Acta Cryst. (1996). D52, 730–737

Structure of Porcine Pancreatic Spasmolytic Polypeptide at 1.95 Å Resolution

THOMAS N. PETERSEN, ANETTE HENRIKSEN AND MICHAEL GAJHEDE*

Centre for Crystallographic Studies, Universitetsparken 5, Department of Chemistry, University of Copenhagen, Denmark. E-mail: michael@xray.ki.ku.dk

(Received 2 June 1995; accepted 30 January 1996)

Abstract

The structure of a trigonal crystal form of porcine pancreatic spasmolytic polypeptide (PSP) has been solved by molecular replacement and refined to 1.95 Å resolution. Three heavy-atom derivatives were prepared, giving unbiased phase information, which was used in the model building of the protein molecules. The final conventional R value is 19.8% with the inclusion of 183 water molecules. PSP crystallizes as a dimer in space group $P3_121$ with a non-crystallographic twofold axis relating the monomers. The monomer consists of two very similar domains each composed of three loop regions. Two clefts are found in the monomer, one in each domain, that are proposed as possible substrate-binding sites. Important interactions have been identified in the proposed substrate-binding sites, where conserved water molecules probably mimic the hydrophilic positions of the substrate. The estimated cleft size is $9 \times 9 \times 12$ Å. Analysis of the charge distribution within the clefts, by an electrostatic potential calculation, shows the clefts to be essentially non-charged.

1. Introduction

Porcine pancreatic spasmolytic polypeptide (PSP) belongs to the trefoil family of loop peptides, which includes at least eight different proteins/peptides originating from mammalian sources and six different frog-skin proteins/peptides (Thim, 1994). A common characteristic for all members of the family is a domain of 38 or 39 residues which includes six disulfide-bridged cysteines in the configuration 1–5, 2–4 and 3–6 when numbering the cysteines from the N-terminal end. The domain is also termed the *P*-domain motif. A large number of conserved residues are found within the characteristic trefoil domain: 14 identical and seven homologous, as seen in Fig. 1.

PSP is a single-chain protein consisting of 106 residues with a pyrrolidone carboxylic acid (PCA) as the N-terminal residue. The protein is made up of two trefoil domains, connected by a stretch of ten amino-acid residues, Lys48–Glu57. There is an additional disulfide bridge outside the trefoil domains, Cys6–Cys104, that keeps the N-terminal residues in contact with the C-terminal tail. The molecular weight of PSP is 11 725 Da, as calculated from the primary sequence.

PSP was originally isolated as a by-product from the production of porcine insulin, and early on it was found that PSP is secreted into the pancreatic juice upon stimulation with enzymes (Thim, Jørgensen & Jørgensen, 1982; Rasmussen *et al.*, 1993). It has further been found that PSP is highly resistant to digestion by intraluminal proteases within the gastrointestinal tract (Jørgensen, Thim & Jacobsen, 1982). The proteins have been suggested to act as naturally occurring healing factors for peptic ulcers (Wright, Pike & Elia, 1990), inflammatory bowel disease and other diseases of the gastrointestinal tract

```
Peptide  Domain              Sequence number

                           10        20        30        40
                            .         .         .         .

PSP      1         RCSRQDPKNRVNCGFPGITSDQCFTSGCCFDSQVPGVPWCFK
PSP      2         EC-VMQVSARKNCGYPGISPEDCAARNCCFSDTIPEVPWCFF
hSP      1         QCSRLSPHNRTNCGFPGITSDQCFDNGCCFDSSVTGVPWCFH
hSP      2         QC-VMEVSDRRNCGYPGISPEECASRKČCFSNFIFEVPWCFF
mSP      1         RCSRLTPHNRKNCGFPGITSEQCFDLGCCFDSSVAGVPWCFH
mSP      2         QC-VMEVSARKNCGYPGISPEDCASRNCCFSNLIFEVPWCFF
pS2      1         TC-TVAPRERQNCGFPGVTPSQCANKGCCFDDTVRGVPWCFY
hITF     1         QC-AVPAKDRVDCGYPHVTPKECNNRGCCFDSRIPGVPWCFK
rITF     1         QC-MVPANVRVDCGYPTVTSEQCNNRGCCFDSSIPNVPWCFK

                              N  F  ITS          D  V
Consensus sequence -C-------R--CG-P------C----CCF------VPWCF-
                              D  Y  VSP          S  I
```

Fig. 1. A sequence alignment of some of the mammalian proteins/peptides belonging to the trefoil family. The conserved residues are shown as the consensus sequence in the bottom row. Partly conserved residues are represented above and below the consensus sequence line. PSP: porcine pancreatic spasmolytic polypeptide (Thim, Thomsen, Christensen & Jørgensen, 1985); hSP: human spasmolytic polypeptide (Thim *et al.*, 1993); mSP: mouse spasmolytic polypeptide (Tomasetto *et al.*, 1990); pS2: human breast cancer associated peptide (Tomasetto *et al.*, 1990); hITF: human intestinal trefoil factor (Podolsky *et al.*,1993); and rITF: rat intestinal trefoil factor (Suemori, Lynch-Devaney & Podolsky, 1991).

(Wright, Poulsom *et al.*, 1990; Wright *et al.*, 1993; Rio *et al.*, 1991).

