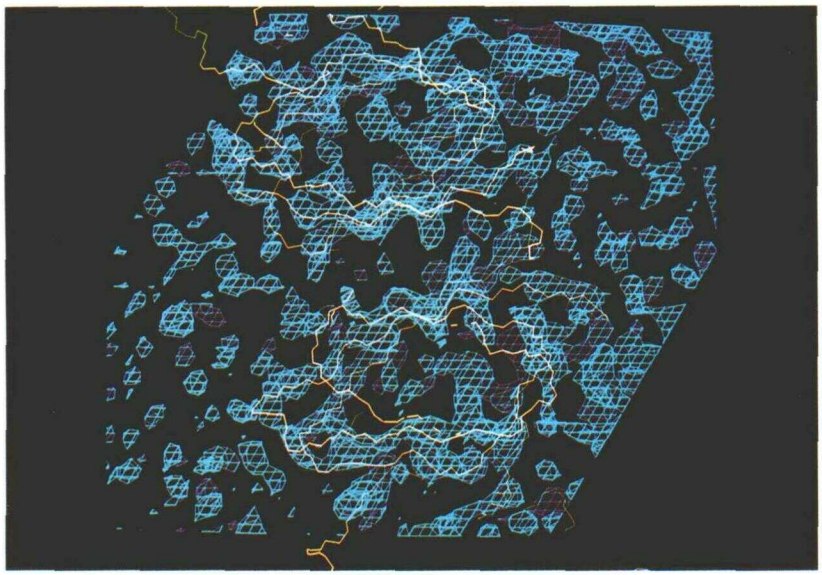

Figure 3. An 8-Å thick section through the MIR-phased electron-density map of canavalin viewed down the 3-fold axis of symmetry. Of distinction in this figure are the two horseshoe-shaped densities on either side of the pseudo-dyad. These correspond to cross-sections of the crystallographically nonequivalent amino- and carboxyl-terminal β-barrels of the two domains that make up the canavalin subunit.

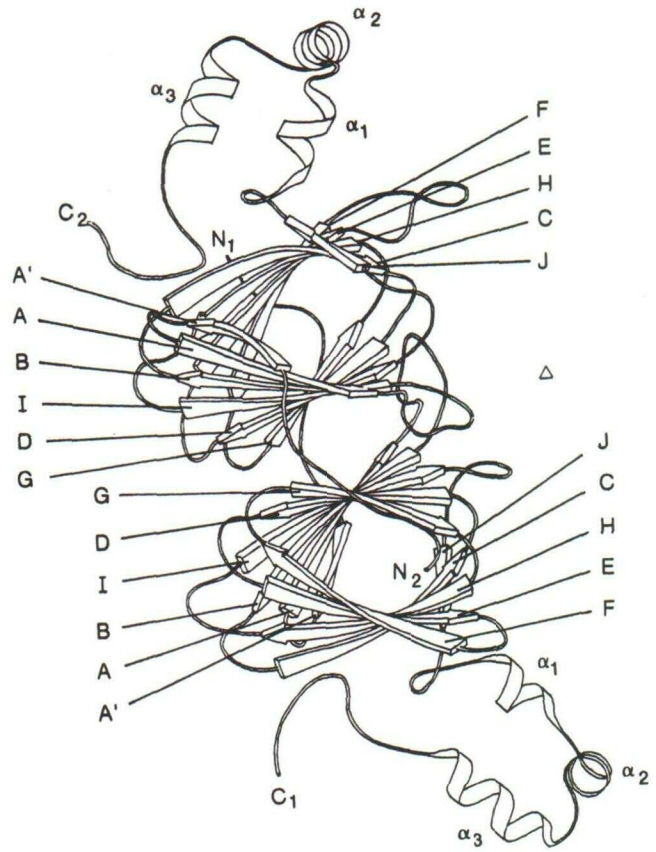

Figure 4. The secondary structural elements of the canavalin subunit are indicated by A′–G and F–J for corresponding strands of the β-barrel domain cores; α_1 through α_3 denote equivalent α-helices. The sequences corresponding to each element are indicated in Figure 5.

```
                                                                                  βA'
Met Ala Phe Ser Ala Arg Phe Pro Leu Trp Leu Leu Leu Gly Val Val Leu Leu Ala Ser  20
Val Ser Ala Ser Phe Ala His Ser Gly His Ser Gly Gly Glu Ala Glu Asp Glu Ser Glu  40
                                                                                  βA'
Glu Ser Arg Ala Gln Asn Asn Pro Tyr Leu Phe Arg Ser Asn Lys Phe Leu Thr Leu Phe  60
Thr Leu Ser Ser Gln Asp Lys Pro Phe Asn Leu Arg Ser     Arg     Asp Pro Ile Tyr 258
                                  βA
Lys Asn Gln His Gly Ser Leu Arg Leu Leu Gln Arg Phe Asn Glu Asp Thr Glu Lys Leu  80
Ser Asn Asn Tyr Gly Lys Leu Tyr Glu Ile Thr Pro Glu Lys Asn     Ser Gln     Leu 276
                                  βB
Glu Asn Leu Arg Asp Tyr Arg Val Leu Glu Tyr Cys Ser Lys Pro Asn Thr Leu Leu Leu 100
Arg Asp Leu     Asp Ile Leu Leu Asn Cys Leu Gln Met Asn Glu Gly Ala Leu Phe Val 295
βC                                βD                                βE
Pro His His Ser Asp Ser Asp Leu Leu Val Leu Val Leu Glu Gly Gln Ala Ile Leu Val 120
Pro His Tyr Asn Ser     Thr Val Ile Leu Val Ala Asn Glu Gly Arg Ala Glu Val Glu 316
                Arg Ala
                                  βF                                βG
Leu Val Asn Pro Asp Gly Arg Asp Thr Tyr Lys Leu Asp Gln Gly Asp Ala Ile Lys Ile 140
Leu Val Gly Leu     Arg Arg Tyr Ala Ala Thr Leu Ser Glu Gly Asp Ile Ile Val Ile 348
  Glu Gln Gln Gln Gln Gln Gln Gly Leu Glu Ser Met Gln Leu
                                  βH
Gln Ala Gly Thr Pro Phe Tyr Leu Ile Asn Pro Asp Asn Asn Gln Asn Leu Arg Ile Leu 160
Pro Ser Ser Phe Pro Val Ala Leu Lys Ala Ala     Ser     Asp Leu Asn Met Val 365
βI                                βJ
Lys Phe Ala Ile Thr Phe Arg Arg Pro Gly Thr Val Glu Asp Phe Phe Leu Ser Ser Thr 180
Gly Ile Gly Val Asn     Ala     Glu     Asn Asn Glu Arg Asn Phe Leu Ala Gly His 382
                α1                                α2
Lys Arg Leu Pro Ser Tyr Leu Ser Ala Phe Ser Lys Asn Phe Leu Glu Ala Ser Tyr Asp 200
Lys Glu     Asn Val Ile Arg Gln Ile Pro Arg Gln Val Ser Asp Leu Thr Phe Pro Gly 401
                α3
Ser Pro Tyr Asp Glu Ile Glu Gln Thr Leu Leu Gln Glu Glu Gln Glu Gly Val Ile Val 220
Ser     Gly Glu Glu Val Glu Glu     Leu Leu Glu Asn Gln Lys Glu Ser Tyr Phe Val 419
Lys Met Pro Lys Asp Gln Ile Gln Glu Ile Ser Lys His Ala Gln Ser Ser Ser Arg Lys 240
Asp Gly Gln Pro Arg His Ile Asp Ala Gly Gly Lys
                    Ala Arg Arg Ala His Leu Pro Asn Leu Phe Arg Thr Phe Tyr 445
```

