Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The refined structure of canavalin from jack bean in two crystal forms at 2.1 and 2.0 Å resolution

Tzu-Ping Ko,[a] John Day[b] and Alexander McPherson[b]*

[a]Institute of Biological Chemistry, Academia Sinica, Taipei 11529, Taiwan, and [b]Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697-3900, USA

Correspondence e-mail: amcphers@uci.edu

Received 5 November 1999
Accepted 7 February 2000

PDB References: canavalin, cubic, 2cau; canavalin, hexagonal, 2cav.

The structure of canavalin was refined to 2.1 and 2.0 Å resolution in cubic and hexagonal crystals of space group $P2_13$ and $P6_3$, respectively. The threefold molecular symmetry is expressed in the symmetry of both crystals, where each identical subunit is an asymmetric unit. The canavalin subunit consists of two very similar domains, each comprised of a core subdomain having Swiss-roll topology with a loop subdomain that contains helices. The refined canavalin models resolved the discrepancy in amino-acid registers of the secondary-structural elements compared with phaseolin. The presence of strand Z in both domains of canavalin was confirmed and a new helix in the loop between strands A and B of each domain was observed. The models were analyzed in terms of the duplicated vicilin domains. Three strictly conserved residues, two glycines and a proline, were identified. The similarity between entire vicilin molecules is greater than that between separate domains of canavalin and phaseolin. Homology modeling of the sucrose-binding protein (SBP) from soybean showed a plausible trimeric assembly of subunits similar to that of vicilins.

1. Introduction

Canavalin constitutes a major fraction of the soluble protein from jack bean (*Canavalia ensiformis*). It is a trimer with a molecular mass of 142 kDa and belongs to the vicilin class of storage proteins (Derbyshire *et al.*, 1976; Smith *et al.*, 1982). This protein was first isolated and, after treatment with trypsin, crystallized in a rhombohedral unit cell by Sumner & Howell (1936). X-ray diffraction revealed a pseudo-32 point-group arrangement of subunits in the molecule, reflected particularly in the rhombohedral crystal (McPherson & Spencer, 1975). The canavalin gene was cloned, sequenced and expressed in *Escherichia coli* (Ng *et al.*, 1993) and MIR phasing of the hexagonal crystal (Ko, Ng & McPherson, 1993) enabled construction of a detailed model. This model was then used successfully to solve the structures of three other crystal forms, rhombohedral, orthorhombic and cubic, by molecular-replacement methods (Ko, Ng, Day *et al.*, 1993). Although the cubic crystal with space group $P2_13$ appeared only occasionally, it contained less solvent and diffracted to a resolution of nearly 2.1 Å, which is significantly higher than other conventionally grown canavalin crystals. Under microgravity conditions (Day & McPherson, 1992), large hexagonal crystals which diffracted to about 2 Å resolution were eventually obtained aboard the US Space Shuttle (Koszelak *et al.*, 1995). Data from those microgravity-grown crystals were used in the refinement of the hexagonal canavalin crystals described here.

The structure of a similar vicilin-class protein, phaseolin, from French bean (*Phaseolus vulgaris*), was determined at 3 Å in one crystal form and refined in a second to 2.2 Å resolution (Lawrence *et al.*, 1990, 1994). Comparison of the X-ray structures of canavalin and phaseolin along with homologous sequence alignment of other vicilin-class proteins revealed canonical structural elements of this family of storage proteins (Lawrence *et al.*, 1994).

An individual protein subunit can be described as a pair of almost identical domains (or 'modules') organized about a pseudo-dyad axis perpendicular to the molecular threefold axis, as in Fig. 1. Each domain contains a 'core' β-barrel subdomain of Swiss-roll topology with an extended 'loop' subdomain that contains three helices. Two core subdomains are tightly associated *via* an extensive hydrophobic interface, possibly sealed from solvent penetration, while loop subdomains are responsible for subunit assembly into trimers (Ko, Ng & McPherson, 1993). A fourth helix joins the N-terminal and C-terminal domains in phaseolin, but in canavalin no linkage was visible. When strands of the β-barrels, particularly the N-terminal strands (designated Z; Lawrence *et al.*, 1994), of canavalin and phaseolin were compared, a few amino-acid residues also appeared to be out of register with respect to the homology sequence alignment.

As an extension of previous work, the model of canavalin was refined further using the 2.1 Å data obtained from the cubic crystal and new 2.0 Å data collected from hexagonal crystals grown under microgravity conditions (Koszelak *et al.*, 1995). Water molecules, which were previously neglected, are now included.

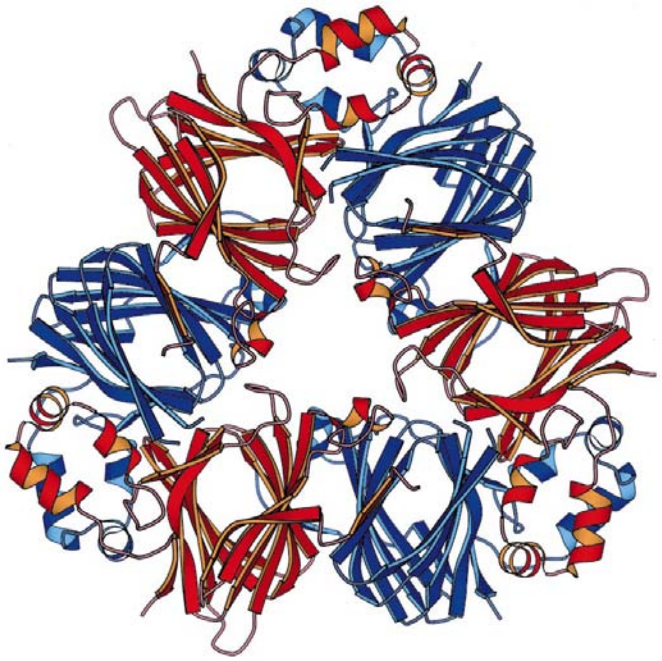

Figure 1
Ribbons diagram of the canavalin trimer as seen along the molecular threefold axis. The N-terminal and the C-terminal domains, colored red and blue, respectively, are arranged about the threefold axis and related by twofold axes perpendicular to the threefold, rendering a pseudo-$D3$ point-group symmetry. This figure was produced using *MOLSCRIPT* (Kraulis, 1991).

Table 1
Data-collection statistics for the canavalin crystals.

Numbers in parentheses are for the highest resolution shell.

Space group	$P2_13$	$P6_3$
Unit-cell parameters (Å, °)	$a = b = c = 105.99$, $\alpha = \beta = \gamma = 90$	$a = b = 126.43$, $c = 51.37$, $\alpha = \beta = 90$, $\gamma = 120$
Resolution (Å)	2.02 (2.18–2.02)	2.00 (2.15–2.00)
No. of observations	71863 (5944)	337907 (29912)
No. unique	22331 (3415)	31848 (6302)
Completeness (%)	85.5 (66.5)	99.7 (99.6)
$R_{\rm merge}$† (%)	4.8 (18.4)	10.1 (31.2)
Average $I/\sigma(I)$	21.4 (3.1)	24.5 (3.4)

† $R_{\rm merge} = \sum_{hkl}\sum_j |I_{\rm ave} - I_{{\rm obs},j}| / \sum_{hkl}\sum_j I_{\rm ave}$, in which $I_{\rm ave}$ is the average intensity of j equivalent reflections $I_{{\rm obs},j}$ and the sum is over all reflections that have equivalents in the unmerged data set.

