

Crystal Structure of Hydrophobic Protein from Soybean; a Member of a New Cysteine-rich Family

Franck Baud^{1,2}, Eva Pebay-Peyroula^{1,2,†}, Claudine Cohen-Addad^{1,3,‡}, Shoji Odani⁴ and Mogens S. Lehmann^{1,‡}

¹*Institut Laue-Langevin, Avenue des Martyrs
F 38042, Grenoble, France*

²*Université Joseph Fourier, BP 53 X, F 38041 Grenoble Cedex, France*

³*Laboratoire de Biologie Structurale, CNRS, URA 1333, CENG
F 38041, Grenoble, France*

⁴*Department of Biology, Faculty of Science, Niigata University
Niigata 950-21, Japan*

(Received 1 October 1992; accepted 11 February 1993)

X-ray diffraction methods have been used to determine the structure of the 8.3 kDa hydrophobic protein from soybean and to refine the atomic co-ordinates to a crystallographic *R*-factor of 18.7% at 1.8 Å resolution. The molecule is a four-helix bundle, which together with the connecting loops and a twisted β -strand form a spiral. The surface contains 70% apolar atoms, and the crystal packing is dominated by hydrophobic interactions, producing a two-dimensional sheet of protein molecules. Most of the 59 water molecules located are involved in hydrophilic contacts and their structural organization does not seem to be affected by the high hydrophobicity of the molecule. From the protein fold it appears that three of the four disulphide bridges are important for keeping the amino and carboxyl-terminal segments in place in the native form, while the central part of the molecule is stabilized by many hydrophobic interactions. Although the protein function is not known, a number of possibilities can be excluded on experimental grounds and by comparison with other members of the family.

Keywords: hydrophobic; helix-bundle; cysteine-rich; soybean

1. Introduction

The title compound, which was identified in a 60% ethanolic extract of soybean (*Glycine max*), has an unusual amino acid composition with a very high content of hydrophobic residues. This prompted a full characterization (Odani *et al.*, 1987), including sequencing (Table I) and a search for comparable proteins. At the time no convincing homology was found, except for a similarity with residues 66 to 144 of rat prolactin, and it was therefore proposed that the hydrophobic protein from soybean (HPS§) would belong to an entirely new protein family.

Shortly after, during a study of the conserved domains near the C terminus of wheat storage proteins (Henrissat *et al.*, 1988) using hydrophobic cluster analysis (Gaboriaud *et al.*, 1987), the HPS was found to be homologous with these, as was a number of non-specific lipid-transfer proteins. Consequently, it was suggested that the conserved domains probably reflect a common ancestor, while the unique properties were associated with the variable domains.

Moreover, some of the proteins, namely the HPS, a number of non-specific lipid-transfer proteins (Takishima *et al.*, 1988; Tchang *et al.*, 1988; Breu *et al.*, 1989; Désormeaux *et al.*, 1992), and a 115 residue hydrophobic protein from wheat (Marion, personal communication), showed a common pattern of eight cysteine residues, ...C....C....CC....CXC....C...., forming four disulphide bridges. Because of the small size of the proteins, typically around 100 residues, and the tight net of disulphide bridges, it is therefore

† Present address: Institut de Biologie Structurale, Avenue des Martyrs, F 38027, Grenoble, France.

‡ Author to whom all correspondence should be addressed.

§ Abbreviations used: HPS, hydrophobic protein from soybean; r.m.s., root-mean-square.

Table 1
Primary sequences

	1	10	20	30
(a)	ALITRPSCPD	LSICLNILGG	SLGTVDDCCA	
	LIGGLGDI EA	IVCLCIQLRA	LGILNLRNL	
	QLILNSCGRS	YPSNATCPRT		
(b)	AIXCGQVXSX	LXPCLXYAXG	XGXGPSAXCC	
	XGVRXLXAA	XTXTXDRRXA	ACNCSLKXAA	
	XRGIXGLNXG	XAASIPXKCG	VXIPYTISXS	
	TDCSRVNN			
(a)	ALITRPSCP-	D-LSICLN	ILGGS LGTV D	
(c)	-----AGKCP R	DALKLGVCAD	VLNLVHN VVI	
(d)	-----AVRTCPI	DTLKL NACVD	VLSGLIHLVI	
	D-----CCA	LIGGLGDI EA	IVCLCIQLRA	
	GSPPTLPCCS	LLEGLVNLEA	AVCLCTAIKA	
	GQEARSKCCP	LVQGVADLDA	ALCLCTTIRA	
	---LGI-LNLN	RNLQLILNSC	GRSYPSNATC	
	NILGKNLNL P	I ALSLVLN NC	GKQVPNGFEC	
	RLLNINIYLP	I ALNLLI-TC	GKHAPSGFQC	
	PRT			
	T			
	PPLYD			

(a) Hydrophobic protein from soybean (Odani *et al.*, 1987). (b) Consensus sequence for phospholipid transfer protein (Désormeaux *et al.*, 1992). (c) Hydrophobic part of protein DC 2-15 from carrot (Aleith & Richter, 1990). (d) Hydrophobic part of proline-rich protein from maize (José-Estanyol *et al.*, 1992).

believed that these proteins have similar three-dimensional structures, possibly with a new fold.

In the case of the HPS the biological function is not known. Nevertheless, it seemed worthwhile to determine the structure, partly to compare with the structure of a phospholipid-transfer protein from wheat (Désormeaux *et al.*, 1992) which is presently being determined by NMR analysis (Simorre *et al.*, 1991), and partly to learn more about the function of the protein by comparison with the various protein structures as they become available.

HPS is insoluble in water at neutral or high pH, but crystals can be grown from an acetic acid solution by increasing pH. Because of the high content of hydrophobic residues and its small size, 80 amino acids, a fair number of hydrophobic residues might be in contact with the solvent. These interactions are expected to lead to a well-ordered water structure due to the reduced number of hydrogen bond configurations available near a hydrophobic surface. A good example of this is part of the water arrangement in the crystal structure of Crambin (Teeter & Whitlow, 1986), where water molecules connect in pentagons to form a cap around Leu18. As a result of the water ordering the entropy will decrease, possibly increasing the total free energy. Consequently, only few cases of such arrangements have been found, and it is therefore important to search for further examples of these interactions.