The first three-dimensional structure of PSP was solved by X-ray diffraction to 2.5 Å resolution (Gajhede *et al.*, 1993) from an orthorhombic crystal form. The structure has also been solved to 2.6 Å resolution (De *et al.*, 1994) and by NMR spectroscopy (Carr, 1992). One major result of the structural analysis was the discovery of a groove in the trefoil-loop domain that seemed of appropriate size for binding with a substrate molecule. The groove is found to be constructed mainly from residues which are conserved within the trefoil family. Circumstantial evidence indicates that the mammalian members of this family of proteins are associated with the mucus layer of the gastrointestinal tract (Frandsen, Jørgensen & Thim, 1986). This mucus layer is primarily made of mucins (Strous & Dekker, 1992) which are high molecular weight glycoproteins with oligosaccharides attached to the side chains of Ser/Thr. The carbohydrates in the peripheral region of these oligosaccharides could be possible substrates for PSP, but so far a natural substrate has not been found and the actual function of PSP is yet to be determined.

The present study is based on a trigonal crystal form of PSP which contains a dimer in the asymmetric unit. This crystal form diffracts beyond 2 Å resolution. The enhanced resolution obtained, compared with the previous studies, enables a more detailed analysis of the grooves. The electrostatic properties of the grooves are characterized and an analysis of the water structure has revealed a few important residues as anchor points between the protein and a potential substrate.

2. Experimental

2.1. *Crystallization and preparation of heavy-atom derivatives*

Orthorhombic crystals of PSP, crystallized in space group $I2_12_12_1$ have been described previously (Gajhede, Thim, Jørgensen & Melberg, 1992). However, the diffraction from these crystals was limited to 2.5 Å, even when using synchrotron radiation (Gajhede *et al.*, 1993). Consequently, crystallization experiments were set up to produce an alternative crystal form. The experiments were carried out using the hanging-drop vapour-diffusion technique set up in Linbro tissue-culture multiwell plates (Flow Laboratories Inc., Mclean, Virginia 22102, USA). The drops were made by mixing 3 µl of the reservoir solution with 3 µl of the protein solution at a concentration of 20 mg ml^{-1}. Well diffracting crystals, with a hexagonal brick-like morphology, appeared after 2–3 weeks. The best results were obtained using a crystallizing solution that contained 2.0 *M* $(NH_4)_2SO_4$, 7%(*w*/*v*) PEG 400 and 0.1 *M* Hepes buffered at pH 7.0. Often a two-phase system appeared with small oily droplets within the main drop and the crystals grew either at this interface or at the edge of the drop. Reproducing these crystals proved difficult as the optimal conditions for crystallization are within a very narrow range, especially with respect to the protein concentration. A screening around these conditions was, therefore, always set up to obtain crystals of optimal size and quality. The dimensions of the largest crystals were approximately 0.3 × 0.2 × 0.2 mm.

Heavy-atom derivatives were searched for with prior knowledge of two successful reagents which bound in the orthorhombic crystal form of PSP (Gajhede *et al.*, 1992). These reagents were $AgNO_3$ and K_2PtCl_4, both of which also bound to the new hexagonal crystal form. The soaking method was used and the final heavy-atom concentration in the drops was 2.5 m*M* in both cases. Later, a third derivative was obtained with $(CH_3)_3PbOOCCH_3$ with a final heavy-atom concentration of 10.0 m*M*. The soaking times for the three derivatives were 24, 6 and 45 h, respectively.

2.2. *Data collection and processing*

All data sets were collected in-house on a Rigaku R-AXIS IIC image-plate system with a rotating anode operating at 50 kV and 180 mA. Using a graphite monochromator and a 0.5 mm collimator, monochromatic Cu $K\alpha$ radiation was produced. The crystal to image-plate distance was 112 mm. The temperature was 274 K during the native data collection while the derivative data sets were collected at 285 K. A detector offset of $2\theta = 15^\circ$ was used while recording the native data in order to obtain the highest possible resolution (24 frames, 2.0° oscillation). No offset was used for the derivatives: $AgNO_3$ (60 frames, 1.5° oscillation), $PtCl_4$ (40 frames, 2.0° oscillation) and $(CH_3)_3PbOOCCH_3$ (22 frames, 2.0° oscillation). The native crystals belong to the trigonal space group $P3_121$ with unit-cell parameters $a = b = 60.6$, $c = 113.1$ Å. There was no significant change in the unit-cell parameters for the derivatives. The crystals contain two molecules in the asymmetric unit and the solvent content was calculated to be 52% by use of Matthews' formula (Matthews, 1968). The volume-to-mass ratio was 2.55 Å^3 Da^{-1}.

The *DENZO* program (Otwinowski, 1993) was used to integrate the raw data and to calculate standard deviations, whereupon programs from the *CCP*4 package (Collaborative Computation Project, Number 4, 1994) were used for scaling, merging and truncation of the data.

The final data-collection statistics for all data sets are summarized in Table 1.

2.3. *Heavy-atom derivatives*

$R_{\rm iso}$ values $[R_{\rm iso} = \sum_{hkl} |F_{PH}(hkl) - F_P(hkl)| / \sum_{hkl} F_{PH}(hkl)]$, were used as an initial indication as to whether a collected set of data was a derivative

Table 1. *Data-collection statistics*

Only fully recorded reflections were used in the data processing for the Pt derivative. The rows are interpreted as follows. Total number of reflections with a partiality above 0.5. Number of rejections after merging. Number of unique reflections within the used resolution range. Completeness is the percentage of reflections used compared to the number of theoretically obtainable measurements. Outermost shell is the highest resolution shell with an internal R factor, $R_{\rm merge}$, below 20%, $R_{\rm merge} = \sum_{hkl}\sum_i |I(hkl)_i - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I(hkl)_i$. $I/\sigma(I) > 2$ is the percentage of data intensities greater than two standard deviations.