Both canavalin and phaseolin have been exploited as models in a number of other studies, for example, methionine enhancement of phaseolin (Dyer *et al.*, 1995), solubility modification of cocoa vicilin (Warwicker & O'Connor, 1995) and mapping IgE-binding epitopes on peanut allergens (Shin *et al.*, 1998). The vicilin structure was also used to build a model of oxalate oxidase, where it predicted a possible metal-binding site (Gane *et al.*, 1998).

Homologous sequence alignment revealed a probable evolutionary relationship between vicilins and legumins, another major class of storage proteins in legume seeds (Gibbs *et al.*, 1989) and Shutov *et al.* (1995) later produced convincing evidence for a common single-domain ancestor. They further extended the homology to a fungal desiccation protein (Baumlein *et al.*, 1995) and to fern-spore protein (Shutov *et al.*, 1998), as well as other seed proteins. The sucrose-binding protein (SBP) of soybean, first characterized by Grimes *et al.* (1992), is also highly homologous to canavalin. This 62 kDa protein is expressed in young leaves, mature phloem and developing cotyledons, and probably mediates non-saturable sucrose uptake in these plant cells (Overvoorde *et al.*, 1996). Although its function is distinct from storage proteins (Overvoorde *et al.*, 1997), we constructed a homology model of SBP based on the vicilin structure, which might possibly be useful in terms of molecular organization.

2. Materials and methods

2.1. Crystallization and data collection

Canavalin was prepared from defatted jack-bean meal as described previously (Smith *et al.*, 1982; Sumner & Howell, 1936). The procedures include treatment with trypsin, bulk crystallization and three recrystallizations. Hexagonal crystals were initially grown by vapor diffusion at 277–281 K from solutions containing 15–20 mg ml^{-1} protein, 1.0% NaCl and 50 m*M* phosphate buffer pH 6.8. Crystals were later obtained aboard the US Space Shuttle under similar conditions but at 293 K by liquid–liquid diffusion (Koszelak *et al.*, 1995). The crystals were mounted in quartz capillaries and X-ray diffraction data were collected and processed using a San Diego Multiwire Systems detector with a Rigaku RU-200 rotating-anode generator and *SDMS* software package as described previously (Ko, Ng, Day *et al.*, 1993). These microgravity-grown crystals allowed extension of the diffraction resolution from 2.6 Å to beyond 2.0 Å. The cubic crystals were difficult to reproduce (Ko, Ng, Day *et al.*, 1993), but substantial improvement was obtained when the diffraction images were reprocessed using new programs and parameters. X-ray data statistics are listed in Table 1.

2.2. Structure refinement

Refinement was carried out using *X-PLOR* (Brünger, 1992*a*) on Silicon Graphics Indigo II, Indy and O2 computers. A randomly selected 8% of the data was set aside for cross validation using the $R_{\rm free}$ value (Brünger, 1992*b*). Low-resolution data between 40 and 8 Å were not included until most solvent molecules were included, after which the bulk-solvent option of *X-PLOR* was employed (Jiang & Brünger, 1994). The refinement utilized data with a $F/\sigma(F)$ ratio greater than 2; however, map calculations consistently included all reflections. Manual rebuilding of the model and addition of solvent molecules were performed using *TOM* (Jones, 1982) and analyses of the models employed *PROCHECK* (Laskowski *et al.*, 1993) and the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994).

Of the canavalin crystals grown on earth, the cubic form diffracted X-rays to highest resolution. Thus, the refinement of the canavalin structure started with the cubic crystal. The coordinates of PDB entry 1cau served as an initial model and yielded an overall R of 0.281 for all data having $F > 2\sigma(F)$. An initial $2F_o - F_c$ map showed weak density for residues 44–51, 207–217, 222–224, 241–247, 321–332 and 421–424, which correspond to terminal strands or flexible loops. Exclusion of these residues gave an R value and an $R_{\rm free}$ of 0.266 and 0.362, respectively, after simulated annealing for data in the resolution range 8–2.1 Å. Subsequent modifications included reconstruction of N-terminal strands and surface loops, adjustment of side chains and addition of water molecules, guided by the $2F_o - F_c$ maps alternating with refinement. Waters having B values greater than 60 and not within 1.5σ density were deleted. After convergence, the model contained residues 46–223, 246–322, 332–421 and 63 waters, and yielded