Figure 5. Alignment of the sequences of the amino and carboxyl domains of canavalin with equivalent amino acids on the top and bottom of each line, respectively. The secondary structure elements to which subsequences correspond are indicated by the bars above. The various secondary elements are labeled in accordance with the diagram in Figure 4.

cleavage site has not been identified chemically, but we believe it can be defined with some confidence by inspection of the electron-density map. This is not entirely straightforward because the molecules making up the crystals used in the analysis contained both intact and cleaved domains. Thus, we would not expect to see a clean break in the polypeptide chain density, but only a weak, perhaps somewhat disordered segment at some susceptible point near the center of either the amino- or carboxyl-terminal domains. Such a segment of weak density consistently appeared in $2F_o$-F_c difference Fourier maps throughout all stages of refinement. This corresponded to a 14-residue loop between strands E and F in the carboxyl-terminal domain. This large loop is greatly abbreviated in the amino-terminal domain. The large loop in the carboxyl domain contains two consecutive Args, 334 and 335, which satisfy the specificity of trypsin, and these two residues appear to be quite accessible. The position of this loop in the sequence is consistent with the lengths of the two fragments produced due to the secondary cleavage by trypsin.

Polypeptides comprising the asymmetric unit were refined, as described above, by simulated annealing and restrained least-square procedures to a final residual of $R = 0.197$ and a correlation coefficient of 0.900 at 2.6-Å resolution using data from the $P6_3$ crystals having an I/σ ratio of 3.0. No water molecules were included in the model. Table III presents the refinement results and the geometry of the structure derived directly from TNT. A Ramachandran plot of the ϕ,ψ angles of the polypeptide chain is shown in Figure 6. The distribution of angles is confined to the allowed regions of the diagram (Morris et al., 1992), although some angles are still marginal. These can be ascribed principally to the free termini, which in this molecule number six and are accompanied by at least some disorder. The average overall error in the model, as estimated from the Luzatti plot (Luzatti, 1952) in Figure 7, is about 0.25 Å.

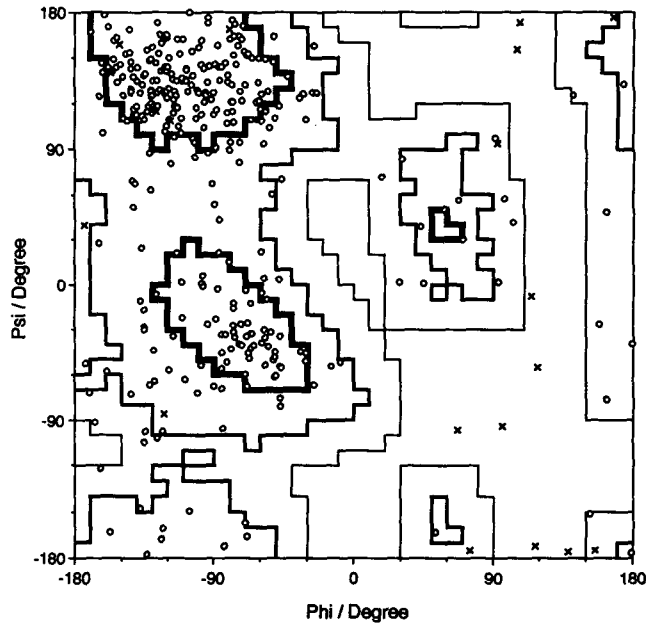

Figure 6. Ramachandran plot of the ϕ,ψ dihedral angles of the canavalin polypeptide chain. The solid lines demarcate the permitted regions. Gly residues are marked by ×, all others by O. The plot is drawn according to Morris et al. (1992), and the heavy lines denote the areas commonly assumed acceptable.

Figure 8, A and B, represent the canavalin subunit with the amino and carboxyl terminal domains shown in yellow and blue, respectively. In Figure 8A, the molecule is viewed along the 3-fold direction, whereas in Figure 8B it is seen perpendicular to the triad and along the pseudo-dyad axis relating two domains. The subunit is about 35 to 40 Å along two directions, but extends nearly 86 Å in the third, thus

Table III. *TNT refinement of the hexagonal $P6_3$ canavalin crystal structure*

Unit cell: $a = b = 126.35$ Å, $c = 51.64$ Å. $\alpha = \beta = 90°$, $\gamma = 120°$. Resolution: 8.0–2.6 Å. $I/\sigma \geq 3.0$. Number of reflections = 10,117. Native $K = 1.578$. Overall $B = 0.0$. $K_{solvent} = 1.0$. $B_{solvent} = 308.8$. Overall $R = 0.197$. Correlation coefficient = 0.9004.

Resolution Breakdown

d_{min}/Å[a]	5.12	4.26	3.79	3.47	3.24	3.06	2.91	2.79	2.69	2.60
No. of F_{obs}[b]	1314	1310	1290	1219	1167	1016	879	769	621	532
% complete	88.9	88.6	87.3	82.5	78.9	68.7	59.5	52.0	42.0	36.0
R_{shell}	0.212	0.154	0.180	0.190	0.198	0.217	0.234	0.234	0.251	0.249
R_{sphere}	0.212	0.181	0.181	0.182	0.185	0.188	0.191	0.193	0.196	0.197

Thermal Parameter Distribution[c]

B_{max}/Å^2	10.0	20.0	30.0	40.0	50.0	60.0	>60.0
No. of atoms	523	749	773	485	244	77	28

Stereochemical Deviation

Category	Bond Length	Bond Angle	Torsional Angle	Trigonal Atom	Planar Group	Bad Contact	Chiral Center
No. of restraints	2978	4012	1793	101	419	300	377
rms deviation	0.020Å	2.909°	24.923°	0.015Å	0.019Å	0.107Å	0

[a] d_{min}, Minimal Bragg spacing. [b] F_{obs}, Observed amplitude. All other quantities defined in Tronrud et al. (1987). [c] Total number of atoms = 2924; average $B_{iso} = 23.40$Å^2.

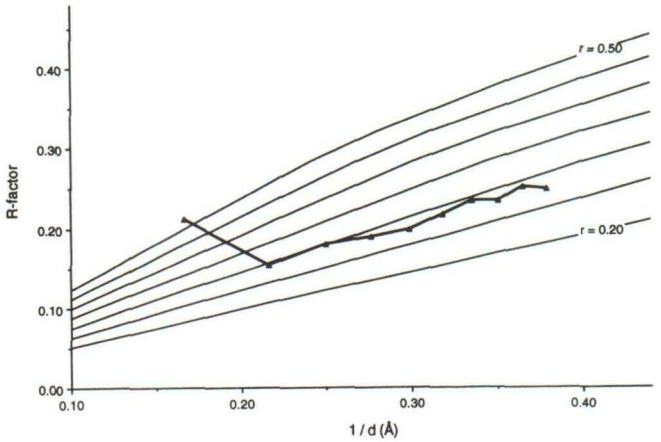

Figure 7. A Luzatti plot showing the R factor as a function of resolution and indicating an average estimated error in the coordinates of the canavalin structure of about 0.25 Å.

exhibiting a highly asymmetric axial ratio. At the level of the polypeptide fold, the inherent dyad symmetry relating amino- and carboxyl-terminal halves is quite striking, and it is clear that an underlying domain motif is strictly maintained in both halves of the subunit. In retrospect, it is hardly surprising that the diffraction pattern from the rhombohedral crystals bore such a high degree of R32 pseudosymmetry.