2. Materials and Methods

(a) Crystallization and data collection

Methods of extraction, purification and crystallization have been discussed elsewhere (Odani *et al.*, 1987;

Lehmann *et al.*, 1989). The hanging drop method was used, the reservoir typically contained 800 ml of 1.25 M ammonium acetate buffer (pH 4.7), 100 μ l of 2 M NaCl solution and 100 μ l of dioxane, while the drop was a 1:1 mixture of the reservoir and a 1.25 M acetic acid solution of the protein (30 mg/ml). Later it was found that larger crystals were obtained if grown in the presence of 0.1% (v/v) octyl- β -glucoside. Typically the size of these crystals were 0.3 mm $\times$ 1.0 mm $\times$ 0.5 mm along the *a*, *b* and *c* axis of the unit cell. They were isomorphous with crystals grown without octyl- β -glucoside, with unit cell *a*=52.1 Å, *b*=43.3 Å, *c*=28.6 Å and space group *P*2₁2₁2₁, and were therefore used for a large part of the heavy-atom search and for the final data collections. As more than half the drops would normally contain a large number of thin, plate-like crystals, attempts were made to improve this, and keeping the hanging drops at about 38°C for a day at the start of crystallization was found to produce fewer and larger crystals, possibly due to a retarded generation of nuclei for the crystal growth.

A precession camera was used for the identification of possible heavy-atom derivatives, while intensity data were recorded on FAST diffractometers (Enraf-Nonius, Delft) employing the MADNES control software (Messerschmidt & Pflugrath, 1987) and peak integration based on profile fitting (Kabsch, 1988). All measurements were done at ambient temperature. The main X-ray source was an Enraf-Nonius FR 571 rotating-anode generator with a Cu target and equipped with a graphite monochromator. Here all derivative and native data used for structure solution were recorded to a resolution of 2.5 Å with a completeness of 88%. For the derivative data *R*_{sym} was in the range from 5 to 9% while *R*_{sym} for the native data was 4.0% for 2163 unique reflections. In addition one set of native FAST diffractometer synchrotron data were recorded at station 9.6 of the Daresbury Laboratory using a wavelength of 0.89 Å. In this case reflections were seen up to a resolution of 1.3 Å. The *R*_{sym} was 3.8% for data limited to 1.8 Å, where the complete-

ness was 78% giving 4943 unique reflections. In the last resolution shell from 1.88 Å to 1.80 Å the completeness was 72%. Beyond this there were 1225 reflections with an intensity of more than 3 times the standard deviation. However, this part of the set was very incomplete, and the data have not yet been included in the analysis. For averaging of the data, as well as for the further analysis involving calculations of Patterson functions and electron densities, refinement of heavy-atom positions and phasing of the reflections, programs from the CCP4 system (Daresbury Laboratory) were used.

Neutron small angle scattering measurements were also carried out to determine the radius of gyration for HPS when dissolved in 1.25 M acetic acid. The D11 instrument of the Institut Laue-Langevin was used with a wavelength of 7.0 Å and a sample to detector distance of 1.4 m. The scattering curve showed no sign of aggregation, and a Guinier plot gave a radius of 12.7 ± 1.0 Å, corresponding to a solution of monomers.

(b) Heavy-atom search and structure solution

A first search for heavy-atom derivatives done by soaking, using crystals grown according to the original recipe, gave no positive results, most probably due to the presence of ammonium ions in the mother liquor. In a second attempt the ammonium acetate in the reservoir was therefore replaced by sodium acetate, and eventually two useful derivatives were found, namely $\text{UO}_2(\text{NO}_3)_2$ and K_2PtCl_4 . They were both single site derivatives, and the essential parameters are given in Table 2. A number of other heavy-atom complexes such as K_2OsCl_6 and KAuCl_4 also gave changes in the intensities, but at the same time lead to large contractions along the *a*-direction of the crystal and very noisy Patterson functions. They were thus abandoned.

The best derivative, $\text{UO}_2(\text{NO}_3)_2$, was used to fix the origin, and the choice of enantiomorph was done by phasing with one derivative employing the anomalous contribution. The correct solution was then chosen to be

the one that gave a single strong peak in the difference Fourier of the other heavy-atom derivative. This choice was later confirmed from the right-handedness of the α -helices.

An electron density based on the combined phases from the two derivatives including the anomalous contribution did not allow an unambiguous location of the protein molecule, but showed continuous bands of molecules in the *b* direction which formed rather tightly packed layers of molecules in the *b*-*c* plane interspaced by an approximately 5 Å thick layer of fluid density corresponding to the water. Attempts of phase improvements using solvent flattening did not improve the clarity of the maps, and in addition only about a quarter of the density could easily be used for tracing of the main-chain. A reason for this was undoubtedly the relatively large number of disulphide bridges, which gave several regions with a complex density pattern, and the sulphur atoms were therefore located directly (Lehmann & Pebay-Peyroula, 1992) using a method originally proposed by Kraut (Straus & Kraut, 1968). For this the anomalous differences in the native data were phased with the experimental phases minus 90°, and the result was a map that clearly showed the location of the four sulphide bridges. Combining this with the original maps it was then quite easy to identify the atoms belonging to the pair Cys28-Cys29 and the triplet Cys43-Leu44-Cys45. This reduced the assignment of the rest of the sulphur atoms to basically two possibilities, and finally from the chain direction one solution could be identified.

Further improvement of the map was then done by the introduction of one more derivative, $\text{Pb}(\text{CH}_3\text{COO})_2$ and a second set of data for $\text{UO}_2(\text{NO}_3)_2$ with two sites for UO_2^{2-} (Table 2). By a combination of density fitting employing versions E4.3 and ESV4.4 of the program FRODO (Jones, 1978), analysis of difference Fourier maps and use of the known primary sequence (Odani *et al.*, 1987) it was then possible to assign about 80% of the atoms in the protein and to start a structure refinement.

Table 2
Parameters for the heavy-atoms used in the phasing procedure

Derivative	$\text{UO}_2(\text{NO}_3)_2^a$	$\text{K}_2\text{PtCl}_4^b$	$\text{Pb}(\text{CH}_3\text{COO})_2^c$	$\text{UO}_2(\text{NO}_3)_2^a$
Unique refl.	1990	2013	1957	1787
Completeness	81%	81%	79%	72%
Resolution (Å)	10.0/2.5	10.0/2.5	7.0/2.5	7.0/2.5
R_{merge}^d	0.097	0.053	0.053	0.051
Isomorphous difference ^e	0.20	0.20	0.22	0.25
Heavy-atoms	1	1	1	2
<i>x</i> (Å)	0.51	7.10	41.07	0.41 7.67
<i>y</i> (Å)	-15.76	7.14	4.85	-15.73 6.67
<i>z</i> (Å)	7.64	9.60	1.10	7.57 8.95
occ. ^f	0.43	0.53	0.51	0.39 0.11
R_{Cullis}^g	0.67	0.78	0.62	0.74
Phasing power ^h	1.93	0.98	2.24	1.49

Statistics are given for the final run of PHARE from the CCP4 system (Daresbury Laboratory). The mean figure of merit was 0.85.

^aSoaked for 1.5 h in 60 mM $\text{UO}_2(\text{NO}_3)_2$ (pH 4).

^bSoaked for 3 h in 20 mM K_2PtCl_4 (pH 4).

^cSoaked for 2 h in 500 mM $\text{Pb}(\text{C}_2\text{H}_3\text{O}_2)_2$.