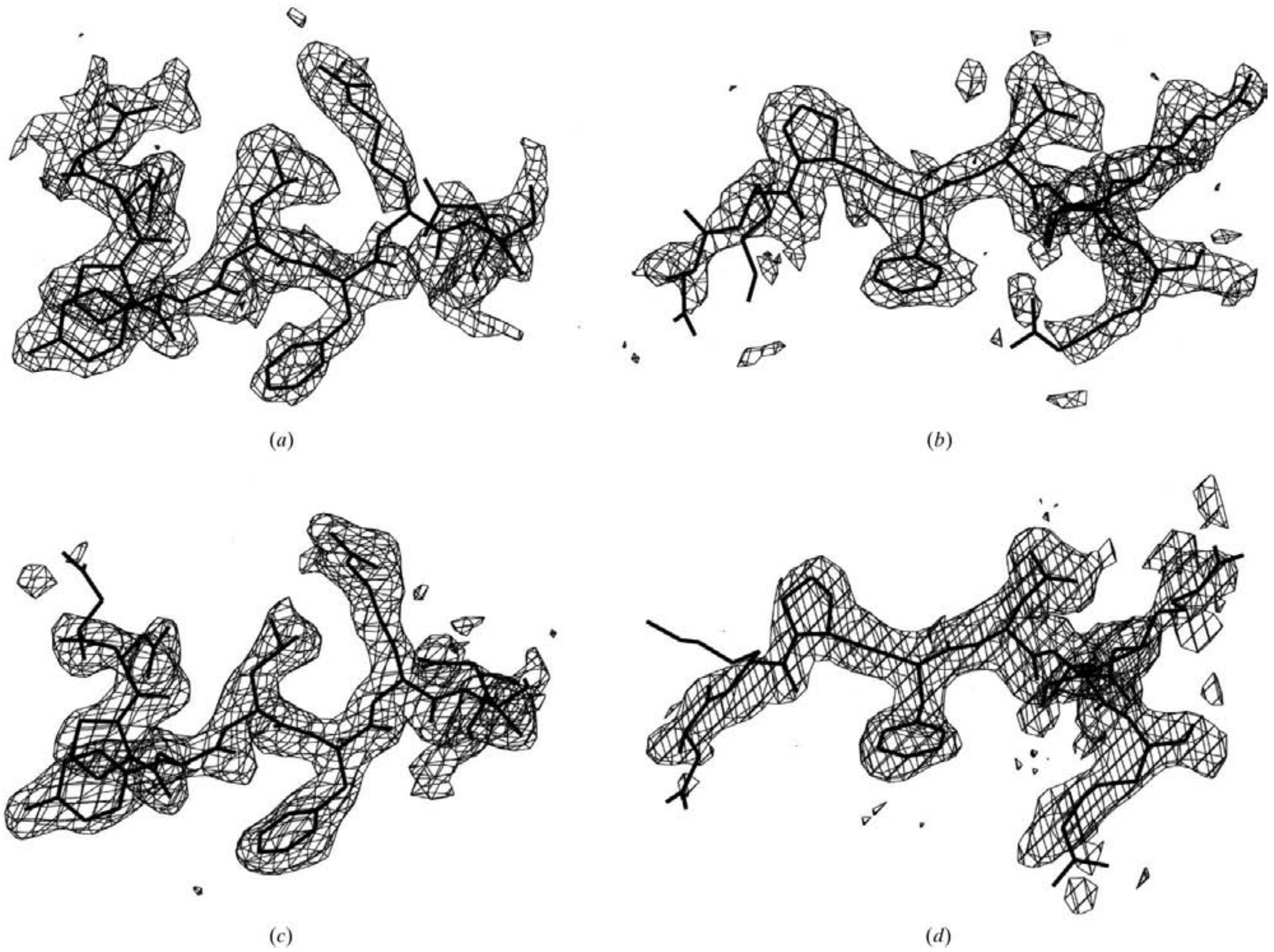

Figure 2
Omit maps of strands Z. Maps were calculated by omitting the N-terminal ten residues from the canavalin domains and simulated annealing at 1000 K using *X-PLOR*. The contour level is 1.0σ and only selected regions of the maps are shown. Protein models (shown in heavy lines) of the strands Z in (*a*) the N-terminal domain of the cubic crystal, (*b*) the C-terminal domain of the cubic crystal, (*c*) the N-terminal domain of the hexagonal crystal and (*d*) the C-terminal domain of the hexagonal crystal are superposed on the maps. The sequences for the strands Z are Asn-Asn-Pro-Tyr-Leu-Phe-Arg-Ser-Asn in (*a*) and (*c*), and Asp-Lys-Pro-Phe-Asn-Leu-Arg-Ser-Arg in (*b*) and (*d*). The strands Z have similar conformations in both crystals. This figure was produced using *TOM*.

an R value of 0.181 and an $R_{\rm free}$ of 0.253. Ten omit maps excluding up to 40 successive residues along the polypeptide chain were computed to recheck precise conformations. In the final modification, there was no change in the number of amino-acid residues, while the number of waters was increased to 101. Using the bulk-solvent correction, the final R value and $R_{\rm free}$ were 0.175 and 0.235, respectively, for data in the resolution range 40–2.1 Å. Other statistics are found in Table 2.

When the model refined in the cubic unit cell was placed in the hexagonal crystal by superimposing it upon the previously determined structure (Ko, Ng & McPherson, 1993), an overall R value of 0.316 was obtained for all $F > 2\sigma(F)$ data after rigid-body refinement. Initial $2F_o - F_c$ maps showed good superposition of all protein atoms, with only a few exceptional side chains. Densities at the C-terminal strands of both domains indicated a possible extension of one or two residues. Near the end of strand E, residues 321–322 displayed weak density and were deleted. The model was adjusted and 80 waters were included. After simulated annealing and B-factor refinement, the R value and $R_{\rm free}$ fell to 0.233 and 0.292, respectively, in the resolution range 8–2.0 Å. Further iterative applications of real-space and reciprocal-space adjustment of the model, as described for the cubic crystal, were performed. The final model contained residues 46–225, 246–320 and 332–422, plus 164 water molecules. It yielded an R value and an $R_{\rm free}$ of 0.208 and 0.264, respectively, for 40–2.0 Å data with bulk-solvent correction. Statistics are also included in Table 2. Samples of omit maps superimposed upon the final models are shown in Fig. 2.

2.3. Homology modeling and comparison

Comparison of structures and modeling of SBP were performed using the program *O* (Jones *et al.*, 1991). The model of phaseolin was obtained from the PDB (entry 2phl). According to the sequence alignment of Overvoorde *et al.* (1997), also shown in Fig. 3, the polypeptide templates of residues 46–222, 246–321 and 332–421 were extracted from the refined cubic model of canavalin. These correspond to residues 107–285, 309–386 and 402–511 in the SBP sequence. After a least-squares fit with the canavalin model, residues 196–210 of phaseolin were extracted to provide helix 4 of SBP (residues 286–300). Side chains were substituted where necessary and torsional angles were adjusted to eliminate unfavorable contacts. Insertions and deletions at residues Leu219, Gln277, Asp334, Ser429 and Asp455 were made using the *LEGO* procedure and the database of *O*. The loop near Asp115 adopted the conformation of phaseolin and that near Glu432 was built with reference to the equivalent loop structure in the N-terminal domain. The resulting SBP model contained 3565 non-H atoms in 363 amino-acid residues. The larger insertions of residues 387–401 and 485–503, the linker residues 301–308 between the N- and C-terminal domains, as well as the terminal residues of 1–106 and 512–524 all remained absent. Molecular dynamics and energy minimization using *X-PLOR* were then carried out while restraining C^α coordinates to the initial model.

3. Results and discussion

3.1. Refinement

The model of canavalin in the cubic unit cell contains 2767 non-H protein atoms in 344 amino-acid residues and 101 water molecules. It yielded a conventional R and an $R_{\rm free}$ of 0.175 and 0.235, respectively, for 19 695 reflections. The model in the hexagonal crystal contains 2784 atoms in 346 residues plus 164 waters. For 30 894 reflections, R and $R_{\rm free}$ were 0.208 and 0.264, respectively. The models, summarized in Table 2, have good geometry and have r.m.s. deviations from ideal bond lengths and angles of only 0.010 Å and 1.57°, respectively, in the cubic crystal, and 0.013 Å and 1.71°, respectively, in the hexagonal crystal. Overall temperature factors are 28.3 and 37.1 Å^2, respectively. In the cubic crystal model, 27 protein atoms have B values greater than 60, while in the hexagonal model there are 81. All outstanding B values occurred in surface loops and terminal strands. The Luzzati plots (Luzzati, 1952) show estimated coordinate errors of less than 0.22 and 0.28 Å for the cubic and hexagonal crystals, respectively. In *PROCHECK* (Laskowski *et al.*, 1993), both models have 91.1% of the 304 non-