Each domain of the subunit has a molecular mass of about 21,000 D and is composed of two structural elements (see also Fig. 4). One of these is a compact, eight-stranded β-barrel having the Swiss roll topology previously seen in a number of protein molecules (Richardson, 1981; Branden and Tooze, 1991). The second feature is a broad loop of extended chain that incorporates several α-helices, the longest being about two and one-half turns. It can be seen that the β-barrels are composed of residues from the primary and secondary amino-terminal portions of the polypeptides, whereas the extended, α-helix-containing loops are composed of carboxyl-terminal residues.

The lumina of the β-barrels enclose a preponderance of hydrophobic amino acid side chains but contain some hydrophilic residues as well. The interior volumes of the β-barrels are not entirely occluded, but they do not appear sufficiently spacious or accessible to contain very many water molecules. The OCMP and K_2HgI_4 heavy-atom derivatives used in the structure solution both bind at locations found just inside the β-barrel cavities; thus, the lumina must be at least partially occupied by solvent. Although views along the pseudo-dyad axes might suggest a channel passing along the axis and through the subunit, this is, in fact, not the case. That volume is filled by hydrophobic side chains, which exclude the passage of water.

The interface between the two domains of the canavalin subunit is shown in part in Figure 9. The interface, which is extensive, involves almost exclusively side chains from the strands of opposing β-barrels. The residues constituting the interface, presented in Table IV, are primarily hydrophobic, and we see little likelihood of water molecules being interposed. The two domains are thus joined and maintained primarily through the apposition of two large and complimentary hydrophobic surfaces. The interface between the two domains is bisected by the pseudo-2-fold axis of the subunit; thus, the two surfaces making up the interface are related by approximate dyad symmetry. The pronounced clustering of hydrophobic side chains on both sides of the interdomain contact area results in a dense hydrophobic core at the center of the canavalin subunit. One might well expect that the division of the canavalin subunit into separate domains would be highly unfavorable from an energetic standpoint because of the hydrophobic area that would be exposed to solvent. There are at least four salt bridges that occur at the interface as well, and these undoubtedly contribute to its formation and stability. In addition, there may also be hydrogen-bonding interactions, but we cannot identify them with certainty at this point.

The degree of structural redundancy within the canavalin subunit is most pronounced when the β-barrels of the two domains are optimally superimposed upon one another as in Figure 10. Although some variation is present in β-turns and connecting loops, the superposition of the individual strands of the barrels is good. The rms deviation in the positions of the 82 pairs of α carbon atoms making up the β-barrel strands

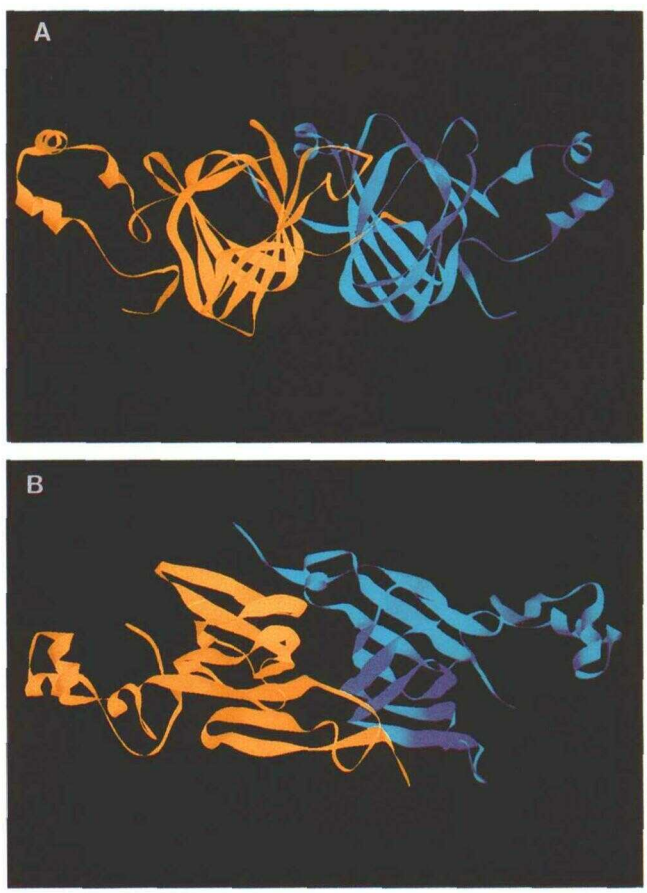

Figure 8. Computer-generated schematic diagrams of the canavalin subunit showing a view along the 3-fold axis (A) and a view along the pseudo-dyad axis (B) of the molecule with the amino-terminal domain in yellow and the carboxyl-terminal domain in blue.

is 0.783 Å. The two barrels, composed of residues 55 through 178 and 255 through 380, contain but a few identical residues, but are otherwise highly homologous in terms of sequence (Fig. 5). We noted that both β-barrels contain a prominent "β-bulge" (Richardson, 1981) at symmetrically equivalent Pro residues 101 and 296. In the "core domains," which exclude terminal strands and extended loops or connections, the 112 pairs of α carbon atoms of homologous residues have an rms deviation of 1.04 Å. The extended, α-helix-containing portions of the domains superimpose with an rms deviation of 1.37 Å.

Prominent in Figure 10 is one of the major differences between the two β-barrels. This involves the connecting loop between strands E and F in the amino-terminal domain contrasted with that between strands E and F in the carboxyl-domain. In one case, the connection contains only 2 residues, whereas in the homologous loop there are 14 residues. Among these additional amino acids are five consecutive Gln. This loop produces a prominent protrusion on one surface of the subunit, and, as described above, may account for the second trypsin cleavage site.

Figure 11 is a schematic representation of the secondary structural motifs of the two domains of the canavalin subunit showing the symmetrical hydrogen bond arrays of the two β-barrels, the α-helices in the extended loops, and hydrogen bonds contributed by the amino- and carboxyl-terminal strands of each domain. Amino acid residues that are identical and in structurally equivalent positions in the two domains and those residues identical between canavalin and five other vicilin proteins (Gibbs et al., 1989; Ng et al., 1993) are indicated.

The structural similarity between vicilins was further confirmed by comparing the α-carbon coordinates of the canavalin model with those of the phaseolin structure (Lawrence et al., 1990) taken from the Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987). This was carried out using the RIGID subroutine in FRODO. We identified 96 pairs of homologous α carbons in the corresponding amino-terminal β-barrels and these agreed with an rms deviation of 0.627 Å. For the carboxyl-terminal β-barrels, the rms deviation was 0.635 Å between 85 pairs of corresponding atoms. The loop regions and terminal strands of both domains were also compared and these had an rms difference of 1.46 Å between the 85 pairs of atoms. The overall rms deviation was 0.978 Å for the positions of 266 pairs of structurally homologous α carbons. In the carboxyl-terminal domain, the loops between strands A to B and E to F of phaseolin were considerably shorter than in canavalin. The fourth α-helix between the two domains of phaseolin was not observed in canavalin, and the amino-terminal strand of the second domain appears to be a little longer for canavalin. We suspect that cleavage by trypsin may account for the observed difference in the connecting region between the two domains of canavalin.