^d $R_{\text{merge}} = \sum |I_j(h) - \langle I(h) \rangle| / \sum I_j(h)$, summed over all measured reflections.

^eIsomorphous difference = $\sum |F_{\text{nat}} - F_{\text{der}}| / \sum (F_{\text{nat}} + F_{\text{der}}) / 2$.

^fOccupation factor from refinement with the known structure.

^g $R_{\text{Cullis}} = \sum |F_{\text{der calc}} - F_{\text{der obs}}| / \sum |F_{\text{der obs}} - F_{\text{nat}}|$, where F_{der} and F_{nat} are the derivative and native structure factor amplitudes and the sum is over all the centric reflections.

^hPhasing power = σ_{der} / E , where σ_{der} is the r.m.s. heavy-atom structure factor amplitude and E is the r.m.s. lack of closure.

(c) Refinement and location of the water

The program X-PLOR, based on combined minimization of agreement factor and energy by simulated annealing (Brünger, 1991), was employed for the refinement and the data used were the set of reflections recorded with synchrotron radiation, where the low-order data beyond 10.0 Å were excluded. The *R*-value at the start was 40.0%, and after 10 cycles of refinement it fell to 32.2% using data to 2.8 Å resolution. At this stage the structure was re-evaluated using difference Fourier techniques, and it was noted that several residues in the α -helix made up by residues 56 to 65 were badly fitted. All these residues were shifted one step towards the start of the chain, modifying the preceding loop, after which another 30 cycles of refinement were done with the data to 2.5 Å. At this stage 7 residues were described as Ala. They were later converted to their true denomination as the difference maps improved.

During the following refinement, where various schemes for cooling and annealing were studied, the agreement factor stayed constant around 32%. A first attempt was therefore made to introduce water molecules. The 41 largest peaks from the maps used in the original chain tracing were included in the refinement as water molecules, employing at this stage only data to 2.8 Å.

After 8 cycles the *R*-value had dropped to 18.8%, at which stage the water molecules were removed. A further 40 cycles of refinement were then done including only the protein atoms while the resolution was slowly increased to 1.8 Å, and during this phase individual isotropic temperature factors were introduced together with an overall anisotropic temperature parameter. Smaller modifications were done based on the Fourier and difference Fourier maps at intervals of about 8 cycles, and finally the *R*-factor reached a value of 25.2% for all 578 non-hydrogen atoms. During this process omit maps were used to check the chain tracing in the loop region, typically by excluding 10 residues at a time. Then water molecules were again included, this time based on the final difference Fourier map with the inclusion criteria that the peak should be higher than 3 times the r.m.s. density in the map and less than 5 Å from the protein molecule. In all 59 water molecules were included, and after 17 more cycles of refinement the final agreement factor was 18.7% at 1.8 Å resolution, including 4647 structure amplitudes, *F*, with $F > 3\sigma(F)$. In order to see how much influence the addition of the water had on the agreement factor, this was calculated for the protein atoms only, and gave 24.4%.

The final model shows clear electron density for all main-chain peptide units except at Gly34 and at the 2 termini, namely residues Ala1-Leu2-Ile3-Thr4 and Arg79-Thr80. These residues were included anyway and placed to best cover the highest part of the density in these regions, but their high temperature parameters with a mean *B* value of 44 Å² clearly shows their bad definition. From the sequence analysis (Odani *et al.*, 1987) it is known that 1/3 of the molecules lack the first 2 residues, but because of the bad definition of these residues, attempts to refine the site occupation factors did not lead to any improvement in the overall agreement factors. For the side-chains the electron density is well-defined except for Gln47, Ile53, Arg69 and Tyr71.

At the final stage the mean temperature parameter for the protein atoms, excluding the badly defined residues quoted above, was 18.5 Å², while the corresponding value for the water was 36.0 Å², ranging from 13.5 Å² to 66.9 Å².

The atomic co-ordinates have been deposited with the Brookhaven Protein Data Bank (reference 1HYP).

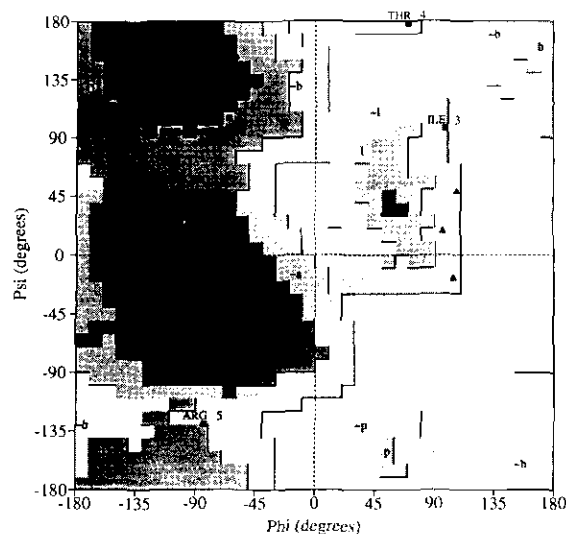


Figure 1. Ramachandran plot for the (ϕ , ψ) torsional angles. The program PROCHECK (R. A. Laskowski, M. W. MacArthur, D. D. Moss & J. M. Thornton, personal communication) was used, and the different regions are defined by Morris *et al.* (1990). PROCHECK was also used to check the stereochemistry, and no unusual features were found, except in the first 5 residues, which are very badly ordered. Filled triangles, Gly.

3. Results

(a) The final model and the helix arrangement

The final refined molecule contains 578 atoms, not including hydrogen atoms, and in addition 59 water molecules were found. For the protein molecule a number of tests were made to ensure that its form and structure matches with standard values. The first of these tests was therefore to ensure that the protein follows the power law $A_s = 6.3 M^{0.73}$ (Miller *et al.*, 1987) relating the solvent-accessible surface A_s (Lee & Richards, 1971) to the molecular weight M for globular proteins. The X-PLOR system was used for the calculation of A_s , and fixing the coefficient to 6.3 the exponent value found was 0.73, in agreement with the above quoted value.

The Ramachandran plot (Ramachandran *et al.*, 1966) is shown in Figure 1, and here again the observations fall within the expected boundaries. To check the agreement with expected bond distances and angles, the refined molecule was compared to an idealized molecule. This gave an r.m.s. error of 0.027 Å for the bond distances and 4.2° for the angles. These are relatively high values for the resolution and data quality, so test refinements using X-PLOR with less weight assigned to the X-ray data were carried out, giving an agreement factor of 20.0% for an r.m.s. error of 0.015 Å. The maximum average co-ordinate errors are given in the Luzzati plot (Luzzati, 1952) shown in Figure 2.