	Native	Ag	Pt	Pb
No. of reflections	31498	44580	48102	16116
No. rejected	427	1325	24	87
No. of unique reflections	15581	8689	11139	5395
Completeness (%)	87.3	99.9	80.9	85.4
Lowest resolution (Å)	50.0	30.0	30.0	30.0
Outermost shell (Å)	2.05–1.95	2.63–2.50	2.27–2.15	2.95–2.80
Completeness outermost shell (%)	73.4	99.5	82.8	86.1
Overall $I/\sigma(I) > 2$ (%)	81.5	95.5	87.4	92.0
Outermost shell $I/\sigma(I) > 2$ (%)	56.3	92.0	72.9	85.6
$R_{\rm merge}$ (%) overall	4.8	4.2	5.1	4.9
$R_{\rm merge}$ (%) outermost shell	18.7	7.7	16.7	11.7

Table 2. *Space-group determination and phasing statistics*

Comparison of the mean figure of merit ⟨FOM⟩ for the two enantiomorphic space groups: $P3_121/P3_221$. The space-group numbers are 152 and 154, respectively, according to *International Tables for Crystallography* (1995, Vol. A). Phase calculations with the *SQUASH* program were made with all the heavy-atom sites included. $R_{\rm Cullis}$ and phasing powers are calculated using root mean squares (r.m.s.). E = lack of closure, E' = r.m.s.(E), $R_{\rm Cullis}$ = r.m.s.(E)/$R_{\rm iso}$, phasing power = r.m.s.(f_H/E). $E = \sum ||F_{PH} \pm F_P| - f_H|$, $E' = (E^2/n)^{0.5}$. R.m.s. $f_H = (\sum f_H^2/n)^{0.5}$, n = number of terms. $R_{\rm iso} = \sum_{hkl} |F_{PH} - F_P| / \sum_{hkl} F_P$.

	MLPHARE		*SQUASH*	
Heavy atoms included	$\langle{\rm FOM}\rangle_{152}$	$\langle{\rm FOM}\rangle_{152}$	$\langle{\rm FOM}\rangle_{152}$	$\langle{\rm FOM}\rangle_{152}$
Pt	0.360	0.269	—	—
Pt, Ag1	0.351	0.302	—	—
Pt, Ag1, Ag2	0.399	0.314	—	—
Pt, Ag1, Ag2, Pb	0.451	0.331	0.631	0.493
Derivative		Ag	Pt	Pb
$R_{\rm Cullis}$ (centric reflections)		0.89	0.70	0.88
Phasing power (centric reflections)	0.40	0.80	0.40	
Phasing power (acentric reflections)	0.60	1.10	0.50	
$R_{\rm iso}$ (%)		19.7	19.5	12.7

Table 3. *Heavy-atom parameters*

Pt and Ag1 are bound at the same positions in the structure of molecule B, between the two residues: Met99 and Met60. Ag2 is bound at an analogous position in molecule A. The Pb atom is bound to the atoms Oδ1 and Oδ2 of Asp87A.

	x	y	z	Isomorphous occupancy	Anomalous occupancy	B (Å^2)
Pt	0.344	−0.037	0.655	0.683	0.633	43.5
Ag1	0.347	−0.037	0.651	0.293	0.189	12.4
Ag2	0.518	0.296	0.582	0.351	0.303	17.9
Pb	0.511	0.182	0.422	0.206	0.250	22.9

($R_{\rm iso} \simeq 15\%$), or if it was differently indexed with respect to the native data, which would give extremely high $R_{\rm iso}$ values ($R_{\rm iso} \simeq 50\%$) due to a possible rotation of 60° in the *hk*0 layer. A difference Patterson map was produced and interpreted for the Pt derivative in space group $P3_121$ keeping in mind that the heavy-atom position might have to be inverted, as the correct enantiomorphic space group ($P3_121/P3_221$) was not yet determined. After refinement of the Pt site with the program *HEAVY* (Terwillinger & Eisenberg, 1983), phases were calculated to 3 Å resolution using *MLPHARE* (Collaborative Computational Project, Number 4, 1994) and afterwards improved with the *SQUASH* program (Zhang & Main, 1990) using the options: solvent flattening, histogram matching and Sayre's equation. A cross Fourier synthesis with the other derivatives revealed two heavy-atom sites in the Ag derivative and one site in the Pb derivative. With the inclusion of the anomalous differences for all derivatives, the phasing statistics, as shown in Table 2, clearly demonstrate that the correct space group is $P3_121$. The final heavy-atom positions, occupancies and B factors are listed in Table 3, where it is also seen that the Pt site and one Ag site are located at the same positions in the unit cell.

With these phases an electron-density map was calculated using the *CCP*4 program *FFT* and the map was thoroughly inspected with the modelling program *O* (Jones, Zou, Cowan & Kjeldgaard, 1991). Unfortunately, the map was of very poor quality without any obvious or recognizable features that could be used as starting points in the tracing of the protein molecules.

Meanwhile, the orthorhombic crystal form of PSP, which also crystallizes as a dimer, had been solved to 2.5 Å resolution (Gajhede *et al.*, 1993) and could therefore be used as a search model in a molecular-replacement search against the trigonal crystal form.

2.4. *Molecular replacement*

Programs from the *X-PLOR* package (Brünger, 1992*a*,*b*) were used for the molecular-replacement search. The dimer from the orthorhombic crystal form of PSP, with the exclusion of water molecules, was used as search model. The rotation search gave a set of solutions, from which 6000 with the highest Rf values were further refined by means of Patterson correlation (PC) refinement (Brünger, 1990). The search model was divided into four rigid bodies, since each monomer contains two homologous domains. The four rigid bodies were defined by the residues 1–53 and 54–106 in both molecules *A* and *B*. After the conventional rotation search, the correct solution had the 12th highest Rf value, but this was dramatically improved after PC refinement when it became the top solution, with a height of 7.5σ. The rotation solution is given by the Euler angles $\theta_1 = 105.57^\circ$, $\theta_2 = 75.03^\circ$, $\theta_3 = 53.44^\circ$ according to *X-PLOR* conventions (Brünger, 1992*a*,*b*). With this rotation the translation search was straightforward using an *X-PLOR* script (Brünger, 1992*a*,*b*). A unique solution for the translation function ($\mathbf{t} = 0.540$, 0.186, 0.018) was found for the space group $P3_121$.