All nonhydrogen atoms of a single canavalin subunit are shown in Figure 12. Three subunits assemble around an exact 3-fold axis to generate the native trimeric molecule, which is seen in Figure 13. Figure 14 shows the canavalin molecule viewed along a pseudo-2-fold axis. The trimer is a toroid with an outside diameter of 86 to 88 Å and a large hole through its center that is approximately 18 Å in diameter. The thickness of the toroid is that of one subunit, or about 35 to 40 Å. Because of the nonexact dyad axes relating

Table IV. *Interaction at the interface between two domains*

Hydrophobic Residues at the Interface					
Domain I			Domain II		
Leu^{50}	Val^{88}	Leu^{160}	Leu^{251}	Leu^{286}	Val^{365}
Phe^{51}	Leu^{109}	Phe^{162}	Ile^{268}	Ile^{305}	Ile^{367}
Phe^{56}	Leu^{111}	Ile^{164}	Leu^{276}	Val^{307}	Val^{369}
Leu^{70}	Leu^{113}		Leu^{279}	Ile^{346}	
Phe^{73}	Ala^{137}		Leu^{282}	Ile^{348}	

Salt Bridges between the Two Domains	
Arg^{84}–Asp^{280}	Arg^{158}–Glu^{342}
Lys^{139}–Asp^{278}	Arg^{252}–Asp^{136}

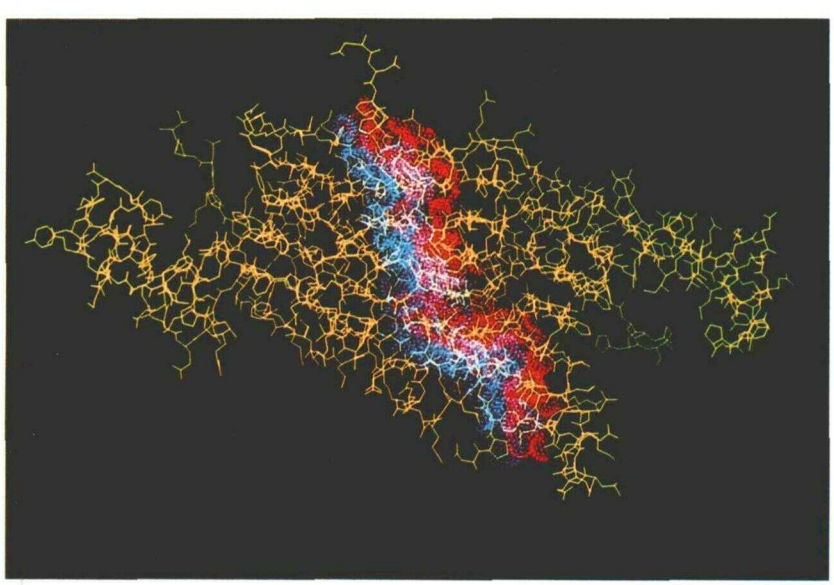

Figure 9. The two domains of a canavalin subunit, shown here in yellow, interface through the apposition of extensive hydrophobic surfaces. The atoms constituting these surfaces are represented in van der Waals form in blue and red. The cross-section shown here is typical of the entire interface and illustrates the likelihood that few if any water molecules could be interposed.

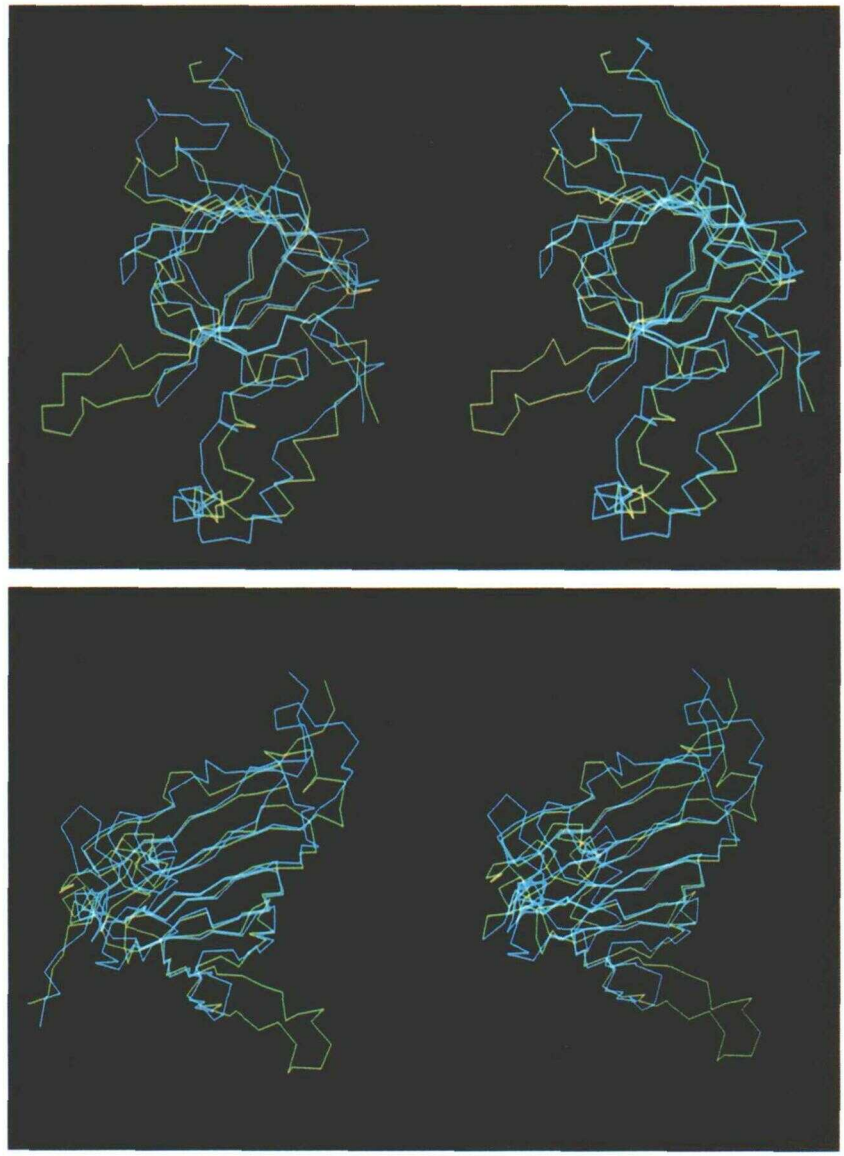

Figure 10. The α carbon tracings of the amino- (blue) and carboxyl-terminal (yellow) domains of a single canavalin subunit are optimally superimposed and shown in two arbitrary orientations. Although the connecting loops vary considerably, the strands of the two β-barrels align with high precision (rms deviation of 0.783 Å). The large yellow loop is that containing five consecutive Glus.

domains within the subunit, the entire molecule exhibits quasi-32-point group symmetry. The 3-fold axis, unlike the pseudo-dyad axes that relate domains, does not appear to be an organizational element in the structure. That is, the molecule does not assume its trimeric form to create a 3-fold symmetric motif, rather the triad arises as a consequence of the cyclic interfaces and interactions that produce subunit aggregation.

The extended, α-helix-containing loops account for the aggregation of the subunits into an oligomer and provide the interfaces between them. The amino acid residues involved are presented in Table V. They arise not only from the "loop" regions but also from one side of the "core" β-barrels. Similar to the domain interface within a subunit, these subunit interfaces involve the apposition of interdigitating hydrophobic surfaces, and these likely drive aggregate formation. Again, because of the pseudo-2-fold axes in canavalin, these inter-subunit interfaces also have an approximate dyad relationship. The area involved in this inter-subunit interface is as extensive as that relating domains within the subunit. However, the apparent complementarity between the two surfaces is not so striking as for the intra-subunit interface. Water molecules could possibly occur in or enter this region; thus, this inter-subunit interface may be more sensitive to its solvent environment. There are, in addition, six salt bridges at each subunit interface, and these might be expected to lend additional stability. We note that the oligomer undergoes dissociation to monomers at elevated pH (Kadima et al., 1990) and suspect that this may be a consequence of deprotonation of some of the basic groups involved in these interactions.