A plot of the molecular structure given in Figure 3 shows that the molecule is formed by four helices and one twisted β -strand, connected by loops and tightly knit by the four disulphide bridges. The secondary structure elements are given in Table 3

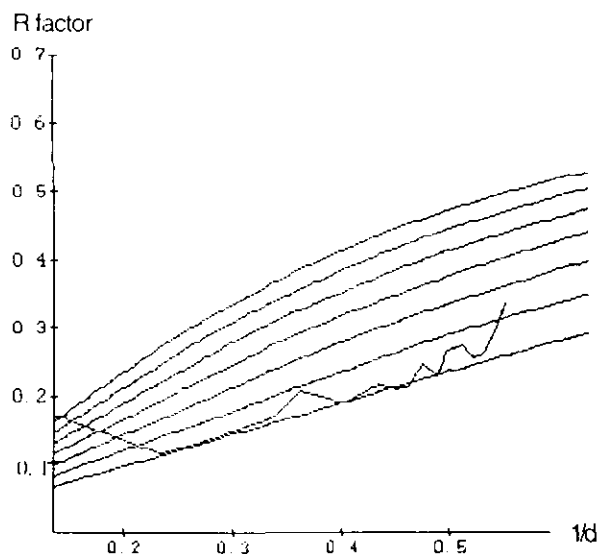


Figure 2. Luzzati plot, showing the final co-ordinate error. The bottom curve corresponds to 0.2 and the interval is 0.05 Å, so the overall error is around 0.25 Å.

together with the geometry of the four disulphide bridges. The protein is best described as a helix bundle with H1/H2 and H3/H4 forming two anti-parallel helix sets. The angles between the helices are given in Table 3. The connecting loop pieces are short, and it is not possible for the two sets of helices to align, thus H1/H2 is twisted with respect to H3/H4. Indeed, the two pairs of helices are nearly orthogonal to each other. The total folding pattern is very straightforward, the helices and connecting loops together creating a spiral. At present we have not identified any similar fold, and the pattern found here is clearly one of the simplest ever observed.

Figure 3(b) shows a view of the molecule along the central helix H3, which consists mainly of hydrophobic components. This is connected with the amphipathic helices H2 and H4 *via* loops L2 and L3, and in addition H2 and H4 are bound together *via* disulphide bridge Cys29–Cys67. H1 is held in place by Cys14–Cys28 at the H2/H4 connection and bound to H3 *via* Cys8–Cys43, while S1 is bound to

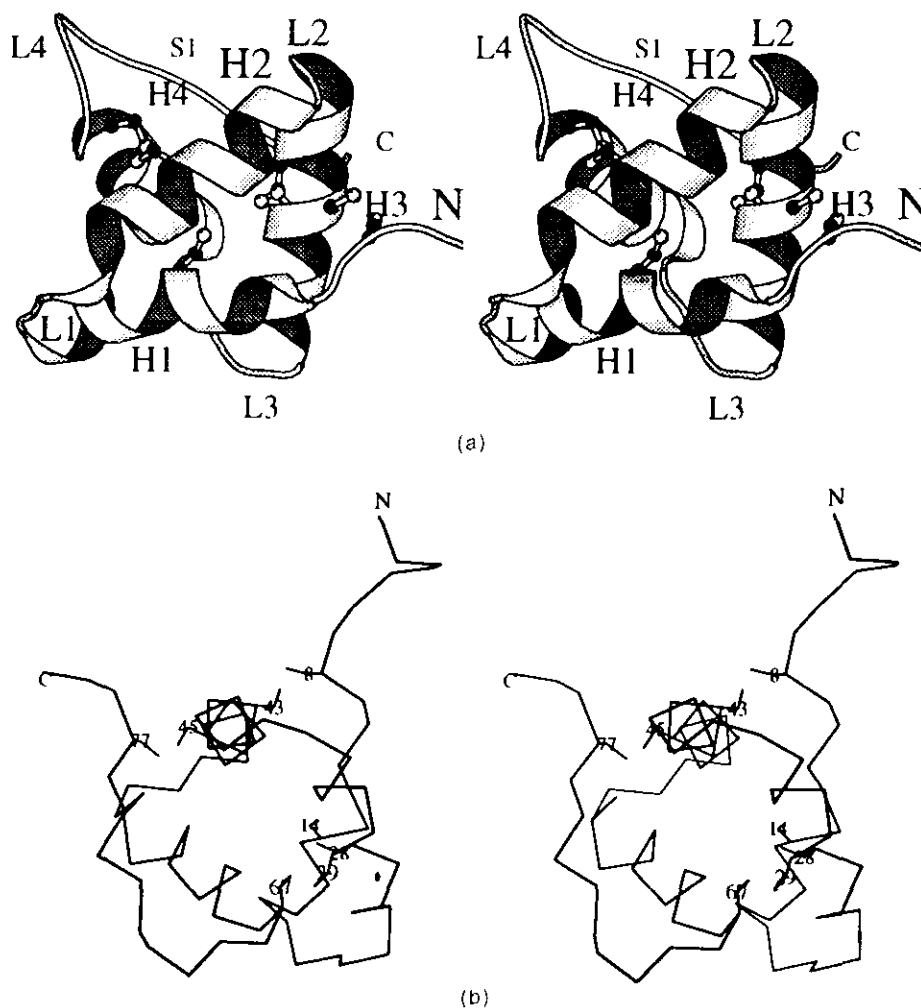


Figure 3. (a) Ribbon plot of the protein molecule. The locations of the sulphur atoms are marked with small spheres, and the name of the helices are indicated. (b) $C\alpha$ plot viewed along the axis of helix H3. An approximate 2-fold axis can be found in the vertical direction through the middle of the molecule. The program MolScript (Kraulis, 1991) was used.

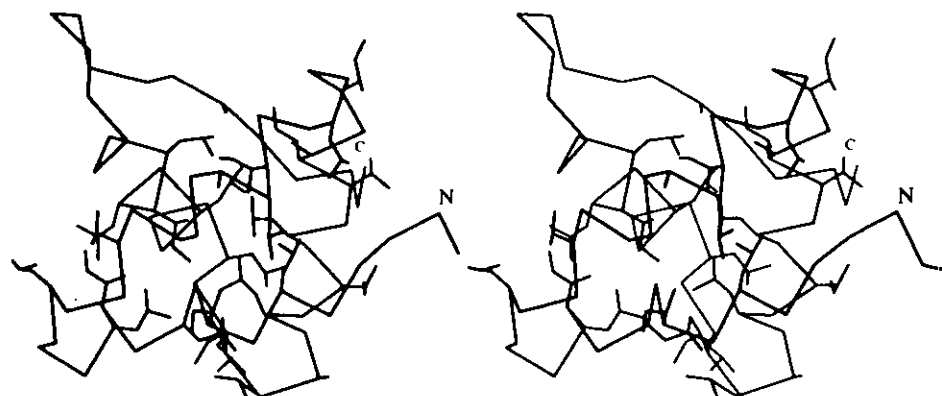


Figure 4. Main-chain with the hydrophobic residues Ala, Leu, Ile, Val and Pro added.