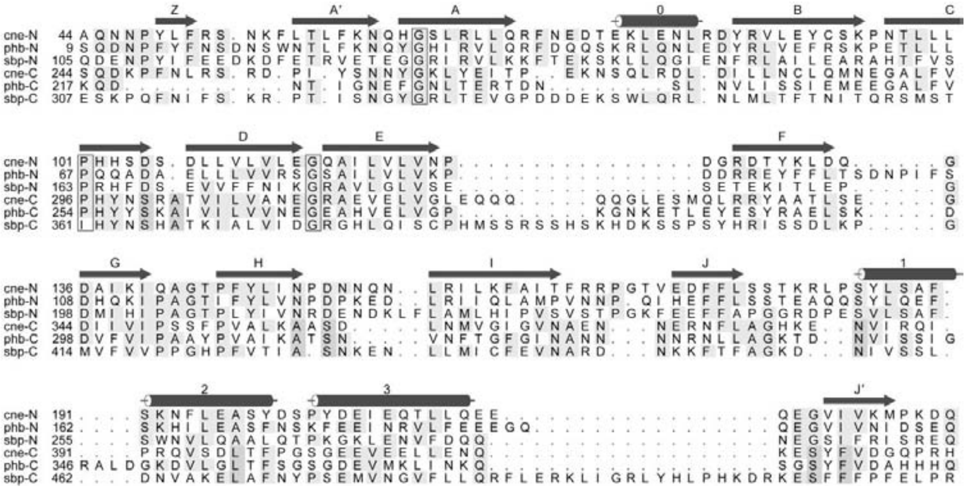
Z
A'
A
0
B
C
cne-N 44 AQNNPYLFRS.NKFLTLFKNQHGSLRLLQRFNEDTEKLENLRDYRVLEYCSKPNTLLL
phb-N 9 SQDNPFYFNSDNSWNTLFKNQYGHIRVLQRFDQQSKRLQNLEDYRLVEFRSKPETLLL
sbp-N 105 QDENPYIFEEDKDFETRVETEGGRIRVLKKFTEKSKLLQGIENFRLAILEARAHTFVS
cne-C 244 SQDKPFNLRS.RD.PI.YSNNYGKLYEITP..EKNSQLRDL.DILLNCLQMNEGALFV
phb-C 217 KQD..........NT.IGNEFGNLTERTDN......SL.NVLISSIEMEEGALFV
sbp-C 307 ESKPQFNIFS.KR.PT.ISNGYGRLTEVGPDDDEKSWLQRL.NLMLTFTNITQRSMST
D
E
F
cne-N 101 PHHSDS.DLLVLVLEGQAILVLVNP...............DGRDTYKLDQ......G
phb-N 67 PQQADA.ELLLVVRSGSAILVLVKP...............DDRREYFFLTSDNPIFS
sbp-N 163 PRHFDS.EVVFFNIKGRAVLGLVSE...............SETEKITLEP......G
cne-C 296 PHYNSRATVILVANEGRAEVELVGLEQQQ....QQGLESMQLRRYAATLSE.....G
phb-C 254 PHYYSKAIVILVVNEGEAHVELVGP..........KGNKETLEYESYRAELSK.....D
sbp-C 361 IHYNSHATKIALVI DGRGHLQISCPHMSSRSSHSKHDKSSPSYHRISSDLKP.....G
G
H
I
J
1
cne-N 136 DAIKIQAGTPFYLINPDNNQN..LRILKFAITFRRPGTVEDFFLSSTKRLPSYLSAF.
phb-N 108 DHQKIPAGTIFYLVNPDPKED..LRIIQLAMPVNNP.QIHEFFLSSTEAQQSYLQEF.
sbp-N 198 DMIHIPAGTPLYIVNRDENDKLFLAMLHIPVSVSTPGKFEEFFAPGGRDPESVLSAF.
cne-C 344 DIIVIPSSFPVALKAASD.....LNMVGIGVNAEN...NERNFLAGHKE..NVIRQI.
phb-C 298 DVFVIPAAYPVAIKATSN.....VNFTGFGINANN...NNRNLLAGKTD..NVISSIG
sbp-C 414 MVFVVPPGHPFVTIA.SNKEN..LLMICFEVNARD...NKKFTFAGKD...NIVSSL.
2
3
J'
cne-N 191 ....SKNFLEASYDSPYDEIEQTLLQEE...............QEGVIVKMPKDQ
phb-N 162 ....SKHILEASFNSKFEEINRVLFEEEGQ.............QEGVIVNIDSEQ
sbp-N 255 ....SWNVLQAALQTPKGKLENVFDQQ................NEGSIFRISREQ
cne-C 391 ....PRQVSDLTFPGSGEEVEELLENQ................KESYFVDGQPRH
phb-C 346 RALDGKDVLGLTFSGSGDEVMKLINKQ................SGSYFVDAHHHQ
sbp-C 462 ....DNVAKELAFNYPSEMVNGVFLLQRFLERKLIGRLYHLPHKDRKESFFFPFELPR

Figure 3
Homology sequence alignment of the duplicated domains. The N- and C-terminal domains of canavalin (cne-N and cne-C), β-phaseolin (phb-N and phb-C) and SBP (sbp-N and sbp-C) are shown juxtaposed. Sequences were aligned with reference to the results of least-squares fitting by the program *O*, in which structurally equivalent residues were identified. The common secondary-structural elements of strands and helices are shown superimposed on the sequences. Helix 4 is not present in N-terminal domains and is therefore omitted. Numbering of phaseolin residues follows that in PDB entry 2phl. This figure was produced using *ALSCRIPT* (Barton, 1993).

Table 2
Crystallographic refinement of canavalin.

Numbers in parentheses are for the highest resolution shell.

Crystal form	Cubic	Hexagonal
Resolution range (Å)	40–2.10 (2.20–2.10)	40–2.00 (2.07–2.00)
Number of reflections [$F > 2\sigma(F)$]	19695 (2109)	30894 (2647)
Completeness (%)	84.0 (72.9)	96.7 (83.5)
R value based on 92% data†	0.175 (0.239)	0.208 (0.333)
$R_{\rm free}$ for 8% test data set†	0.235 (0.291)	0.264 (0.370)
Average B (Å^2)/No. of atoms		
Total	28.3/2868	37.1/2948
Backbone atoms‡	25.8/1376	34.6/1384
Side-chain atoms‡	30.0/1391	37.6/1400
Solvent atoms	40.4/101	53.1/164
R.m.s.d.s		
Ideal bond length (Å)	0.010	0.013
Bond angle (°)	1.569	1.707
Dihedral angle (°)	27.8	28.7
Improper angle (°)	0.729	0.812

† $R = F_{hkl}|F_{\rm obs} - F_{\rm calc}|/\sum_{hkl} F_{\rm obs}$, where $F_{\rm obs}$ and $F_{\rm calc}$ are the structure-factor amplitudes measured from X-ray diffraction and calculated from the model, respectively. ‡ Backbone atoms include N, C$^\alpha$, C and O; other atoms are side-chain atoms.

glycine, non-proline residues with φ, ψ angles lying in the most favored regions, while 8.2% are in the additional allowed regions. The only exceptions are Ser351 and Tyr417, which have similar conformations in both crystal forms, with (φ, ψ) = (73.5, −15.5°) or (76.5, −22.4°) for Ser351 and (64.8,−64.3°) or (72.8, −70.4°) for Tyr417. Ser351 is the third residue of a four-membered type-II-like turn, where a glycine is preferred (Richardson, 1981), and its backbone NH group is also hydrogen bonded to the CO group of Asn299. Tyr417 is similar to Tyr376 in phaseolin (Lawrence *et al.*, 1994). It assumes the second position of a γ-turn in which a very tight hydrogen bond forms between the CO group of the first residue and the NH group of the third (Richardson, 1981). Also, its CO group is hydrogen bonded to the NH of Val295. Both regions had well defined electron densities in both the cubic and the hexagonal crystals.

The previous cubic crystal structure of canavalin to 2.3 Å resolution yielded an R value of 0.193 for 11 968 reflections having $F/\sigma(F) > 3$. For the hexagonal, orthorhombic and rhombohedral crystals, the resolution was 2.6 Å and the R values were 0.197, 0.185 and 0.194, respectively (Ko, Ng & McPherson, 1993; Ko, Ng, Day *et al.*, 1993). In the current study, the resolution was increased to 2.1 and 2.0 Å for the cubic and hexagonal crystals, respectively. The numbers of reflections currently used are 19 695 and 30 894, respectively, which are nearly two and three times as many as previously. With the improved crystallographic R values and $R_{\rm free}$ as well as improved stereochemical properties, the refined models presented here provide a decidedly more precise description of canavalin.

The β-strand Z of both domains now appears well defined and has a similar conformation in both crystals, as shown in Fig. 2. Attempts to include the EF-loop of the C-terminal domain, which contains five consecutive glutamine residues, were unsuccessful. Omission of this probably disordered section resulted in lower $R_{\rm free}$ values. The side chain of Met331 may interact with a hydrophobic pocket formed by Leu320 and Leu333 of the same subunit along with Phe176 and Leu183 of a neighboring subunit, but no density was observed for this residue in either crystal. It may be that the entire loop is disordered, as were the connections between domains and the N- and C-terminal extensions.