The inside surface of the prominent channel that passes along the 3-fold axis of canavalin is composed predominantly of hydrophilic side chains. Only 5 of roughly 25 to 30 side chains provided by each subunit on the inside of the channel are hydrophobic, and these occur in clusters that minimize the contact area with solvent. Six of the hydrophilic side

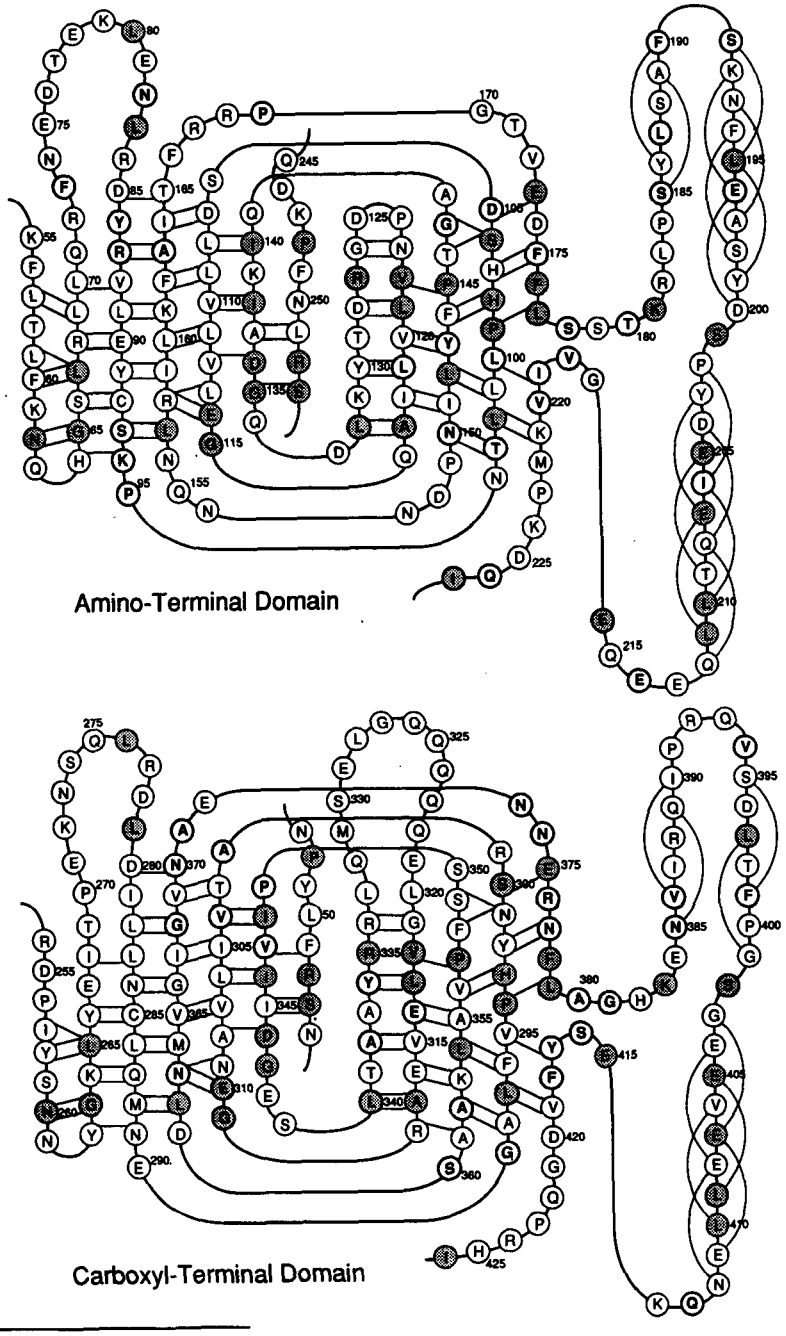

Figure 11. At top is a secondary structure diagram of the amino-terminal domain, where horizontal or diagonal bars represent likely main-chain hydrogen bonds. Residues that are conserved between domains are shaded, and those conserved among vicilins are circled in bold lines. Amino acids that are conserved both between domains and among vicilins carry both indications. At bottom is the equivalent secondary structural diagram for the carboxyl-terminal domain.

chains from each subunit are charged. Thus, the hole almost surely contains bulk water, and likely ordered water as well. Free exchange would be unimpeded.

No electron density was observed on the 3-fold axis that would indicate the presence of any ion, as was suggested by previous studies (McPherson, 1980; Smith et al., 1982). Indeed, the triad is virtually "out of reach" of any potential side chain ligands. This appears to be true for the canavalin molecule in all of its known crystal forms (T.-P.Ko., J.D. Ng, A. McPherson, unpublished data).

DISCUSSION

The most striking feature of the three-dimensional structure of canavalin is the repetitive domain motif, which reflects in turn the genetic duplication and the internal redundancy of the amino acid sequence. Although the repetition of the domains is explicable in terms of a tandemly duplicated primordial gene, what does not directly follow is their 2-fold spatial relationship. Although multiple and varied sequences can yield similar three-dimensional structures (Richardson,

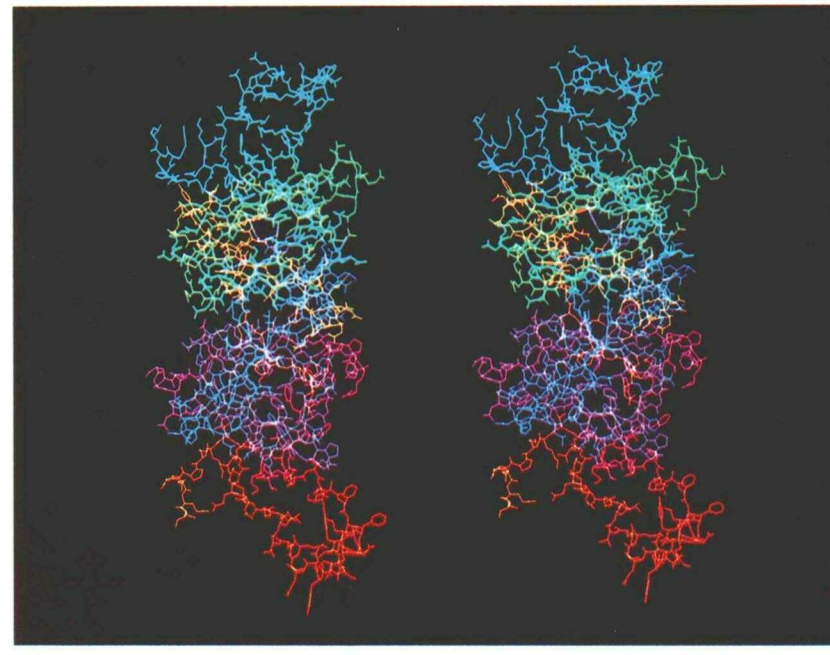

Figure 12. All nonhydrogen atoms of entire canavalin subunit are shown here with the polypeptide color coded according to the spectrum from amino to carboxyl terminus.

1981; Branden and Tooze, 1991), even within the same protein molecule, there is no reason why they should necessarily assume any symmetrical disposition to one another. That is, however, what is observed in canavalin. The question then is whether the dyad relationship and the associated symmetrical interfaces were an inherent feature of the primordial protein. That is, was it originally a dimeric protein having two identical subunits that became incorporated into one, or did the original double domain protein initially lack such a 2-fold relationship that only developed through evolutionary change?

Other proteins, among them enzymes, have been shown by x-ray diffraction analysis to contain multiple domains exhibiting similar relationships. Bovine liver rhodanese, for example, has two structurally homologous domains of the α/β class having an rms deviation of their α carbon atoms of about 1.95 Å (Ploegman et al., 1978). The two domains in this case, however, are not related by any symmetry axis.