H3 via Cys77–Cys45. The disulphide bridges are thus arranged at two ends of the molecule, namely around the midpoint of H3 and in the knot binding H2 to H4. The amino and carboxyl terminals are both found near one of the disulphide knots, but they are not bound to each other. It is worth noting, that the protein has partial 2-fold symmetry with the 2-fold axis found orthogonal to H3, going through its midpoint and through the disulphide bridge Cys29–Cys67. If we assume that the disulphide bridges are symmetrically distributed as well, then we note that one is lacking, namely the one binding S1 to H4, which therefore gives more flex-

ibility to this region than to the region of H1 and H2.

(b) Hydrophobic surfaces and the crystal packing

The hydrophobic residues are shown in Figure 4, while Figure 5 gives the correlation diagram between all the hydrophobic groups. Clearly visible is the antiparallel arrangement in the two pairs of helices H1/H2 and H3/H4. Because the two sets are twisted away from each other, H1 also has contact with H3 and H4, and this is obvious from smaller patches in the diagram.

There are two main narrow hydrophobic clefts. The larger one of these is found in the gap formed by helix pair H1/H2 being twisted away from pair

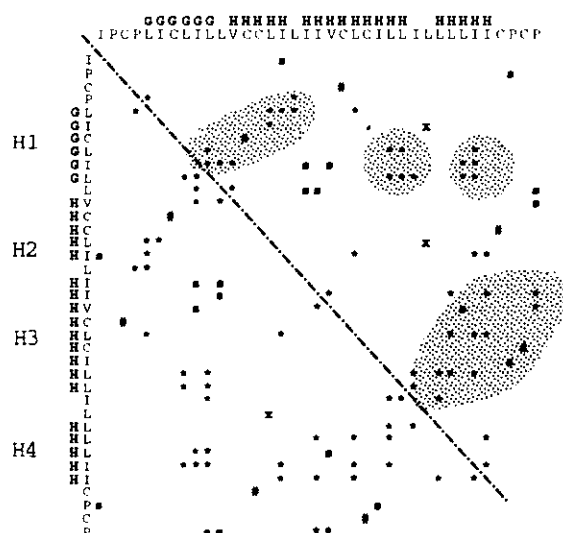


Figure 5. Connection diagram for the main hydrophobic interactions. Only the hydrophobic groups are listed, and the locations of the helices are indicated. Two residues are in contact if a distance between 2 side-chain carbon atoms is less than 5 Å. G stands for the 3_{10} -helix, H for α -helices, an asterisk represents intramolecular contacts, s for contacts between molecules related by a the screw-axis along *b* combined with a translation along *c*, x for intermolecular contacts along *c* and # indicates the disulphide bridges. Major regions of intramolecular contacts are dotted for the upper half of the symmetric diagram. Elongated patches with positive slope show antiparallel interactions. For identification of a given residue number please refer to Table 1.

Table 3

Secondary structure elements, torsional angles (deg.) and C^α – C^α distance (Å) in the disulphide bonds, and angles (deg.) between the helices

A. Names of secondary structure elements						
Residue number	Name				Type	
1-9					Undefined	
10-19	H1				3 ₁₀ -Helix	
20-23	L1				Loop	
24-34	H2				α -Helix	
35-37	L2				Loop	
38-49	H3				α -Helix	
50-55	L3				Loop	
56-65	H4				α -Helix	
66-71	L4				Loop	
72-78	S1				β -Strand	
79-80					Undefined	
B. Torsion angles and distance in S-S bridges						
S-S bridge	χ^1	χ^2	χ^3	χ^4	χ^5	C $^\alpha$ -C $^\alpha$
Cys8-Cys43	-62	-63	-77	-70	-62	5.45
Cys14-Cys28	-69	-69	95	95	160	4.82
Cys29-Cys67	-55	-81	-85	-52	-74	5.49
Cys45-Cys77	-71	-89	-86	-63	-67	5.70
C. Angles between helices						
Helix	H2		H3		H4	
H1	-137		-65		100	
H2			127		-101	
H3					126	

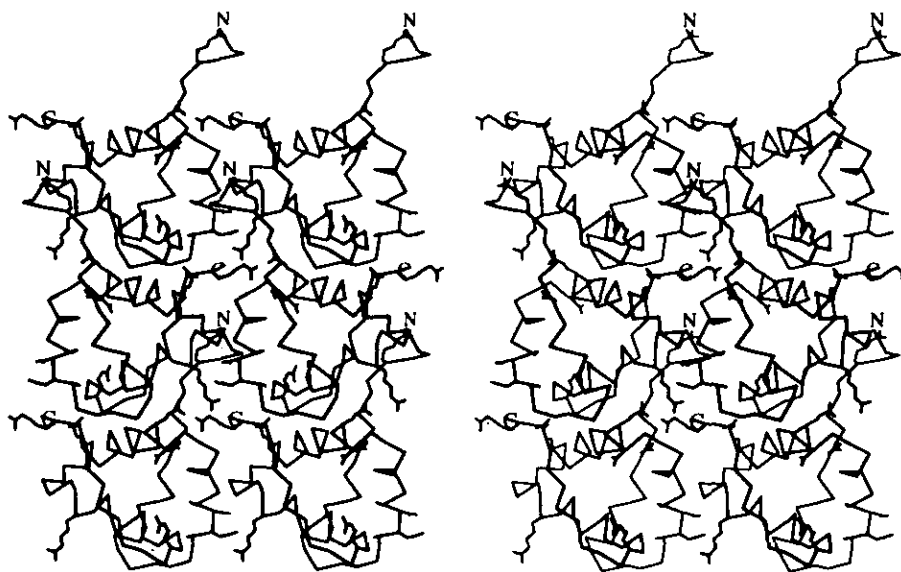


Figure 6. Packing arrangement in the *b*-*c* plane with *b* pointing up and *c* to the right in the Figure. The C^{α} atoms are marked as well as the Arg, Asp and Glu residues. In the *b*-direction neighbouring molecules are related *via* a screw axis along *b* combined with translations along *c*, and in the *c*-direction neighbouring molecules are related by the *c*-translation.

H3/H4. It involves Leu15, Ile17, Leu18, Leu22, Val25, Leu44, Leu48, Leu51, Ile53, Leu62 and Ile63. A smaller hydrophobic region is found on the other side of the molecule and is made by Ile38, Leu41, Leu42, Ile46 and Pro78, and finally a number of other hydrophobic groups are found on the protein surface. The accessible surface was analysed using the method of Lee & Richards (1971) using a water search sphere with a radius of 1.4 Å. We found the contacts to be distributed as 70%

apolar, 18% polar and 12% charged surface, so obviously the protein deserves its name.