3.2. Tertiary structure of canavalin

Consistent with the previously determined crystal structures of canavalin (Ko, Ng & McPherson, 1993; Ko, Ng, Day *et al.*, 1993) and phaseolin (Lawrence *et al.*, 1990, 1994), a monomer subunit is comprised of two very similar domains, as seen in Fig. 4, and each can be subdivided into a core and a loop subdomain. In addition to the strands A′, A–I, J and J′ in the β-barrels, core subdomains also contain a strand Z at the N-terminus and a helix 0 (or 1′) between strands A and B. Helix 0 of the N-terminal domain is a 3_{10} helix, but its equivalent in the C-terminal domain is an α-helix. Loop subdomains contain three helices, numbered 1, 2 and 3. Helix 1 of either an N- or a C-terminal domain was clearly seen to have the 3_{10} conformation, while helices 2 and 3 were α-helical. Several residues were shifted in register with respect to the secondary-structure pattern of the starting model; Fig. 5 is a revised hydrogen-bonding diagram for the polypeptide chain. Strand Z of the N-terminal domain is antiparallel to and forms hydrogen bonds with strand G of the C-terminal domain. Similar interactions were observed between strand Z of the C-terminal domain and strand G of the N-terminal domain.

In addition to the helices and strands shown in Fig. 5, there are numerous short turns in connecting loops. Each N-terminal strand Z is linked by a 3_{10} turn to the first strand A′ of the β-barrel. Specifically, the CO groups of Arg52, Ser53 and Asn250 are hydrogen bonded to the NH groups of Lys55, Phe56 and Ser253, respectively. In the A–B loops, the CO groups of Arg72, Leu83 and Thr269 are hydrogen bonded to the NH groups of Glu75, Tyr86 and Lys272, respectively. Other turns in the loops between β-strands involve the CO groups of Asp133, Gln141, Arg168, Asn289, Ser341, Pro349 and Ala371, which are hydrogen bonded to the NH groups of Asp136, Thr144, Thr171, Ala292, Asp344, Phe352 and Asn374, respectively. In the junctions between core β-barrel subdomains and loop-helical subdomains, several additional hydrogen bonds were also observed involving backbone atoms. Besides those of Ser351 and Tyr417 mentioned above, most notable among these were the NH groups of Thr180, Lys181, Phe378, His382 and Val386 with the CO groups of Leu183, Glu216, Glu384, Glu415 and Phe378, respectively. These probably stabilize β-strands J and J′, as well as the inserted loop subdomains. In addition, there are numerous other hydrogen bonds, salt bridges and van der Waals contacts in the structure involving side chains, which further contribute to the stability of the protein.

3.3. Domain interactions

In the trimeric molecule, the canavalin subunits are disposed about the threefold axis with N-terminal and C-terminal domains related by perpendicular pseudo-dyad axes to create a pseudo-D_3 point-group symmetry (Fig. 1). In the refined cubic crystal structure, domain interfaces bury surface areas of 1409 Å^2 on an N-terminal domain and 1430 Å^2 on a C-terminal domain. From the model in the hexagonal crystal, the buried surface areas are calculated to be 1391 and 1397 Å^2 for the N- and C-terminal domains, respectively. Besides the backbone hydrogen bonds of strands Z and G described above, interdomain interactions also involve many hydrogen bonds between side chains. These include salt bridges between Arg52, Lys79 and Lys139, and Asp344, Glu321 and Asp278, respectively. Hydrophobic interactions also play significant roles in stabilizing domain interfaces. These involve Pro48, Tyr49, Leu50, Phe51, Phe56, Leu70, Phe73, Leu80, Leu83, Tyr86, Val88, Leu109, Leu111, Leu113, Tyr130, Leu132, Ala137, Ile138, Leu160, Phe162 and Ile164, a total of 21 residues in the N-terminal domain, and Pro248, Phe249, Leu251, Leu276, Leu279, Ile281, Leu283, Ile305, Val307, Leu317, Tyr336, Ala338, Leu340, Ile345, Ile346, Val347, Pro349, Phe352, Val365, Ile367 and Val369, a total of 21 residues in the C-terminal domain. Most of these residues are found in the ZA′ABIDG sheets of the core β-barrels.

The trimer of canavalin in Fig. 1 forms by head-to-tail association of subunits; thus, interfaces in the trimeric assembly join the N-terminal and C-terminal domains of adjacent monomers. In the cubic crystal structure, these interfaces bury 2307 Å^2 area on the N-terminal domain of one subunit and 2263 Å^2 on the C-terminal domain of the opposing subunit; in the hexagonal structure, the corresponding areas are 2292 and 2262 Å^2, respectively. Many water-mediated interactions were observed in the intersubunit interface, along with a number of hydrophobic interactions. The residues involved are Leu98, Leu100, Pro101, Val122, Pro124, Pro145, Tyr147, Phe166, Phe175, Leu177, Leu183, Pro184, Tyr186, Leu187, Ala189, Phe190, Phe194, Ala197, Tyr199, Ile206, Leu210, Val220, Met222 and Pro223, a total of 24 residues in the N-terminal domain, and Leu293, Val295,