Gamma crystallin, a lens protein, shows a similar structural redundancy to canavalin in that its two "Greek key" β-barrel domains also originate from the amino- and carboxyl-terminal portions of the polypeptide chain. These too are related by an approximate dyad axis of symmetry. The rms deviation of α carbon atoms in this case is about 1.4 Å for equivalent residues (Miller et al., 1983). This is about the same as for the two halves of the canavalin subunit that includes both

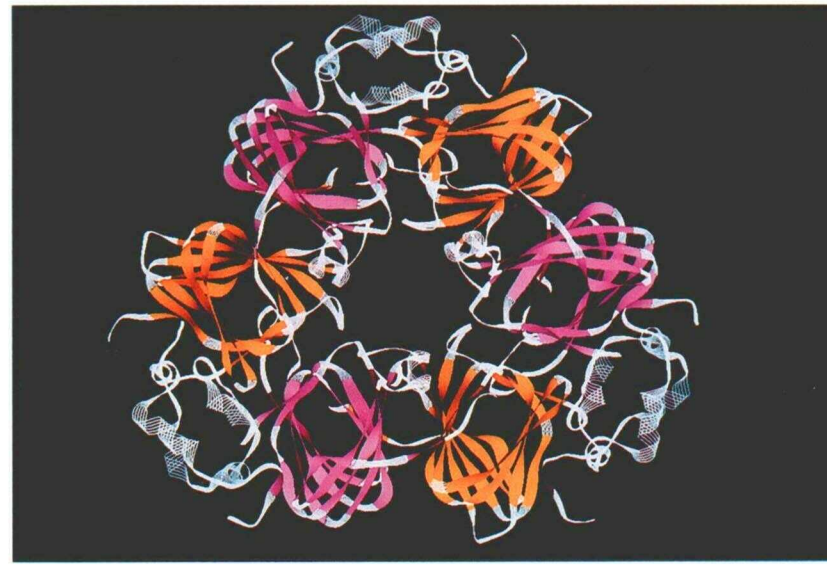

Figure 13. Computer-generated schematic diagram of the entire canavalin trimer with amino-terminal domains in orange and carboxyl-terminal domains in purple. The strands of the Swiss roll β-barrels that make up the cores of the two domains are solid colored, and the α-helices and extended chain elements forming the subunit-subunit interfaces are shown as transparent ribbons. The central hole along the triad is about 18 Å in diameter, and the circumference has a diameter of about 86 to 88 Å.

the β-barrel and the extended α-helix-containing chain, but is not nearly as close as for the β-barrel strands of the two canavalin domains, which show an rms deviation of their α carbon positions of only 0.78 Å.

Canavalin and the homologous protein phaseolin (Johnson et al., 1982; Gibbs et al., 1989; Lawrence et al., 1990) are, we believe, the first proteins solved by x-ray diffraction analysis to show a nearly identical double domain structure involving β-barrels having the Swiss roll motif. Furthermore, the two domains are of greater homology, as judged by rms deviations of α carbon atoms, than any previous examples and the domains are related by the most exact 2-fold axes so far observed. Thus, the relationship of the domains in vicilin proteins are an illustration of what is probably the most extreme case yet of structural duplication and symmetry within a single polypeptide chain.

The 3-fold axis of the canavalin trimer appears not to be an organizing element in the molecule, but it is curious nonetheless because it marks a broad channel of nearly 18 Å diameter running unobstructed through the protein. The volume of the channel is about 4% of the entire trimer. The canal is lined with hydrophilic and charged amino acid residues, and is almost certainly a conduit for water.

We do not know for certain whether the large intramolecular channel has functional significance, as for storage and transport of solvent or metabolites, but it is an attractive idea. In three of the four crystal forms of canavalin (T.-P.Ko., J.D. Ng, A. McPherson, unpublished data), the disk-shaped molecules are stacked like plates so that the channels passing through the trimers are contiguous throughout the crystals. Thus, crystals of canavalin are, in a sense, a vast network of solvent-filled microtubules that penetrate the infrastructure along 3-fold directions.

The stacking of the trimers observed in the different crystal forms undoubtedly reflects a preferred packing motif. Such arrangements could be assumed as well when the canavalin molecules are packed and stored inside the seed. Thus, the naturally occurring protein array may be similarly provided with a natural aqueduct system to "pipe in" water to the interior core of the protein inclusion bodies as imbibition occurs, germination begins, and storage reserves are needed for development.

Table V. *Interaction between two adjacent subunits*

Hydrophobic residues at the interface

Leu^{98}	Phe^{194}	Met^{222}	Phe^{352}	Ile^{390}
Val^{122}	Leu^{195}	Leu^{293}	Ala^{355}	Leu^{397}
Phe^{176}	Ala^{197}	Val^{318}	Phe^{378}	Phe^{399}
Leu^{183}	Ile^{206}	Leu^{320}	Ala^{380}	Val^{406}
Leu^{187}	Leu^{210}	Met^{331}	Val^{386}	Leu^{409}
Ala^{189}	Leu^{211}	Leu^{333}	Ile^{387}	Leu^{410}
Phe^{190}				

Salt bridges between adjacent subunits

Arg^{167}–Glu^{81}	Arg^{301}–Asp^{107}	Lys^{414}–Asp^{200}
Arg^{168}–Glu^{321}	Arg^{376}–Asp^{125}	Arg^{424}–Glu^{196}

The vicilin proteins, as exemplified by canavalin and phaseolin, provide an ideal system for the comparative analysis of amino acid sequence with three-dimensional structure and for the investigation of evolutionary divergence. Canavalin and phaseolin are highly homologous (Gibbs et al., 1989), exhibiting 60% identity of amino acids, and they have very similar three-dimensional structures, including the internally duplicated domains. Gibbs et al. (1989) have further shown that the similarities are not limited to these proteins but extend to other vicilins such as conglycinin and pea vicilin. The homology among conglycinin, pea vicilin, and phaseolin was also noted and reported by Doyle et al. (1986).

Identification and description of amino acids that are conserved between domains and among vicilins defines the specific residues or constellations required for structural integrity, for proper polypeptide folding, or for ensuring some functional property. Figure 15 shows the three-dimensional distribution of the conserved amino acid residues in the

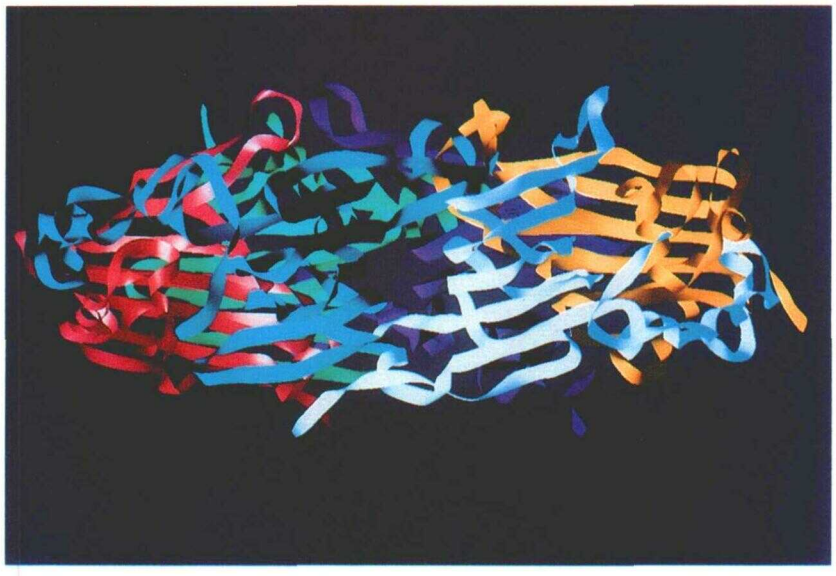

Figure 14. A schematic diagram of the canavalin trimer seen perpendicular to the 3-fold axis and along one of the pseudo-dyad axes. In this illustration, each domain is a different color and the weaving of the strands of the β-barrels is clearly seen. The disk-shaped molecule has a thickness of about 35 to 40 Å.