Despite this high degree of hydrophobicity there is, however, relatively little contact between the solvent region and the apolar residues. This is due to the packing arrangement shown in Figure 6. Along the *b*-direction the hydrophobic residues of helix H3 found at the outside of the molecule are covered up by similar parts of loop L1/helix H4, and in the *c*-direction H2 connects with L3 *via* a few hydrophobic contacts. Further contacts are obtained as the N terminus covers the smaller of the pockets listed above, and the final surface produced by this packing is therefore much like the one found in much less hydrophobic molecules.

This leads to some very tightly bound sheets in the *b*-*c* plane, which may explain why the crystals grow as very elongated plates along the *b*-axis.

(c) Hydrogen bonding and charge distribution

Figure 7 shows the arrangement of hydrogen bonds in the main-chain forming the different helix arrangements, where H2, H3 and H4 are α -helices and H1 is a 3_{10} -helix. As mentioned above the helices are held together by disulphide bridges, but beyond this some hydrogen bonds are also involved in stabilizing intramolecular interactions. These are listed in Table 4 together with the hydrogen bonds forming intermolecular bonds, the two sets being about equal in size. Another important group of hydrogen bonds are with the solvent, and they are discussed below.

In addition there is a group of potential hydrogen bond donors or receptors, which show no structural contacts. All of these were studied in detail to ensure that they were in contact with the disordered solvent region and thus potentially could be

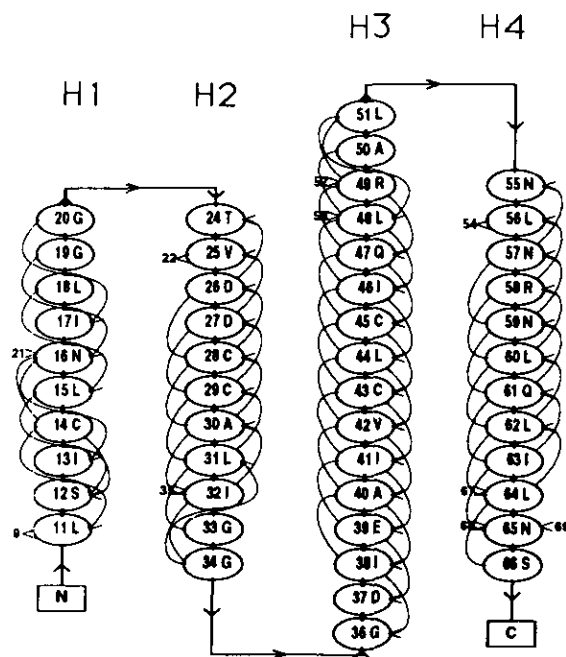


Figure 7. Hydrogen bond connections in the main-chain, drawn with the program HERA (Hutchinson & Thornton, 1990).

Table 4

Intra- and intermolecular hydrogen bonds in A

A. Intramolecular		Distance
Leu11-N	Gln47-O ^{e1}	2.98
Ile13-O	Thr24-O ^{γ1}	2.78
Asn16-N ^{δ2}	Thr24-O ^{γ1}	3.21
Arg49-N ^{η2}	Pro78-O	3.07
Asn57-O ^{δ1}	Cys77-N	2.96
Gln61-N ^{ε2}	Ala75-O	3.01
Asn65-O ^{δ1}	Ser73-N	2.86
B. Intermolecular		Distance
Ala1-N	Asn57-N ^{δ2}	2.51
Ser7-N	Gln61-O ^{ε1}	2.84
Asp26-O ^{δ1}	Asn55-N ^{δ2}	2.91
Asp26-O ^{δ2}	Arg58-N ^{η2}	3.18
Gly34-O	Arg79-N ^{η1}	2.99
Glu39-O ^{ε1}	Arg58-N ^{η1}	3.00
Glu39-O ^{ε2}	Arg58-N ^ε	2.90
Leu51-O	Asn74-N ^{δ2}	3.00

involved in hydrogen bonding; no anomalies were found.

The charge distribution shown in Figure 6 is asymmetric, with positive and negative charges grouped on each side of the molecule. The negative charges from the carboxyl groups are located in the regions near the helices H1 and H2, while the positive guanidinium groups are near H4 and S1. In the packing of the molecules these two sides touch, leading to a concentration of charges from two Arg, two Asp and the Glu in the sheet discussed above. This can also be seen in Figure 6, and the charge interactions with solvent or with themselves are clearly an important factor for the stabilization of the molecular arrangement.

The resulting two Arg and two Asp residues are spread regularly over the surface of the protein sheets.

(d) Arrangement of water molecules

It was expected that the structure analysis would reveal some ordered water structure induced by hydrophobic regions on the protein surface, but as a whole these regions are covered by the molecular packing arrangement. Indeed, the water structure was found to be quite limited, only involving a few larger clusters.

All together the unit cell contains 37% solvent ($V_m = 1.9 \text{ \AA}^3/\text{dalton}$), and although there might be molecules other than water present, such as for example the dioxane molecule used in the crystal growth, none of these were observable in the difference Fourier maps. The expected number of water molecules is therefore around 200 and only approximately 30% of these can be located as ordered water. This is probably due to the crystal packing where the protein molecule forms sheets interspaced by relatively thick layers of water. This leaves a

Table 5

List of groups of water molecules connected by hydrogen bonds

No connection to other water molecules: 8, 18, 19, 25, 32, 35, 40, 46, 47, 54, 58
 Pairs of water molecules: 1, 26; 6, 45; 30, 44; 34, 36
 Groups of three: 3, 48, 50; 9, 10, 11; 22, 41, 56
 Groups of four: 5, 12, 13, 37
 Groups of five: 2, 14, 23, 24, 31; 15, 16, 38, 52, 55; 20, 33, 42, 43, 59
 Groups of six: 4, 27, 28, 29, 51, 53; 7, 17, 21, 39, 49, 57

Water molecules are numbered from 1 to 59, and the number given is the order of inclusion in the list as the refinement progressed. Water molecules were assumed to be connected if the O-O distance was less than 3.1 Å.

smaller than normal surface for water contacts and a larger volume where liquid water can be present.

The water molecules can be roughly distributed into a first and second layer. In the first layer a fairly large number of water molecules are bound directly to the main-chain, namely 19 as donors in the hydrogen bond and only 12 as acceptors, showing a clear preference for the first kind. This is in agreement with expectations, as the steric freedom near the main-chain is limited. It will thus need much less free space to place a hydrogen at a distance of 1.8 to 2.0 Å from the surface rather than an oxygen atom.

A similar number of water molecules has contact with the side-chains, namely 27, distributed with 14 donors and 13 acceptors, and in this case it is obviously equally easy for the water molecules to place themselves as donor and acceptor. For the side-chain there are around 60 locations if all places are filled, and 22 of these are already engaged in inter- and intramolecular hydrogen bonds. The observed number of contacts with water molecules therefore seems reasonable, showing that a good part of the available molecular surface is covered with ordered water.