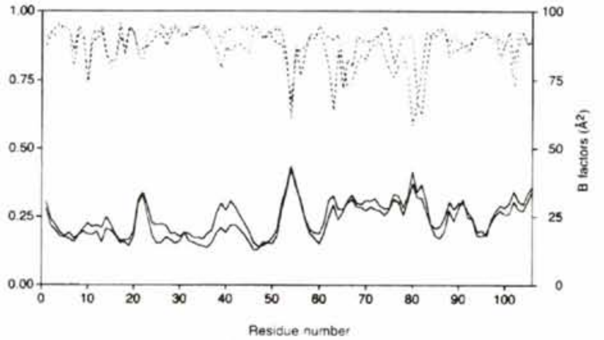

Fig. 2. *Rsfit* and *B*-factor plot. The molecules *A* and *B* are represented by the colours magenta and cyan, respectively. *Rsfit* values were calculated using all atoms within a residue and these are shown as dashed lines. The average main-chain *B* factors are shown as full lines. The higher thermal mobility of the C-terminal domain compared to the N-terminal domain can clearly be seen. The region connecting the two domains is Lys48–Glu57.

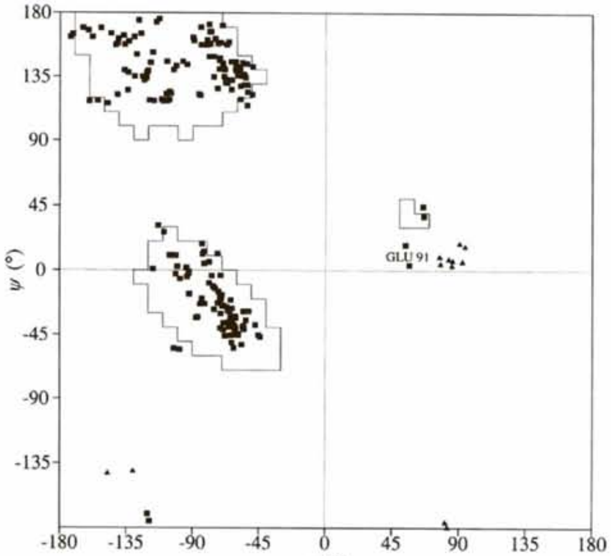

Fig. 3. Ramachandran plot of the dimer. The Ramachandran plot for all 212 residues is shown. Glycine residues are shown as triangles and other residues as squares. The only non-glycine residue in the disallowed region is Glu91*A*. Line borders are shown for the most favored regions, as defined by *PROCHECK*. There are 172 non-glycine/proline residues of which 89.0% are inside the most favoured regions, 10.5% of the residues in the additional allowed regions and 0.6% are in the generously allowed regions.

Table 4. *R values as a function of resolution*

The R and R_{free} values are calculated from reflections within the specified resolution ranges whereas the accumulated R and R_{free} values are based on reflections within the range 50 Å to the highest specified resolution. The test set comprises 10% of randomly selected reflections over the whole resolution range and these reflections have been omitted for all refinement steps. The R factors are defined as $R = \sum_{hkl} |F(\text{obs})_{hkl} - F(\text{calc})_{hkl}| / \sum_{hkl} F(\text{obs})_{hkl}$.

	Working set			Test set		
$d_{\min}$–$d_{\max}$ (Å)	N_{obs}	R	R value accum.	N_{obs}	R_{free}	R_{free} accum.
3.90–50.00	1915	0.1422	0.1422	226	0.2016	0.2016
3.10–3.90	1923	0.1533	0.1473	202	0.2310	0.2142
2.70–3.10	1890	0.2025	0.1597	235	0.2848	0.2304
2.46–2.70	1851	0.2308	0.1696	226	0.2927	0.2397
2.28–2.46	1839	0.2677	0.1800	212	0.3103	0.2470
2.15–2.28	1651	0.2688	0.1870	193	0.2854	0.2502
2.04–2.15	1537	0.2864	0.1928	155	0.3366	0.2542
1.95–2.04	1461	0.2980	0.1976	136	0.3432	0.2576

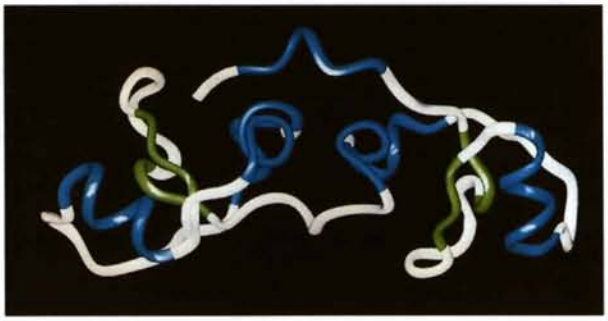

Fig. 4. Secondary-structure elements of the monomer. A view of the monomer seen along the internal twofold axis. The secondary-structure elements in PSP are very few and short. It is predominantly loops (grey) and very short β-strands (green), including only four residues each. The α-helices and 3_{10}-helices (blue) lie on each side of the antiparallel β-sheets and in the C-terminal tail of PSP.