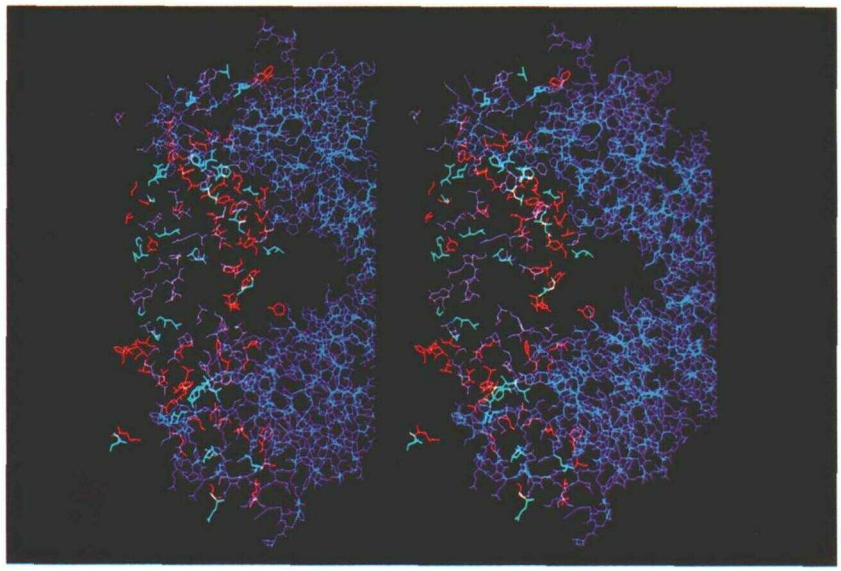

Figure 15. Distribution in space of the conserved residues in canavalin. The purple side chains represent those conserved between the two domains of the subunit, the red side chains are those conserved between canavalin and five other vicilin proteins, and the green side chains are those conserved both between domains and among vicilins. The distribution of conserved residues has been placed appropriately in the volume occupied by a subunit in a canavalin trimer, where the additional two protein subunits are shown in blue.

canavalin subunit. Those shown in green may be thought of as hyperconserved in that they are not only found to be identical at equivalent positions in both domains of canavalin, but are conserved in both domains of five other vicilin proteins as well (Gibbs et al., 1989; Ng et al., 1993).

Of the 90 amino acid residues that show conservation, roughly two-thirds occur at the interfaces between subunits within the trimer. These include an equal proportion of hydrophobic and hydrophilic side chains. The large number of conserved residues at the interfaces suggest that oligomer formation is a crucial property of the polypeptide. On the other hand, there are very few, if any, conserved residues at the extensive hydrophobic interface between domains. If anything, the great majority of conserved side chains are on the exterior of the subunit.

Of the hyperconserved amino acids, two are Glys and two are Pros, five others are hydrophilic, and the last seven are hydrophobic. Although Gly and Pro residues are frequently conserved among homologous proteins because of their unique effects upon local polypeptide conformation (Richardson, 1981; Branden and Tooze, 1991), we see no ready explanation for the preservation of the others. There does not appear to be any clustering of conserved residues in three-dimensional space, although inspection of Figure 15 suggests some clustering of conserved residues in terms of primary and secondary structure. As noted by Ng et al. (1993), some of the conserved amino acids have been implicated in protein sorting and trafficking; thus, their conservation has less bearing on their structural role than on their recognition and interaction with other proteins in the cell. One curious observation is that, of the 18 residues strongly conserved between domains and among vicilins, 7 are Leu, a number out of proportion with their occurrence in the sequence. These do not, in general, appear at the inter-subunit contact areas, but are distributed quite arbitrarily throughout the three-dimensional structure.

A fundamental objective of this investigation was to delineate regions of the protein that might be insensitive to structural alteration and to identify sites where substitutions or insertions might reasonably be made. Our finding that among the most highly conserved residues between domains and among vicilins are seven Leu that fall at what appear to be arbitrary locations illustrates that this objective may be more difficult to attain than one might otherwise have thought. Nonetheless, there are sites that seem attractive and we have, in fact, begun efforts to modify the protein.

The large channel along the 3-fold axis of canavalin, for example, exposes 25 to 30 principally hydrophilic side chains to solvent. Most of these are not conserved residues. In addition, the volume of this cavity, even though it would have to accommodate 3-fold symmetrical changes, is still sufficiently great that substantial amino acid insertions in the sequence might be tolerated.

The loop of extended chain between residues 320 to 333 that contains five consecutive Glus, but is absent at the corresponding location in the amino-terminal domain and in all other vicilins, is another promising site. Change or insertion in this loop could be easily accommodated in the exterior solvent and would not, so far as we can see, disturb the packing of the protein molecules in the crystals.

Certain regions of the molecule, on the other hand, appear particularly sensitive both in terms of conserved residues and for structural reasons. These are the regions at the interface between domains, and even more so, the residues constituting the interfaces between the subunits within the trimer. Other amino acids, not indicated by conservation or structure, could play essential roles in either the maintenance of structure or function, but these we would hope to identify by observing changes in their levels of expression or changes in the mutant structures as visualized in difference Fourier experiments.

The extensive and detailed structural similarity that we observe between canavalin and phaseolin, our understanding of the structure, and the close genetic relationship between vicilins all suggest that other vicilin proteins may be constructed from these two examples by careful modeling. We have shown the rms deviation between 266 pairs of homol-

ogous α carbons of canavalin and phaseolin to be 0.978 Å. A more detailed structural comparison between these two proteins in conjunction with the known sequences of several vicilin-class proteins is underway. Through these analyses, we may arrive at some canonical rules regarding the structures of vicilin proteins. Substitution of appropriate amino acid sequences with subsequent application of molecular dynamics methods (Brunger, 1988, 1991) hopefully would allow one to specify those alterations that distinguish the various storage proteins and to produce suitable structural models.

The structure presented here is that of the proteolytically modified canavalin molecule, and the question might well be raised whether it is entirely representative of the structure of the native molecule. Although we have no assurance that every detail is the same, the similarity of the canavalin structure to that of the intact phaseolin structure (Lawrence et al., 1990) suggests that the molecule presented here is essentially that of the native plant canavalin. Why the native molecule cannot be crystallized, whereas the cleaved form can, remains unknown.

We believe that the protein structure that we have determined will provide us with a good foundation for the engineering of the vicilin proteins (Oxender and Fox, 1987) through the use of site-directed mutagenesis. Given the propensity of the protein to crystallize and the rapidity with which we can carry out difference Fourier studies of altered forms, we believe that the systematic improvement of the protein is now possible. Indeed, we have already begun experiments to demonstrate the value of canavalin in this regard.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the assistance of John Day, Aaron Greenwood, Marie Greene, and Debora Felix.

Received August 7, 1992; accepted November 23, 1992.
Copyright Clearance Center: 0032-0889/93/101/0729/16.