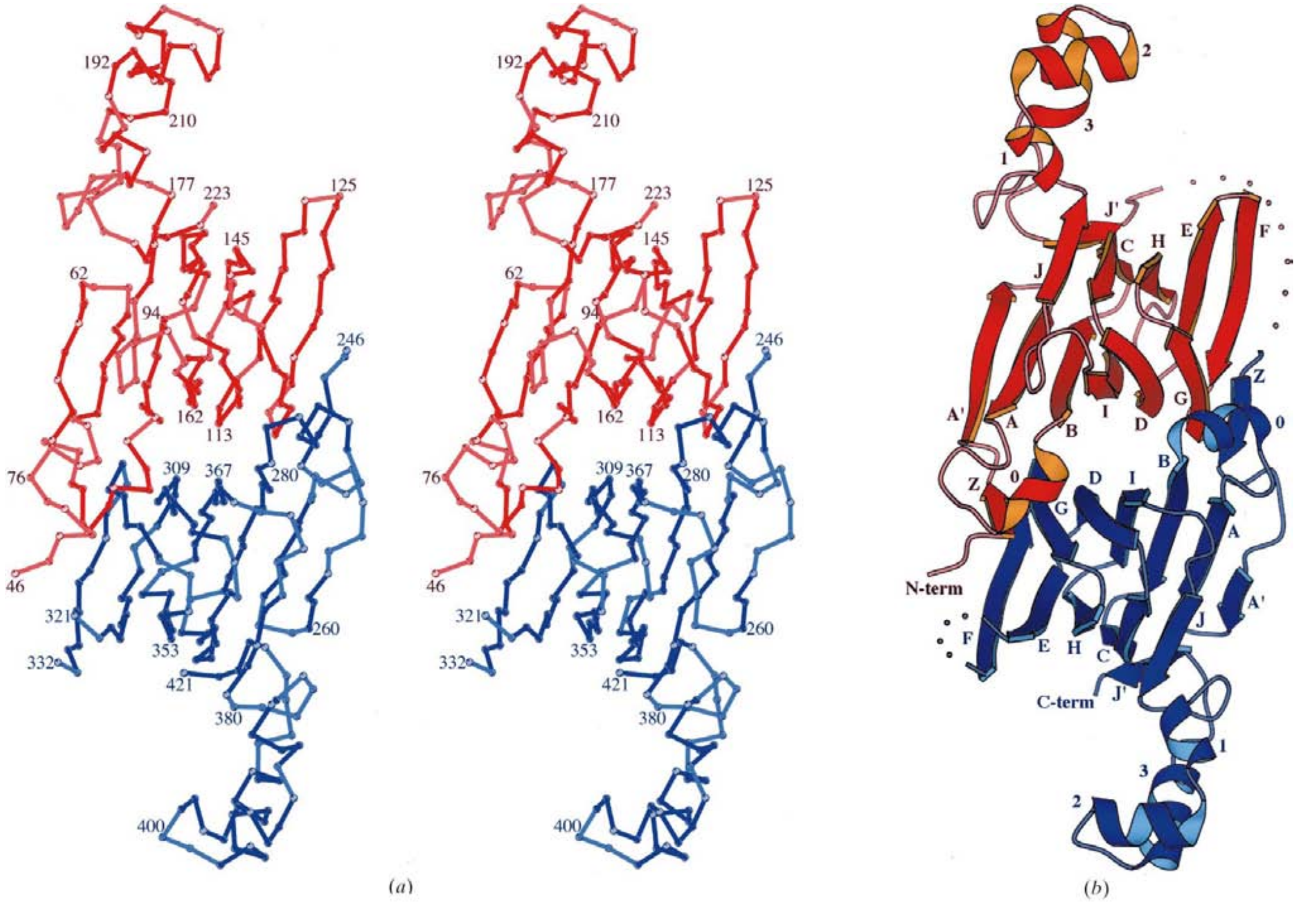

Figure 4
(*a*) Stereo diagram of the C$^\alpha$ tracing of a canavalin subunit based on coordinates from the cubic crystal. Appropriate residue numbers are labelled. (*b*) Ribbon diagram of the canavalin subunit. Strands ZA′ABCDEFGHIJJ′ and helices 0123 are labelled, as well as the N- and C-termini. The unobserved connection between domains and the EF-loop of the C-terminal domain are shown as a string of beads. The view is along the pseudo-dyad of the subunit. The molecular triad lies approximately horizontal in the plane of paper. The N- and C-terminal domains are colored red and blue, respectively. This figure was produced using *MOLSCRIPT* (Kraulis, 1991).

Pro296, Tyr298, Val318, Leu320, Leu333, Phe352, Pro353, Ala355, Leu379, Ala380, Val386, Ile387, Ile390, Pro391, Val394, Leu397, Phe399, Pro400, Val406, Leu409, Leu410 and Val419, a total of 24 residues in the C-terminal domain. Most are in the J/J′CHEF sheets of the core subdomains and in the loop subdomains. An intersubunit interface covers about half again as much surface area as an intrasubunit domain interface, yet the numbers of hydrophobic residues involved are about the same.

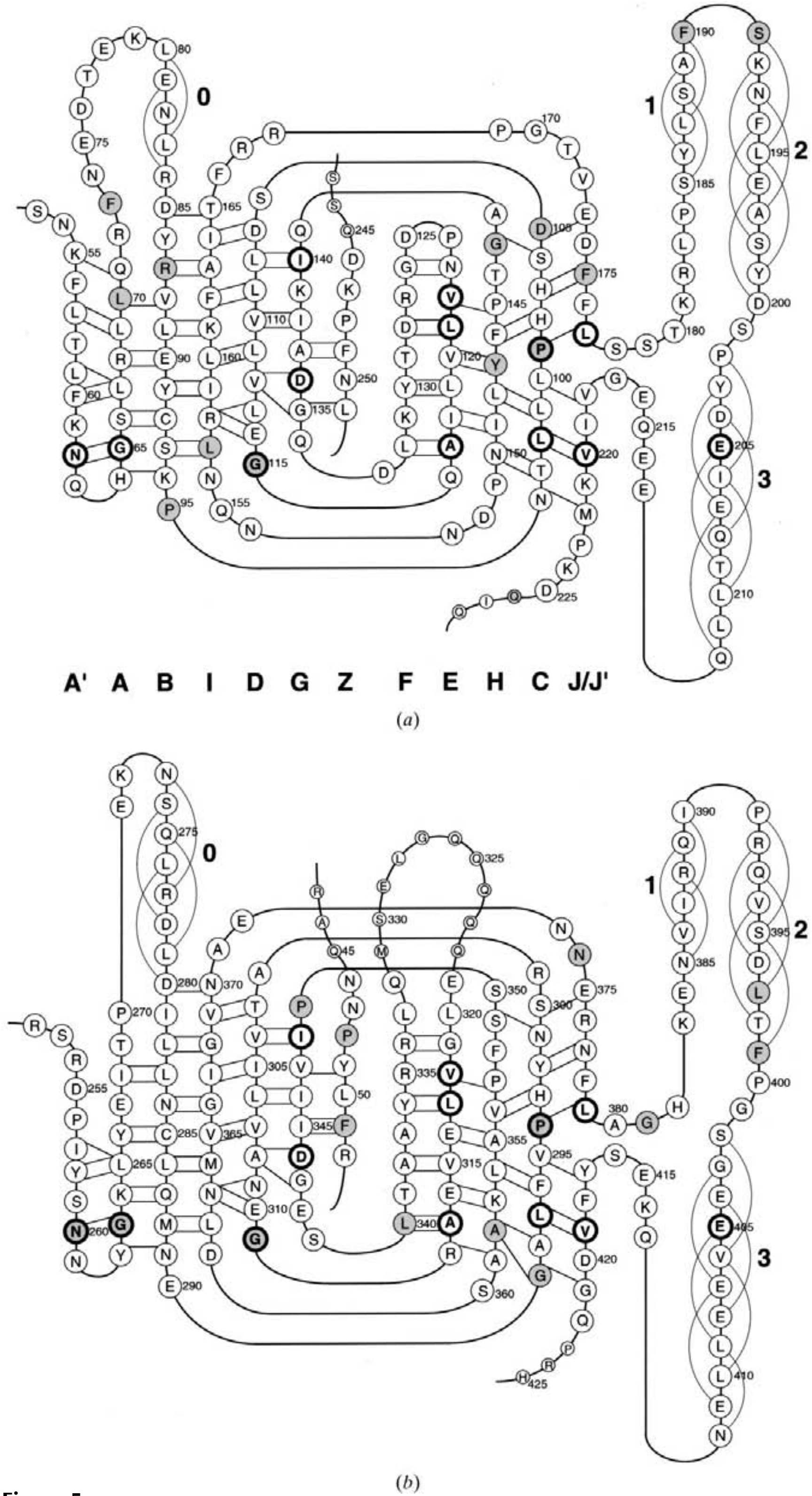

Figure 5
Hydrogen-bond diagram of the canavalin subunit. (*a*) The N-terminal and (*b*) the C-terminal domain. Shaded circles denote the 28 strictly conserved residues among vicilins (Lawrence *et al.*, 1994). Heavy lines emphasize the 13 structurally identical residues in the two domains of canavalin. Residues in smaller circles were not observed in the crystal structures. Residues are numbered in intervals of 5. The secondary-structural elements are labelled as in Fig. 4.

3.4. Crystal packing and solvent molecules

Arrangement of canavalin trimers in four different unit cells has been presented previously (Ko, Ng, Day *et al.*, 1993). The cubic crystal exhibits a molecular packing which is more involved than any other crystal. There, each trimer contacts 12 neighbors and makes two types of intertrimer interactions. The first type involves seven residues on one molecule and nine on another and buries 260 Å^2 surface area on one molecule and 249 Å^2 on a second. The second type of interface involves six residues on one molecule and eight on another and buries 284 and 301 Å^2. These sum to 3282 Å^2 of surface area per trimer of 142 kDa. The residues involved are all hydrophilic, with the exception of Ile118. Its side chain is in contact with that of Asn154 of another trimer.