Table 5. *Assignment of secondary-structure elements*

β-strands 1 and 2 form a short antiparallel β-sheet in the N-terminal domain and likewise β-strands 3 and 4 in the C-terminal domain. The 3_{10}-helix *D* in the C-terminal tail lies outside the two trefoil domains.

N-terminal domain		C-terminal domain	
α-Helix *A*	4–10	3_{10}-Helix *A'*	55–59
α-Helix *B*	25–32	α-Helix *B'*	74–81
3_{10}-Helix *C*	12–16	3_{10}-Helix *C'*	61–65
β-strand 1	35–37	β-strand 3	84–86
β-strand 2	45–47	β-strand 4	94–96
		3_{10}-Helix *D*	100–104

The following rigid-body refinement resulted in a satisfactory packing of the symmetry-related molecules, confirmed with the modelling program *O* (Jones *et al.*, 1991). The *R* value after rigid-body refinement was 40.6% to 4.0 Å resolution.

2.5. *Model building and refinement*

The dimer model of PSP has been refined using the program *X-PLOR* (Brünger, 1992*a*,*b*). All data within the resolution limits 50.0–1.95 Å was used in the refinement. To help avoid model bias in the refinement, 10% of the reflections were randomly selected and kept out of the refinement procedures. This gave a test set of 1585 reflections, from which $R_{\rm free}$ values could be calculated (Brünger, 1992*a*,*b*), while the working set comprised 14 067 reflections. The reflections in the test set are not completely unbiased by the model, because of the non-crystallographic twofold axis. The refinement was carried out using the standard slow-cool protocol and non-crystallographic symmetry (NCS) restraints were used on the main-chain atoms. The side chains can have very different orientations in the two monomers, therefore, NCS restraints were not used for these atoms. An NCS energy term weight of 100 kcal mol^{-1} Å^{-2} (418.4 kJ mol^{-1} Å^{-2}) was found to be useful, as it allowed some deviations in the main-chain trace. *B* factors were refined using restrained individual *B*-factor optimization. A bulk solvent correction was performed after calculation of a bulk solvent mask and determination of the optimal scale and *B* factor for the solvent. The scale and *B* factor were 0.355 and 120 Å^2, respectively. After a refinement cycle the molecules were manually adjusted in the program *O* (Jones *et al.*, 1991).

A symmetry-averaged multiple isomorphous replacement (MIR) electron-density map was used whenever the $2F_o - F_c$ electron-density map was difficult to interpret. The initial mean figure of merit of the MIR map, after solvent flattening, histogram matching and Sayre's equation, was 0.631 to 3.0 Å resolution. A better envelope was made with the program *MAMA* (Kleywegt & Jones, 1993), based on the partially refined coordinates of the *B* molecule and the NCS operator was determined with the *Lsq_improve* option within the program *O*. The monomers are related by the Euler angles $\theta_1 = 241.5$, $\theta_2 = 73.5$, $\theta_3 = 299.4^\circ$. Averaging of the electron-density map with the *CCP*4 program *DM* (Cowtan, 1994) greatly improved the map and the figure of merit increased to 0.791. The correlation coefficient between the two electron-density regions related by non-crystallographic symmetry, increased from 0.265 to 0.853 after 50 cycles of averaging.

Water molecules were included in the structure if they appeared in both the $2F_o - F_c$ and the $F_o - F_c$ electron-density maps and if reasonable hydrogen bonds could be established. The $2F_o - F_c$ electron-density map was contoured at 1σ whereas the $F_o - F_c$ was contoured at 3.2σ. After a refinement cycle, water molecules were examined and rejected from the model if the *B* factor had exceeded 70 Å^2. A total of 183 water molecules are included in the final structure with a mean *B* factor of 39.3 Å^2 and the highest *B* factor being 66.5 Å^2. The *RSFit* correlation coefficients and the *B* factors for all residues in the *A* and *B* molecule are shown in Fig. 2. It can be seen that the higher *B* factors correspond to the less well defined electron-density regions. The highest temperature factors occur for the residues Phe21, Pro22 and Glu54 in both molecules and Gly80 in molecule *A*. There is no electron density for the side chain of Glu54 in both molecules, whereas it is poorly defined for Phe21, Pro22 and Gly80*A*. The average *B* factor for 1624 non-H protein atoms is 24.6 Å^2. An analysis of the protein geometry was performed with the program *PROCHECK* (Laskowski, MacArthur, Moss & Thornton, 1993) and the Ramachandran plot, Fig. 3, shows that the only non-glycine/proline in the disallowed regions is Glu91*A*, ($\varphi = 56.2$, $\psi = 4.2^\circ$). This residue is involved in an intermolecular hydrogen bond, which is probably responsible for the non-standard geometry of the residue. The r.m.s. deviations from ideality for bond length, bond angles, dihedrals and impropers are 0.014 Å, 1.685, 26.76 and 1.635°, respectively (Engh & Huber, 1991). The final *R* values and the $R_{\rm free}$ values are shown as a function of the resolution in Table 4.

2.6. *Description of structure*

The secondary-structure assignment (Kabsch & Sander, 1983) is shown for the monomer in Table 5. It has been divided into two domains, the N-terminal and the C-terminal domain, to emphasize the intramolecular similarities. There are three helices and one short antiparallel β-sheet in each domain and an additional 3_{10}-helix in the C-terminal tail. In the N-terminal domain, the two α-helices *A* and *B*, lie on each side of the β-sheet (β-strands 1 and 2). The same arrangement is found in the C-terminal domain, but the counterpart of α-helix *A* has become a 3_{10}-helix with two fewer residues. Fig. 4 shows the secondary-structure elements of the monomer. Molecule *B* is, in general, marginally better defined by the electron density, compared to molecule *A*, therefore, molecule *B* will be referred to in the forthcoming analysis.