A few of the water molecules in the first layer are involved in stabilizing the protein molecule. This is notably the case for waters 9, 10 and 11, which are placed inside loop L1 and waters 1 and 26, which support loop L4. In addition water 18 connects the start of helix H1 with helix H3, and water 19 makes a bridge between helix H4 and strand S1. Finally one water molecule, 47, is caught in a hydrophobic region between two molecules.

The second layer, consisting of water molecules which are only bound to other water molecules, holds 33 water-water bonds. Together with the other molecules it forms a number of straight or bifurcated chains, which are summarized in Table 5. There is only one ring, a pentagon involving water molecules from the first group of six connected water molecules in the list, and one of the oxygen atoms in this ring is the carbonyl oxygen from Leu64.

Altogether, the water structure is therefore very similar to the one found around much less hydrophobic molecules.

4. Discussion

In small, disulphide-rich proteins one might expect these structural elements to be important for maintaining the correct fold. An inspection of the HPS clearly shows this to be true for the amino segment of the molecule, which starts with a 3_{10} -helix, H1, held together by relatively weak hydrogen bonds (Kabsch & Sander, 1983), followed by a loop containing three Gly residues, and bound to the rest of the molecule *via* three hydrogen bonds. The main stabilizing factor is therefore undoubtedly the two disulphide bridges Cys8–Cys43 and Cys14–Cys28.

The carboxyl segment of the chain is also fixed near the terminus by a disulphide bridge, Cys45–Cys77, and here again the preceding strand, S1, is connected to the core of the protein by three hydrogen bonds. The stability of this end is therefore also mainly determined by the disulphide bridge. Collectively this ensures that the two termini are held near together, and it is tempting to speculate that this might be important for the as yet unknown function of the protein.

The other end of strand S1 is bound by Cys29–Cys67, and this bridge also participates in forming the spiral that is the core of the molecule, made principally by H2, H3 and H4 and the connecting loops. In addition, these three helices are also held together by many contacts in the very hydrophobic interior. The disulphide bridge in question does, therefore, seem to be less essential for the overall form of the molecule. Indeed, one case has been observed in an α -amylase inhibitor from Indian finger millet (Campos & Richardson, 1984), which has 50% homology with the non-specific lipid-transfer protein from wheat (Svensson *et al.*, 1986; Breu *et al.*, 1989), where the Cys corresponding to Cys29 has been mutated to Gln. Even though there is at present only this single example, and although normally both Cys residues would be mutated (Thornton, 1981; Désormeaux *et al.*, 1992), disappearance of this bond might not completely destabilize the molecule.

The structural changes occurring during a reduction of the disulphide bonds have not yet been studied for HPS, but recently the non-specific phospholipid-transfer protein from wheat was analysed in detail (Désormeaux *et al.*, 1992). It was found by infrared techniques that 41% of the structure was helix, 19% β -sheet, and 40% other forms, and this is in reasonable agreement with our observations from the three-dimensional structure. When the disulphide bridges were cleaved, the extended conformations were increased by 15% at the expense of the helix structure. From the above considerations we would expect that H1, the first part of H2 and S1 would come loose, releasing about 15% of the helix structure, so this is in numerical agreement with spectroscopy. Moreover, from structure predictions on the wheat protein it was suggested that the changes would occur near the amino terminus, again in agreement with expect-

tations from the crystal structure analysis. There is thus reason to believe that Cys8–Cys43 and Cys14–Cys28 holding H1 as well as Cys45–Cys77 holding S1 are essential for keeping the protein in its native form, while Cys29–Cys67 perhaps can be left out.

The members of this putative new family can be divided into two groups. On one side is the soluble, hydrophilic proteins, which have been assigned functions such as enzyme inhibition or phospholipid-transfer, and on the other side are the insoluble, hydrophobic proteins, which are fewer, and which until now have been less well characterized functionally. The reason for this is probably the difficulty in solubilizing the compounds. For HPS a number of tests have been done (Odani *et al.*, 1987), which include inhibition of a number of common enzymes as well as haemagglutinating and haemolytic activity. The protein has also been tested for anti-fungal activity (Hilgenfeld, personal communication), but none of the tests showed significant results, so there is still no indication of the biological activity.

Because of its small size and its tightly knit structure it is clearly inviting to believe that HPS, as so many other small disulphide-rich proteins, has some blocking or disrupting function, i.e. it has toxic activity. Because of its low solubility at physiological pH it is, however, somewhat difficult to imagine how it gets transported to its point of action. Another possibility is of course to think of HPS as a transfer protein, similar to the function of many of the soluble proteins in the family, but it has, for technical reasons, not been possible to test this at low pH. However, in the recent study of wheat lipid-transfer protein (Désormeaux *et al.*, 1992) a consensus sequence for a number of these proteins has been identified (see Table 1), and it was proposed that invariant Ser and Thr residues found in the middle of the sequence between Cys–Cys and Cys–X–Cys might be involved in hydrogen bonding to the lipids. In HPS there are no Ser and Thr in the same stretch, which is also shorter, and except for an Asp and a Glu completely hydrophobic, so if the above suggestion is correct this rules out the HPS as a transporter of lipids.

Another possible region for substrate binding is the cleft formed between the planes of H2/L2/H3 and H4/L4/S1. The bottom of this rather wide cleft is covered by hydrophobic residues, while the residues likely to interact with a substrate are found at the middle of the two edges, namely Asp37 on L2 and Arg69 and Tyr71 on L4. Comparing this with the consensus sequence for the transfer proteins (Désormeaux *et al.*, 1992) gives an Asp and a Tyr at similar locations, but there is no Arg near the Tyr, and the Asp is flanked on one side by two Arg residues. There are thus too many differences to assume similar functions. Finally, the argument of insolubility at neutral pH should again be invoked, and although the cleft might be a candidate for binding, the only way of function would be that the

complex between HPS and whatever it transfers renders HPS soluble.

As mentioned before, the HPS is soluble at low pH, and this might well be related to the protonation of the five acid groups. Comparison of the radius of gyration of 12.7 ± 1.0 Å from neutron small-angle scattering, with the value of 12.4 Å calculated from the crystal structure, indicates that this is probably not accompanied by any major rearrangement of the molecule. This is unlikely anyway considering its rigidity due to the many disulphide bridges. The solubilization would therefore mainly relate to a neutralization of the acid groups, and this would then be done by whatever molecule is transported. Such a molecule would have to contain a number of positive groups. The candidates are few, if any, and this probably rules out the molecule as a transfer protein of a smaller molecule.

The above arguments have not included the possibility that the hydrophobic protein is part of a larger complex. During extraction the pH at the chromatography stages was 6.5 and the protein should therefore have been insoluble. Nevertheless, it was the pass-through fraction of the DEAE-cellulose column that gave the HPS (Odani *et al.*, 1987), which crystallized out shortly after. This might very well be due to the fact that during the previous steps the protein was part of a larger assembly. The ragged amino terminus, where 1/3 of the two first residues are lacking, might hint at such a possibility.