Packing in the hexagonal crystal is more similar to that in rhombohedral and orthorhombic crystals, where disk-shaped trimers are stacked with rims in contact (Ko, Ng, Day *et al.*, 1993). In the hexagonal crystal, each canavalin trimer contacts six neighbors with a single type of interface. It encompasses at least eight residues on one molecule and 14 on another, burying 411 and 369 Å^2 surface areas, respectively, summing to 2340 Å^2 per trimer. Again, interactions are exclusively hydrophilic. These observations for the cubic and hexagonal canavalin crystals are typical of protein crystal contacts, according to Janin & Rodier (1995), who estimated a mean of 280 Å^2 buried surface area per contact, with a range of 100–600 Å^2 and a total of between 1100 and 4400 Å^2 per molecule.

Of the 101 and 164 water molecules, respectively, in the models of the cubic and the hexagonal crystals, most had well defined density in the final $2F_o - F_c$ maps, where many were seen bound to loops and surface residues. When the two models are superimposed, 36 corresponding waters are within 1.0 Å distance and are considered equivalent. All are involved in multiple hydrogen bonds. Their average temperature factor is 26.6 $Å^2$ in the cubic crystal and 34.6 $Å^2$ in the hexagonal crystal. Of particular interest are Wat507/505,[1] Wat540/527 and Wat550/535, which hydrogen bond to adjacent backbone atoms and form water-mediated β-sheets. Two others, Wat512/510 and Wat545/531, are involved in intersubunit interactions. Inside the lumina of the β-barrels, there are six waters in both the cubic and hexagonal models. Five are equivalent, Wat506/504 and Wat508/506 in the N-terminal barrel and Wat504/502, Wat505/503 and Wat515/501 in the C-terminal barrel. No common waters were found at crystal contacts. In the hexagonal crystal, more than 60 waters (20 per subunit) formed a cluster about the threefold axis of the trimer.

3.5. Structure comparison

The overall root-mean-square deviation (r.m.s.d.) of coordinates between the canavalin models of the cubic and the hexagonal crystals is 0.82 Å for all 2755 pairs of equivalent protein atoms when fitted by *X-PLOR*. For 1372 matched pairs of backbone atoms the r.m.s.d. is 0.33 Å and for 1383 pairs of side-chain atoms it is 1.11 Å. Loop subdomains have higher *B* values, but while the r.m.s.d. is not significantly larger than for core subdomains, it is positively correlated with the distribution of *B* values. The largest difference occurs in the terminal strands and surface loops. In particular, the side-chain atoms of residues Lys181 and Arg182 have an r.m.s.d. of 5 Å and their backbone atoms have an r.m.s.d. of more than 1 Å. *B* values in this region are also high, with a mean of about 50 $Å^2$. This loop connects strand J of the β-barrel and helix 1 of the loop subdomain. It is exposed to the bulk solvent and is probably a secondary cleavage site by trypsin. Regions including residues Asn153 and Ser360 also had large r.m.s.d., but those differences arise from differential involvement in crystal contacts.

In previous models of canavalin, there were several regions having different registers of amino acids in the descriptions of secondary-structure elements (Ko, Ng & McPherson, 1993; Ko, Ng, Day *et al.*, 1993) with respect to the refined phaseolin model (Lawrence *et al.*, 1994). The refined canavalin model presented here is now in precise agreement with the phaseolin model (PDB entry 2phl) with respect to homologous sequence alignment and hydrogen-bonding patterns of strands and helices. When the models of these two vicilins were fitted by least squares using *O* at a cutoff of 3.8 Å, 324–328 pairs of C^α atoms matched with an r.m.s.d. of 1.16–1.32 Å, depending on the particular crystal form of canavalin and the specific subunit of phaseolin. When more stringent matching criteria of 2.0 and 1.5 Å were applied, 289–299 and 251–277 pairs of atoms continued to match with r.m.s.d.s of 0.92–1.10 Å and 0.82–0.88 Å, respectively. If the limit was only 1.0 Å, matching would bias towards the C-terminal domain, but when fitted separately with a cutoff of 1.0 Å, the r.m.s.d. was 0.49–0.52 Å for 112–125 atom pairs in the N-terminal domain and 0.49–0.51 Å for 115–131 atom pairs in the C-terminal domain.

Although the overall structures are very similar, there are some noteworthy differences between canavalin and phaseolin. First, an equivalent for helix 4 in phaseolin was not observed in canavalin. As has been discussed in previous studies, the helix may be lost upon treatment with trypsin. Second, strand Z of the C-terminal domain is present in canavalin but not in phaseolin. Judging from the homology sequence alignment, this is reasonable because the corresponding residues are missing in phaseolin. Therefore, the connection between the two domains must be different. Phaseolin can form a larger assembly of dodecamers, but canavalin has not been observed to do so; this is a possible reflection of the different interdomain linker. Third, helix 0 (or 1′) in the loop between strands A and B of the C-terminal domain of canavalin is also absent in phaseolin, again because of a shorter sequence. There are also, as noted previously, insertions and deletions in other loops connecting strands and helices.

All of the 27 strictly conserved residues (shaded in Fig. 5) in the vicilin subunits as reported by Lawrence *et al.* (1994) occupy exactly the same positions in the secondary-structure diagrams of canavalin and phaseolin. In addition, Ala305 and Tyr376 of phaseolin share similar conformations with Ser351 and Tyr417 of canavalin. There are 69 water molecules in the phaseolin model (PDB entry 2phl). Upon superposition of the protein models, 21 and 19 waters of phaseolin are within 1 Å distance of those of canavalin in the cubic and hexagonal crystals, respectively, while 17 are identical. Among correlated waters, 11 have equivalents in other subunits. It is worth noting that four water molecules within the lumina of barrels in the canavalin crystals were observed in phaseolin, as well as two of the waters mediating hydrogen bonds of β-strands. These include Wat507 (corresponding to Wat504/502 in canavalin), Wat512/525 (Wat505/503), Wat538/561 (Wat506/504), Wat505/537 (Wat508/506), Wat517/547/567 (Wat507/505) and Wat520/523 (Wat540/527), all of which make hydrogen bonds to structurally equivalent residues.

As pointed out in previous analyses, the two vicilin domains are structurally homologous. This is illustrated in Fig. 5, where 13 pairs of identical residues are emphasized. When the N- and C-terminal domains of canavalin are fitted by the least-squares procedure of *O* with a cutoff of 2.5–3.0 Å, all secondary-structure elements can be superimposed, including strand Z and the helices in the loop subdomain. The number of atom pairs range from 123–134 and the r.m.s.d. is 1.17–1.31 Å, depending on the combination of models. With a cutoff below 2.0 Å, only core subdomains, including strands A′–J′ but not Z, can be fitted; with a limit of 1.5 Å, 91 or 92 atom pairs in the barrels superimpose with an r.m.s.d. of 0.63–0.69 Å.

Similar results are obtained by cross-domain superposition of the phaseolin and canavalin models. At a cutoff of 2.5–3.0 Å, 118–137 atom pairs in the entire domain fit with an r.m.s.d. of 1.16–1.50 Å; however, strands Z and J′ of the N-terminal domain of phaseolin do not fit well with those of

[1] The common water residues are denoted by Wat*i*/*j*, in which *i* and *j* are their numbers in the cubic and the hexagonal crystals, respectively.

the C-terminal domain of canavalin. At 1.5 Å cutoff, 84–91 pairs of C$^\alpha$ atoms in the core subdomains superimpose with an r.m.s.d. of 0.66–0.73 Å. The sequence homology between the two vicilin domains and secondary-structure elements are shown in Fig. 3, which suggests that the similarity between subunits of canavalin and phaseolin as a whole is greater than between the N- and C-terminal domains. This is in agreement with the proposed evolution of vicilins by domain duplication (Shutov *et al.*, 1998), which would have occurred prior to the divergence of jack bean and French bean.