Superpositioning the Cα atoms in the two domains (5–53 and 54–102), using the *Lsq_improve* option in the program *O*, gives a root-mean-square deviation (r.m.s.d.) of 1.94 Å. Each domain contains three loop regions. In the N-terminal domain loop 1 is defined by residues 8–18, loop 2 by residues 19–34 and loop 3 by residues 35–46 and in the C-terminal, the loops are defined as: loop 1 residues 58–67, loop 2 residues 68–83 and loop 3 residues 84–95. There is one residue less in loop 1 of the C-terminal domain compared to loop 1 in the N-terminal domain. This probably causes the change of the first helix in the C-terminal domain, from an α-helix to a 3_{10}-helix. A cleft is situated between loop 1 and loop 2 in each domain and the walls of the cleft in the N-terminal domain are defined both by backbone atoms (residues 19–26, 42–46) and side-chain atoms (residues 18, 22–26, 29, 36, 44–46). Nearly all residues in the cleft are part of the consensus sequence as seen in Fig. 1. Trp45 is completely conserved within the family and the side chain makes a hydrophobic wall suitable for substrate stacking. Another important and conserved residues is Arg16*B* which stabilizes the area to the back of the cleft through hydrogen bonds between Arg16 Nη1 ··· Pro44 O (2.9 Å) and Arg16 Nη2 ··· Asp37 Oδ2 (2.7 Å). The $2F_o - F_c$ electron density is shown for this area in Fig. 5.

Very few residues are involved in the intermolecular interactions between the two monomers. Three intermolecular hydrogen bonds are found Asn15*A* Oδ1 ··· Asn67*B* Nδ2 (3.2 Å), Val62*A* N ··· Glu91*B* Oε1 (3.1 Å) and Glu91*A* Oδ1 ··· Val62*B* N (3.3 Å). Also a hydrogen-bonded chain, involving a water molecule, is found between Glu91*A* Oδ1 ··· HOH96 (2.9 Å) and HOH96 ··· Met60*B* O (3.0 Å). This interaction stabilizes loop 3 of molecule *A*, which fits into the interdomain region of molecule *B*, as seen in Fig. 6.

Attempts have been made to find a substrate model to complex with PSP, but at present it has only been possible to soak oligosaccharides into the groove of the orthorhombic crystal form of PSP (manuscript in preparation). The same experiment did not work with the trigonal crystal form, possibly because of a lower solvent content and different packing of the protein molecules. However, these experiments have not revealed the natural substrate of PSP, which is likely to be a complicated branched oligosaccharide (Strous & Dekker, 1992). Hence, soaking experiments with simple oligosaccharides are only expected to give limited information about the substrate binding. In contrast to this, a detailed analysis of the water structure in the binding pocket will give information about hydrophilic positions of the true substrate. A general hydrogen-bonding pattern was found in the four grooves of the dimer trigonal crystal form. Hydrogen bonds between four water molecules and four residues were found in the cleft. The interactions in the N-terminal cleft are Asn18 Oδ1 ··· HOH40 (3.4 Å), Gly23 N ··· HOH85 (2.6 Å), Ile24 O ··· HOH95 (3.0 Å) and Cys46 N ··· HOH173 (3.4 Å). A view from the top into the N-terminal cleft is shown with a solvent-accessible surface (Connolly, 1983) in Fig. 7. The important protein atoms involved in hydrogen bonding to the four water molecules are shown as balls.

3. Discussion

The structural investigation of PSP has revealed that it is a two-domain protein with a cleft area in each domain. From the sequence alignment of the known proteins belonging to the trefoil family, Fig. 1, it is found that all conserved residues are in the vicinity of the cleft areas.

An analysis of the structure shows a high internal symmetry. Superimposing the residues defining the N-terminal cleft upon those in the C-terminal cleft reveals a high degree of positioning similarity. There is, however, a minor difference in the main-chain progression because of an altered carbonyl O-atom orientation at Cys19 in the N-terminal cleft compared to Cys68 in the C-terminal cleft. This difference is seen in both molecules *A* and *B* and also in both molecules from the orthorhombic crystal form of PSP. The carbonyl O atom of Cys68, in molecule *B*, is kept in place through a hydrogen bond to Arg81 Nη1 (2.7 Å) but a similar hydrogen bond cannot be established in the N-terminal cleft to Cys19 as a serine, Ser32, is present at the analogous position to Arg81. The main-chain differences affect the width of the clefts in a way that enlarges the N-terminal cleft, Pro22 Cα–Trp45 Cα (9.8 Å), compared to the C-terminal cleft, Pro71 Cα–Trp94 Cα (8.6 Å).

The discovery of a general hydrogen-bond pattern in the binding pockets, has provided new information about the true substrate of PSP and the possible residues involved in substrate binding. Four important residues have been identified as possible anchor points for a substrate, Asn18, Gly23, Ile24 and Cys46. The three residues Gly23, Ile24 and Cys46 make hydrogen bonds through a main-chain atom, while Asn18 makes a hydrogen bond through a side-chain atom. Cys46 is the only residue that is completely conserved within the family, as it makes a disulfide bridge to Cys29, which is essential for the formation of the binding pocket. Gly23 and Ile24 are not conserved residues, but the main-chain atoms might be at almost the same positions, whether it is a Thr or a His, which are the residues found in other proteins from the trefoil family of loop peptides, at the position of Glu23 or if it is a Val which can be present at the position of Ile24. An Asp can also be found at the position of Asn18. This means that the donor/acceptor property of the residue is conserved and a hydrogen bond to the substrate can still be formed. The size of the binding pocket and the locations of the anchor points in the protein, suggests that two monosaccharide sites are present in the groove. Asn18 Oδ1 is most likely to make a hydrogen bond to the terminal unit of an oligosaccharide as well as Cys46 N, while Gly23 N and Ile24 O are most likely to interact with the second oligosaccharide unit. The side chain of Trp45 will act as a hydrophobic wall and stack correctly against the second oligosaccharide unit.