Further credence to such an hypothesis can be given by a recent identification of an embryo-specific gene of a hybrid proline-rich protein in carrot (Aleith & Richter, 1990) and maize (José-Estanyol *et al.*, 1992), where the C terminus section has high homology to HPS with the eight Cys residues conserved (see Table 1). For the carrot protein the genes were temporarily transcribed from the second to a bit beyond the 20th day of somatic embryogenesis, and it was pointed out, that the carboxy terminus had resemblances to membrane attachment segments, although no specific function was suggested. For the maize protein it was proposed that the full complex could contribute temporarily to the stability and defence of the developing embryo. Until now we have found no experimental evidence that the soybean protein has a defence function, but this could clearly be changed if it was part of a larger complex.

Therefore, at present we cannot, from the structure, propose the nature of the protein function. A number of common possibilities can be excluded for HPS, but obviously the skeleton of disulphide bridges common to this new family can be used for a considerable number of different functions.

We are very indebted to Dr D. Marion, INRA, Nantes, for many useful discussions on protein function, to Dr R. Hilgenfeld, Hoechst AG, Frankfurt, for the test of antifungal activity. We also thank Professor C.-I. Brändén for helpful comments on protein folding and Drs

S. Cusack and L. C. Sieker for advice, ideas and criticism all along this analysis. Measurements at Daresbury were supported by a grant from the EEC large scale facilities program, and we thank Dr M. Papiz for help and advice. One of us (F.B.) gratefully acknowledges a doctoral grant from the French Ministry for Research and Industry.

References

- Aleith, F. & Richter, G. (1990). Gene expression during induction of somatic embryogenesis in carrot cell suspensions. *Planta*, **183**, 17–24.
- Breu, V., Guerbette, F., Kader, J.-C., Gamini Kannangara, C., Svensson, B. & von Wettstein-Knowles, P. (1989). A 10 kD barley basic protein transfers phosphatidylcholine from liposomes to mitochondria. *Carlsberg Res. Commun.* **54**, 81–84.
- Brünger, A. T. (1991). Simulated annealing in crystallography. *Annu. Rev. Phys. Chem.* **42**, 197–223.
- Campos, F. A. P. & Richardson, M. (1984). The complete amino acid sequence of α -amylase inhibitor I-2 from seeds of ragi (Indian finger millet, *Eleusine coracana* Gaertn.). *FEBS letters*, **167**, 221–225.
- Désormeaux, A., Blochet, J.-E., Pézolet, M. & Marion, D. (1992). Amino acid sequence of a non-specific wheat phospholipid transfer protein and its conformation as revealed by infrared and Raman spectroscopy. Role of disulfide bridges and phospholipids in the stabilization of the α -helix structure. *Biochim. Biophys. Acta*, **1121**, 137–152.
- Gaboriaud, C., Bissery, V., Benchetrit, T. & Mornon, J. P. (1987). Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Letters*, **224**, 149–155.
- Henrissat, B., Popineau, Y. & Kader, J.-C. (1988). Hydrophobic-cluster analysis of plant protein sequences. *Biochem. J.* **225**, 901–905.
- Hutchinson, E. G. & Thornton, J. M. (1990). HERA—A program to draw schematic diagrams of protein secondary structures. *Proteins: Struct. Funct. Genet.* **8**, 203–212.
- Jones, T. A. (1978). A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11**, 268–272.
- José-Estanyol, M., Ruiz-Avila, L. & Puigdomènech, P. (1992). A maize embryo-specific gene encodes a proline-rich and hydrophobic protein. *Plant Cell*, **4**, 413–423.
- Kabsch, W. (1988). Evaluation of single crystal X-ray diffraction data from a position-sensitive detector. *J. Appl. Crystallogr.* **21**, 916–924.
- Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
- Kraulis, P. J. (1991). MolScript: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.
- Lee, B. K. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
- Lehmann, M. S. & Pebay-Peyroula, E. (1992). Location of the sulphur atoms from the phased anomalous map using native protein data can be very helpful in tracing the peptide chain. *Acta Crystallogr. sect B*, **48**, 115–116.
- Lehmann, M. S., Pebay-Peyroula, E., Cohen-Addad, C. & Odani, S. (1989). Crystallographic data for soybean hydrophobic protein. *J. Mol. Biol.* **210**, 235–236.

- Luzzati, V. (1952). Traitement statistique des erreurs dans la détermination des structures cristallines. *Acta Crystallogr. sect. A*, **5**, 802-810.
- Messerschmidt, A. & Pflugrath, J. W. (1987). Crystal orientation and X-ray pattern prediction routines for area-detector diffractometer systems in macromolecular crystallography. *J. Appl. Crystallogr.* **20**, 306-315.
- Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.
- Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). Stereochemical quality of protein structure co-ordinates. *Proteins: Struct. Funct. Genet.* **12**, 345-364.
- Oda i, S., Takehiko, K., Ono, T., Seto, Y. & Tanaka, T. (1987). Soybean hydrophobic protein. Isolation, partial characterization and the complete primary structure. *Eur. J. Biochem.* **162**, 485-491.
- Ramachandran, G. N., Venkatachalam, C. M. & Krimm, S. (1966). Stereochemical criteria for polypeptide and protein conformations. *Biophys. J.* **6**, 849-872.
- Simorre, J.-P., Caille, A., Marion, D., Marion, D. & Ptak, M. (1991). Two and three-dimensional ^1H NMR studies of a wheat phospholipid transfer protein. Sequential resonance assignments and secondary structure. *Biochemistry*, **30**, 11600-11608.
- Strahs, G. & Kraut, J. (1968). Low-resolution electron-density and anomalous-scattering-density maps of *Chromaticum* high-potential iron protein. *J. Mol. Biol.* **35**, 503-512.
- Svensson, B., Asano, K., Jonassen, I., Poulsen, F. M., Mundy, J. & Svendsen, I. (1986). A 10 kD barley seed protein homologous with an α -amylase inhibitor from indian finger millet. *Carlsberg Res. Commun.* **51**, 493-500.
- Takishima, K., Watanabe, S., Yamada, M., Suga, T. & Mamiya, G. (1988). Amino acid sequences of two nonspecific lipid-transfer proteins from germinated castor bean. *Eur. J. Biochem.* **177**, 241-249.
- Tchang, F., This, P., Stiefel, V., Arondel, V., Morch, M.-D., Pages, M., Puigdomènech, P., Grellet, F., Delseny, M., Bouillon, P., Huet, J.-C., Guerbette, F., Beauvais-Cante, F., Duranton, H., Pernollet, J.-C. & Kader, J.-C. (1988). Phospholipid transfer portein: full-length cDNA and amino acid sequence in maize. *J. Biol. Chem.* **263**, 16849-16855.
- Teeter, M. M. & Whitlow, M. D. (1986). Hydrogen bonding in the high resolution structure of the protein crambin. *Trans. Amer. Cryst. Ass.* **22**, 75-89.
- Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261-287.