LITERATURE CITED

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J (1987) Protein Data Bank. *In* FH Allen, G Bergerhoff, R Sievers, eds, Crystallographic Databases Information Content, Software Systems, Scientific Applications. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp 107–132

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol **112:** 535–542

Bewley JD, Black M (1978) Physiology and Biochemistry of Seeds. Springer-Verlag, New York

Blow DM, Crick FHC (1959) The treatment of errors in the isomorphous replacement method. Acta Crystallogr **12:** 794–804

Branden C, Tooze J (1991) Introduction to Protein Structure. Garland, New York

Brayer GD, McPherson A (1983) Refined structure of the gene 5 DNA binding protein from bacteriophage fd. J Mol Biol **169:** 565–596

Brunger AT (1988) Crystallographic refinement by simulated annealing: application to a 2.8 Å resolution structure of aspartate aminotransferase. J Mol Biol **203:** 803–816

Brunger AT (1991) Simulated annealing in crystallography. Annu Rev Phys Chem **42:** 197–223

Brunger AT, Kuriyan J, Karplus M (1987) Crystallographic R factor refinement by molecular dynamics. Science **235:** 458–460

Carson M, Bugg CE (1986) An algorithm for ribbon models of proteins. J Mol Graph **4:** 121–122

Creamer LK, Jimenez-Flores R, Richardson T (1988) Genetic modification of food proteins. Trends Biotechnol **6:** 163–169

Day J, McPherson A (1992) Macromolecular crystal growth experiments on International Microgravity Laboratory-1. Protein Sci **1:** 1254–1268

DeLucas LJ, Suddath FL, Snyder R, Naumann R, Broom MB, Pusey M, Yost V, Herren B, Carter D, Nelson B, Meehan EJ, McPherson A, Bugg CE (1986) Preliminary investigations of protein crystal growth using the space shuttle. J Crystallogr Growth **76:** 681–693

Derbyshire E, Wright DJ, Boulter D (1976) Legumin and vicilin, storage proteins of legume seeds. Phytochemistry **15:** 3–24

Doyle JJ, Schuler MA, Godette WD, Zenger V, Beachy RN, Slightom JL (1986) The glycosylated seed storage proteins of *Glycine max* and *Phaseolus vulgaris*. J Biol Chem **261:** 9228–9238

Gibbs PEM, Strongin KB, McPherson A (1989) Evolution of legume seed storage proteins: a domain common to legumins and vicilins is duplicated in vicilins. Mol Biol Evol **6:** 614–623

Hamlin R, Cork C, Howard A, Nielson C, Vernon W, Matthews D, Xuong NH (1981) Characteristics of a flat multiwire area detector for protein crystallography. J Appl Crystallogr **14:** 85–89

Hennig M, Schlesier B, Dauter Z, Pfeffer S, Betzel C, Hohne WE, Wilson KS (1992) A TIM barrel without enzymatic activity? Crystal structure of narbonin at 1.8 Å resolution. FEBS Lett **306:** 80–84

Johnson S, Grayson G, Robinson L, Chahade R, McPherson A (1982) Biochemical and crystallographic data for phaseolin, the storage protein from *Phaseolus vulgaris*. Biochemistry **21:** 4839–4843

Jones TA (1982) FRODO: a graphics fitting program for macromolecules. *In* D Sayre, ed, Computational Crystallography. Clarendon Press, Oxford, pp 303–317

Jones TA (1985) Interactive computer graphics: FRODO. Methods Enzymol **115:** 157–171

Kadima W, McPherson A, Dunn MF, Jurnak FA (1990) Characterization of precrystallization aggregation of canavalin by dynamic light scattering. Biophys J **57:** 125–132

Lawrence MC, Suzuki E, Varghese JN, Davis PC, Van Donkelaar A, Tulloch PA, Colman PM (1990) The three-dimensional structure of the seed storage protein phaseolin at 3 Å resolution. EMBO J **9:** 9–15

Luzatti PV (1952) Traitement statistique des erreurs dans la determination des structures cristallines. Acta Crystallogr **5:** 802–810

McPherson A (1980) The three-dimensional structure of canavalin at 3.0 Å resolution by X-ray diffraction analysis. J Biol Chem **255:** 10472–10480

McPherson A (1982) Crystallization. *In* Preparation and Analysis of Protein Crystals. John Wiley & Sons, New York, pp 82–159

McPherson A (1990) Current approaches to macromolecular crystallization. Eur J Biochem **198:** 1–23

McPherson A, Greenwood A, Day J (1991) The effect of microgravity on protein crystal growth. Adv Space Res **11:** 343–356

McPherson A, Rich A (1973) X-ray crystallographic study of the quaternary structure of canavalin. J Biochem (Tokyo) **74:** 155–160

McPherson A, Spencer R (1975) Preliminary structure analysis of canavalin from jack bean. Arch Biochem Biophys **169:** 650–661

Miller L, Lindley P, Blundell T (1983) X-ray analysis of the eye lens protein gamma crystallin at 1.9 Å resolution. J Mol Biol **170:** 175–202

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM (1992) Stereochemical quality of protein structure coordinates. Proteins **12:** 345–364

Morrison R, Delozier G, Robinson L, McPherson A (1984) Biochemical and X-ray diffraction analysis of concanavalin B crystals from jack bean. Plant Physiol **76:** 175–183

Ng J, Ko T-P, McPherson A (1993) Cloning, expression, and crystallization of jack bean (*Canavalia ensiformis*) canavalin. Plant Physiol **101:** 713–728

Oxender DL, Fox CF, eds (1987) Protein Engineering. Alan R. Liss, New York

Ploegman JH, Drenth G, Kalk KH, Hol WGJ (1978) Structure of

bovine liver rhodanese. I. Structure determination at 2.5 Å resolution and a comparison of the conformation and sequence of its two domains. J Mol Biol **123:** 557–594

Ramachandran GN, Sassiekharan V (1968) Conformation of polypeptides and proteins. Adv Protein Chem **28:** 283–437

Reeke GN Jr, Becker JW, Wang JL, Cunningham BA, Waxdal MJ, Edelman GM (1975) The covalent and three-dimensional structure of concanavalin A (part I–IV). J Biol Chem **250:** 1490–1547

Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem **34:** 63–100

Rossmann MG (1960) The accurate determination of the position and shape of heavy-atom replacement groups in proteins. Acta Crystallogr **13:** 221–226

Smith SC, Johnson S, Andrews J, McPherson A (1982) Biochemical characterization of canavalin, the major storage protein of jack bean. Plant Physiol **70:** 1199–1209

Sumner JB (1919) The globulins of the jack bean, *Canavalia ensiformis*. J Biol Chem **37:** 137–142

Sumner JB (1926) The isolation and crystallization of the enzyme urease. J Biol Chem **69:** 435–441

Sumner JB, Howell SF (1936) The isolation of a fourth crystallizable jack bean globulin through the digestion of canavalin with trypsin. J Biol Chem **113:** 607–610

Sussman JL (1985) Constrained-restrained least-squares (CORELS) refinement of proteins and nucleic acids. Methods Enzymol **115:** 271–303

Sussman JL, Holbrook SR, Church GM, Kim SH (1977) A structure factor least-squares refinement procedure for macromolecular structures using constrained and restrained parameters. Acta Crystallogr **A33:** 800

Ten Eyck L, Weaver LA, Matthews BW (1976) A method of obtaining a stereochemically acceptable protein model which fits a set of atomic coordinates. Acta Crystallogr **A32:** 349–355

Terwilliger T, Eisenberg D (1983) Unbiased three dimensional refinement of heavy atom parameters by correlation of origin removed Patterson functions. Acta Crystallogr **A39:** 813–817

Terwilliger TC, Weissman L, Eisenberg D (1982) The structure of melittin in the form I crystals and its implication for melittin's lytic and surface activities. Biophys J **37:** 353–363

Tronrud DE, Ten Eyck LF, Matthews BW (1987) An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Crystallogr **A43:** 489–501

Wang B-C (1985) Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol **115:** 90–112

Watenpaugh KD (1985) Overview of phasing by isomorphous replacement. Methods Enzymol **115:** 3–15

Xuong N-H, Nielson C, Hamlin R, Anderson D (1985) Strategy for data collection from protein crystals using a multiwire counter detector diffractometer. J Appl Crystallogr **18:** 342–360