Analysis of the charge distribution shows essentially non-charged cleft areas, Fig. 8, but the minor differences in the main-chain trace as well as the Ser32/Arg81 'substitution', makes the N-terminal cleft slightly negatively charged and the C-terminal cleft slightly positively charged. This difference might result in a substrate specificity that is in accordance with binding studies made with PSP and mucus from the gastrointestinal tract from rats (Frandsen *et al.*, 1986). A high-affinity low-capacity binding site and a low-affinity high-capacity binding site was proposed with K_d's of 1.3×10^{-7} and 4.2×10^{-6} mol l^{-1}, respectively. The unequal K_d's might also reflect the fact that binding in one cleft could introduce steric hindrance towards substrate binding in the other cleft area. This is plausible as the mucus consists of very large glycoproteins with molecular weights up to 1.6 MDa (Strous & Dekker, 1992). A comparison of this high-resolution trigonal crystal form and the orthorhombic crystal form shows a very high agreement with an r.m.s.d. of the Cα atoms of 0.930 Å between the two *A* molecules and a r.m.s.d. value at 0.827 Å for the *B* molecules. Minor differences occur in the assignment of the secondary-structure elements. α-Helix *A* (4–10) has been extended by one residue in the N-terminal end and a former hydrogen-bonded turn (100–103) has now been assigned as a 3_{10}-helix *D* (100–104).

The atomic coordinates and the structure factors have been deposited in the Brookhaven Protein Data Bank.*

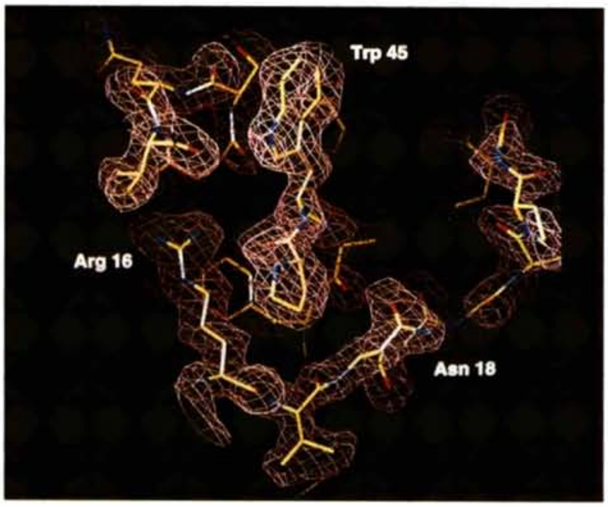

Fig. 5. The final electron-density map. The $2F_o - F_c$ electron-density map is shown for a region at the back of the binding pocket. Arg16 stabilizes the back of the binding pocket and Trp45 is one of the most important residues in the binding pocket. The map is made with the *map_cover* option in the program *O*, which means that density that is more than 0.95 Å away from a protein atom, is not shown.

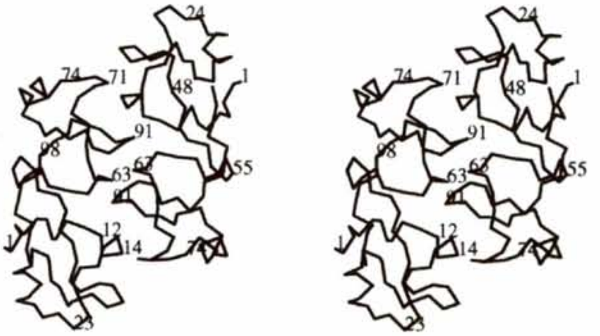

Fig. 6. A stereoview of the Cα atoms in the dimer. The two monomers in the asymmetric unit pack with a limited number of interactions. The tip of loop 3 in one molecule fits into the interdomain region of the other molecule. This seems to be the driving force in the dimerization of PSP.

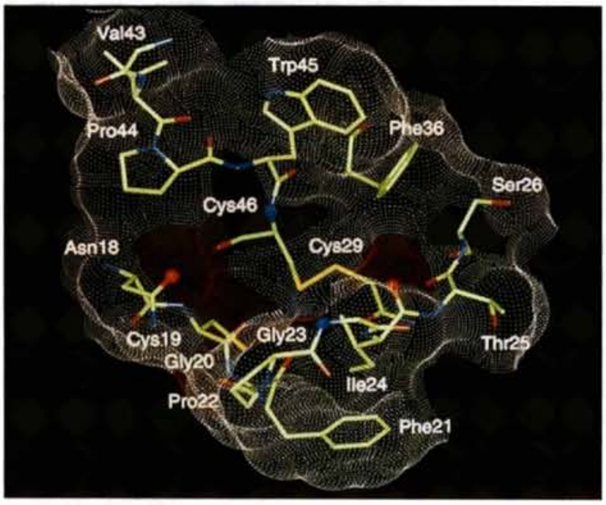

Fig. 7. The substrate-binding area of PSP. A view from the top of the N-terminal cleft area is shown. The four protein atoms involved in hydrogen bonding to water molecules are showed as balls, with N atoms in blue and O atoms in red. The size of the cleft is approximately 9 × 9 × 12 Å being the width, depth and length, respectively.

* Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory (Reference: 2PSP, R2PSPSF). Free copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England (Reference: JN0015).