3.6. SBP model

The primary sequence of SBP has 20–37% identity and 44–61% similarity with the vicilin-class proteins (Overvoorde *et al.*, 1997). This indicates that the tertiary structure of SBP is closely related to those of canavalin and phaseolin and, as shown below, the potential of SBP subunits to assemble into trimers has also been assessed by model building. The model of SBP we constructed lacks the large N-terminal extension, the short octapeptide connection between the two domains, two insertions in the C-terminal domain between strands E and F and between helix 3 and strand J′, and a C-terminal extension. Overvoorde *et al.* (1997) suggested that the 19-residue insertion in the C-terminal loop subdomain may be important for functional divergence from vicilins. In vicilins, such an insertion has not been observed in the C-terminal domain but, according to the sequence alignment of Lawrence *et al.* (1994), there is an equivalent insertion in the N-terminal domain of, for example, *P. vicae* vicilin. SBP probably follows a different pathway of evolution to vicilins; nevertheless, all other extensions and insertions have equivalents in a number of vicilin sequences. Insertions are located on the surface and can therefore be accommodated upon trimerization of the subunits without disturbing interdomain and intersubunit interactions. The interdomain interface of the SBP subunit buries about 1400 Å^2 surface area on each domain and consists of at least 34 and 32 residues from the N- and C-terminal domains, respectively, of which 22 and 21 have non-polar side chains. Presumably, the two domains of SBP are maintained by hydrophobic interactions as in canavalin. There are also two possible salt bridges, one between Lys176 and Asp375 and the other between Glu190 and Lys309. The intersubunit interface buries 2300 Å^2 on the N-terminal domain of one subunit and 2400 Å^2 on the C-terminal domain of another. At least 49 and 48 residues, respectively, from the two subunits are involved, where 24 and 25 are non-polar, respectively. Again, the intersubunit interactions are similar to canavalin in which loop subdomains are tightly associated. There are four possible salt bridges, Glu169–Arg444, Glu186–Lys447, Lys191–Asp462 and Arg343–Asp445. The last one is near the threefold axis and is between C-terminal domains.

With inclusion of the SBP sequence in the interdomain homology alignment of vicilins, only one strictly invariant residue remains, the glycine between strands D and E. The dihedral angles are $\varphi = 90–100^\circ$ and $\psi = -165 \pm 10^\circ$, which are not allowed except for glycine. This flexible hinge may be required to initiate bending of a nascent β-ribbon of BCDEFGHI, which eventually coils on itself to form a Swiss roll, as in the folding of Greek-key barrels (Richardson, 1981). Two other rigorously conserved residues have only a single exception, Val51 of soybean CG4 conglycinin (Lawrence *et al.*, 1994), whose corresponding residue is a glycine in all other vicilin domains and which is located near the N-terminus of strand A. It has a completely extended conformation with both dihedral angles near 180°. In both domains of canavalin and phaseolin, the backbone N atom is likely to form a hydrogen bond with the side chain OD1 of an asparagine at the end of strand A′, although this Asn is not strictly conserved. The other highly conserved residue is Ile361 of SBP (Overvoorde *et al.*, 1997), which is a proline in all other vicilin domains. The proline has dihedral angles $\varphi = 60^\circ$ and $\psi = 130–140^\circ$ and interacts with hydrophobic residues on helices 1 and 2 of the counter subunit upon trimerization. Whether replacing a Pro with an Ile would have a significant effect on the trimer structure is unclear, but it may endow SBP with a more flexible hinge for effective sugar transport or, as suggested by Overvoorde *et al.* (1997), may coordinate with the function of the 19-residue insertion.

The authors thank Professor Hanna S. Yuan for access to her computing facilities and software at the Institute of Molecular Biology, Academia Sinica. This research was supported by NASA and DOE.

References

Barton, G. L. (1993). *Protein Eng.* **6**, 37–40.

Baumlein, H., Braun, H., Kakhovskaya, I. A. & Shutov, A. D. (1995). *J. Mol. Evol.* **41**, 1070–1075.

Brünger, A. T. (1992*a*). *X-PLOR: A System for X-ray Crystallography and NMR.* New Haven, CT: Yale University Press.

Brünger, A. T. (1992*b*). *Nature (London)*, **355**, 472–474.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Day, J. & McPherson, A. (1992). *Protein Sci.* **1**, 1254–1268.

Derbyshire, E., Wright, D. J. & Boulter, D. (1976). *Phytochemistry*, **15**, 3–24.

Dyer, J. M., Nelson, J. W. & Murai, N. (1995). *J. Protein Chem.* **14**, 665–678.

Gane, P. J., Dunwell, J. M. & Warwicker, J. (1998). *J. Mol. Evol.* **46**, 448–493.

Gibbs, P. E. M., Strongin, K. B. & McPherson, A. (1989). *Mol. Biol. Evol.* **6**, 614–623.

Grimes, H. D., Overvoorde, P. J., Ripp, K., Franceschi, V. R. & Hitz, W. D. (1992). *Plant Cell*, **4**, 1561–1574.

Janin, J. & Rodier, F. (1995). *Proteins*, **23**, 580–587.

Jiang, J. S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.

Jones, T. A. (1982). *Computational Crystallography*, edited by D. Sayre, pp. 303–317. Oxford: Clarendon Press..

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Ko, T.-P., Ng, J. D., Day, J., Greenwood, A. & McPherson, A. (1993). *Acta Cryst.* D**49**, 478–489.

Ko, T.-P., Ng, J. D. & McPherson, A. (1993). *Plant Physiol.* **101**, 729–744.

Koszelak, S., Day, J., Leja, C., Cudney, R. & McPherson, A. (1995). *Biophys. J.* **69**, 13–19.

Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Lawrence, M. C., Izard, T., Beuchat, M., Blagrove, R. J. & Colman, P. M. (1994). *J. Mol. Biol.* **238**, 748–776.

Lawrence, M. C., Suzuki, E., Varghese, J. N., Davis, P. C., van Donkelaar, A., Tulloch, P. A. & Colman, P. M. (1990). *EMBO J.* **9**, 9–15.

Luzzati, P. V. (1952). *Acta Cryst.* **5**, 802–810.

McPherson, A. & Spencer, R. (1975). *Arch. Biochem. Biophys.* **169**, 650–661.

Ng, J. D., Ko, T. P. & McPherson, A. (1993). *Plant Physiol.* **101**, 713–728.

Overvoorde, P. J., Chao, W. S. & Grimes, H. D. (1997). *J. Biol. Chem.* **272**, 15898–15904.

Overvoorde, P. J., Frommer, W. B. & Grimes, H. D. (1996). *Plant Cell*, **8**, 271–280.

Richardson, J. S. (1981). *Adv. Protein Chem.* **34**, 167–339.

Shin, D. S., Compadre, C. M., Maleki, S. J., Kopper, R. A., Sampson, H., Huang, S. K., Burks, A. W. & Bannon, G. A. (1998). *J. Biol. Chem.* **273**, 13753–13759.

Shutov, A. D., Braun, H., Chesnokov, Y. V. & Baumlein, H. (1998). *Eur. J. Biochem.* **252**, 79–89.

Shutov, A. D., Kakhovskaya, I. A., Braun, H., Baumlein, H. & Muntz, K. (1995). *J. Mol. Evol.* **4**1, 1057–1069.

Smith, S. C., Johnson, S., Andrews, J. & McPherson, A. (1982). *Plant Physiol.* **70**, 1199–1209.

Sumner, J. B. & Howell, S. F. (1936). *J. Biol. Chem.* **113**, 607–610.

Warwicker, J. & O'Connor, J. (1995). *Protein Eng.* **8**, 1243–1251.