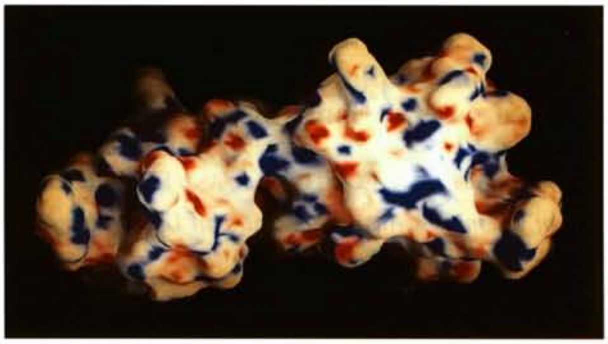

Fig. 8. Topology and electrostatic potential charge distribution. The electrostatic potential charge distribution of molecule *B* as calculated using the *DelPhi* program (*InsightII Version* 2.3.0, Biosym Technologies, 1993). The environment of the molecule is that present in the crystalline state, *e.g.* pH, ionic strength and dielectricity of the crystallization media. The mapping was performed with the *GRASP* program (Nicholls, Bharadwaj & Honig, 1993). At the right side of the molecule, a sideview of the N-terminal cleft is seen and the C-terminal cleft is seen from the top at the left side of the picture. The dark red colour corresponds to a potential $<-6\,\mathrm{kT\,e^{-1}}$, blue colour to a potential $>+6\,\mathrm{kT\,e^{-1}}$.

We gratefully thank Dr Lars Thim who provided the purified protein. We also thank Dr Anders Kadziola and Dr Paul Rowland for many helpful discussions. This work was supported by the Danish National Research Foundation.

References

Biosym Technologies (1993). *DelPhi InsightII* version 2.3.0. Biosym Technologies, 9685 Scranton Road, San Diego, CA, USA.

Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR version 3.1. A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, Connecticut, USA.

Carr, M. D. (1992). *Biochemistry*, **31**, 1998–2004.

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

Connolly, M. L. (1983). *Science*, **221**, 709–713.

Cowtan, K. (1994). *Jnt CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **31**, 34–38.

De, A., Brown, D. G., Gorman, M. A., Carr, M., Sanderson, M. R. & Freemont, P. S. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 1084–1088.

Engh, R. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Frandsen, E. K., Jørgensen, K. H. & Thim, L. (1986). *Regul. Peptides*, **16**, 291–297.

Gajhede, M., Petersen, T. N., Henriksen, A., Petersen, J. F. W., Dauter, Z., Wilson, K. S. & Thim, L. (1993). *Structure*, **1**, 253–262.

Gajhede, M., Thim, L., Jørgensen, K. H. & Melberg, G. S. (1992). *Proteins*, **13**, 364–368.

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

Jørgensen, K. H., Thim, L. & Jacobsen, H. E. (1982). *Regul. Peptides*, **3**, 207–219.

Kabsch, W. & Sander, C. (1993). *Biopolymers*, **22**, 2577–2637.

Kleywegt, G. J. & Jones, T. A. (1993). *ESF/CCP4 Newslett.* **28**, 56–59.

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

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Nicholls, A., Bharadwaj, R. & Honig, B. (1993). *Biophys. J.* **64**, A166–A166.

Otwinowski, Z. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, pp. 56–62. Warrington: Daresbury Laboratory.

Podolsky, D. K., Lynch-DeVaney, K., Stow, J. L., Oates, P., Murgue, B., De-Beaumont, M., Sands, B. E. & Mahida, Y. R. (1993). *J. Biol. Chem.* **268**, 6694–6702.

Rasmussen, T. N., Harling, H., Thim, L., Pierznowsky, S., Weström, B. R. & Host, J. J. (1993). *Am. J. Physiol.* **264**, G22–G29.

Rio, M.-C., Chenard, M.-P., Wolf, C., Marcellin, L., Tomatsetto, C., Luthe, R., Bellocq, J.-P. & Chamson, P. (1991). *Gastroenterology*, **100**, 375–379.

Strous, G. J. & Dekker, J. (1992). *Crit. Rev. Biochem. Mol. Biol.* **27**, 57–92.

Suemori, S., Lynch-Devaney, K. & Podolsky, D. K. (1991). *Proc. Natl Acad. Sci. USA*, **88**, 11017–11021.

Terwillinger, T. C. & Eisenberg, D. (1983). *Acta Cryst.* A**39**, 813–817.

Thim, L. (1994). *Digestion*, **55**, 353–360.

Thim, L., Jørgensen, K. H. & Jørgensen, K. D. (1982). *Regul. Peptides*, **3**, 221–230.

Thim, L., Norris, K., Norris, F., Nielsen, P. F., Bjørn, S. E., Christensen, M. & Petersen, J. (1993). *FEBS Lett.* **318**, 345–352.

Thim, L., Thomsen, J., Christensen, M. & Jørgensen, K. H. (1985). *Biochim. Biophys. Acta*, **827**, 410–418.

Tomasetto, C., Rio, M. C., Gautier, C., Wolf, C., Hareveni, M., Chambon, P. & Lathe, R. (1990). *EMBO J.* **9**, 407–414.

Wright, N. A., Pike, C. & Elia, G. (1990). *Nature (London),* **343**, 82–85.

Wright, N. A., Poulsom, R., Stamp, G. W. H., Hall, P. A., Jeffery, R. E., Longcraft, J. M., Rio, M.-C., Tomasetto, C. & Chamson, P. (1990). *J. Pathol.* **162**, 279–284.

Wright, N. A., Poulson, R., Stamp, G., Vannoorden, S., Saffaf, C., Elia, G., Ahnen, D., Jeffery, R., Longcraft-Pike, C., Rio, M.-C. & Chamson, P. (1993). *Gastroenterology*, **104**, 12–20.

Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 41–45.