Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure of aspergillopepsin I from *Aspergillus phoenicis*: variations of the S1′–S2 subsite in aspartic proteinases

Sang Woo Cho, Nam-june Kim, Myung-Un Choi and Whanchul Shin*

School of Chemistry and Molecular Engineering, and Center for Molecular Catalysis, Seoul National University, Seoul 151-742, South Korea

Correspondence e-mail: nswcshin@plaza.snu.ac.kr

Received 8 December 2000
Accepted 5 April 2001

PDB Reference: aspergillopepsin I, 1ibq.

The crystal structure of aspergillopepsin I (AP) from *Aspergillus phoenicis* has been determined at 2.18 Å resolution and refined to R and $R_{\rm free}$ factors of 21.5 and 26.0%, respectively. AP has the typical two β-barrel domain structure of aspartic proteinases. The structures of the two independent molecules are partly different, exemplifying the flexible nature of the aspartic proteinase structure. Notably, the 'flap' in one molecule is closer, with a largest separation of 4.0 Å, to the active site than in the other molecule. AP is most structurally homologous to penicillopepsin (PP) and then to endothiapepsin (EP), which share sequence identities of 68 and 56%, respectively. However, AP is similar to EP but differs from PP in the combined S1′–S2 subsite that is delineated by a flexible ψ-loop in the C-terminal domain. The S1′ and S2 subsites are well defined and small in AP, while there is no definite border between S1′ and S2 and the open space for the S2 subsite is larger in PP. Comparison of the structures indicates that the two amino-acid residues equivalent to Leu295 and Leu297 of AP are the major determining factors in shaping the S1′–S2 subsite in the fungal aspartic proteinases.

1. Introduction

Aspartic proteinases are a group of proteolytic enzymes in which the scissile peptide bond is attacked by a nucleophilic water molecule activated by two aspartic residues in a DT(S)G motif at the active site (Barrett *et al.*, 1998). They have a similar fold composed of two β-barrel domains, as revealed from numerous X-ray structures either in native forms or in complexes with substrate analogues (Davies, 1990). Between the N-terminal and C-terminal domains, each of which contributes one catalytic aspartic residue, there is an extended active-site cleft capable of interacting with multiple residues of a substrate. Although members of the aspartic proteinase family of enzymes have very similar three-dimensional structures and catalytic mechanisms, each has a unique substrate specificity. Prediction of the binding preferences from the amino-acid sequence is difficult because the specificity is determined by extensive interactions occurring at multiple subsites and the shape of each subsite is delineated by the structures of the flexible loops as well as the individual amino-acid residues in the loops (Dunn & Hung, 2000). Site-directed mutagenesis showed that a single mutation at the binding site alters the substrate specificity (Scarborough & Dunn, 1994; Shintani *et al.*, 1997).

Aspartic proteinases (aspergillopepsins; E.C. 3.4.23.18) are found in various *Aspergillus* species, together with *Penicillium*, the most economically important genus of fungi (Pitt & Samson, 1990). They may serve as virulence factors or as

industrial aids. For instance, aspergillopepsin F from *A. fumigatus* is involved in invasive aspergillosis owing to its elastolytic activity (Lee & Kolattukudy, 1995) and aspergillopepsins from the mould *A. saitoi* are widely used in the Asian fermentation industry (Ichishima, 1998). X-ray structures are not available for any aspergillopepsins, while those of penicillopepsin (PP) from *P. janthinellum* with and without inhibitors have been extensively characterized (James & Sielecki, 1983; Khan *et al.*, 1998; Fujinaga *et al.*, 2000). In this study, we have determined the crystal structure of aspergillopepsin I (AP) from *A. saitoi* (now designated *A. phoenicis* ATCC 14332) at 2.18 Å resolution.

AP shows a broad primary substrate specificity (Shintani & Ichishima, 1994). It favours hydrophobic residues at the P1 and P1′ positions, but also accepts a lysine residue in the P1 position, leading to the activation of trypsinogen and chymotrypsinogen A (Abita *et al.*, 1969; Shintani *et al.*, 1996). Because AP shares a high sequence identity of 68% with PP, it has generally been assumed that both proteins have very similar structures (Ichishima, 1998). However, they differ in some properties. For instance, PP clots milk while AP does not, indicating that their overall structures may be quite similar but their substrate-binding sites may be different (Ichishima, 1998). The present study revealed that not only the structures of a flexible ψ-loop in the C-terminal domain but also the S1′ and S2 subsites are significantly different despite the high sequence homology of the two proteins. In this article, the first structure of an aspartic proteinase from the *Aspergillus* species is described and the S1′ and S2 subsites in the aspartic proteinases are compared in detail.

2. Materials and methods

2.1. Purification and crystallization

AP was purified from a crude fungal protease (Protease type XIII, P-2143) from Sigma (St Louis, Missouri), following the procedure of Ichishima & Yoshida (1965) with minor modifications. The purified enzyme solution was concentrated to a final concentration of 20 mg ml^{-1}. Crystallization experiments were carried out at 277 K by sitting-drop vapour diffusion. The trapezoid-shaped crystals could be reproducibly grown in drops containing 2–5 µl protein solution as well as 18% PEG 8000, 0.1 *M* sodium cacodylate pH 6.5 and 200 m*M* zinc acetate.

2.2. Data collection

The data set from several single crystals, with typical dimensions of 0.3 × 0.4 × 0.7 mm, were collected at 291 K on an Enraf–Nonius FAST area-detector system coupled with a Rigaku RU-200 X-ray generator running at 40 kV and 70 mA. The data set was processed with the *MADNES* software (Messerschmidt & Pflugrath, 1987) and scaled with the Fourier scaling program (Weissman, 1982). In all, 88 563 observations of 26 402 unique structure amplitudes to 2.18 Å resolution were obtained. The $R_{\rm sym}$ is 6.8% for all data and 12.1% in the highest resolution shell (2.22–2.18 Å), with an overall completeness of 86.6% and 43.7% completeness in the last shell. The space group is $P2_1$, with two molecules in the asymmetric unit. The unit-cell parameters are a = 82.19 (8), b = 36.62 (4), c = 104.94 (9) Å, β = 113.5 (1)°. The calculated Matthews volume is 2.22 Å^3 Da^{-1} and the solvent content is 44.5%, both being comparable with those commonly observed for protein crystals.

2.3. Structure determination and refinement

The structure was determined by molecular-replacement methods using the program *CNS* (Brunger *et al.*, 1998). The coordinates of PP obtained from the Protein Data Bank (PDB code 3app) were used as a search model. In the initial rotational search, the PP models with and without the side-chain atoms produced one prominent peak instead of the two expected for the two molecules in the asymmetric unit. A correct solution could be obtained only when the comparatively modelled AP structure was used as a probe. The AP model was built by homology modelling based on the PP structure using the Swiss-Model server and has essentially the same backbone structure as PP. A search was carried out with data in the resolution range 10–3.5 Å and a centre of mass cutoff at 20 Å; the solution had a correlation coefficient of 0.48 and an R factor of 47.5%. A following rigid-body refinement, using 10.0 and 3.0 Å data, resulted in R and $R_{\rm free}$ values of 43.7 and 48.3%, respectively.

The refinement of the model was then continued using *CNS* with positional refinement followed by simulated annealing to 4000 K and finally an individual B-factor refinement. Manual rebuilding was carried out between each run with the σ_A-weighted $2F_o - F_c$ maps using the program *Xfit* from the *XtalView* package (McRee, 1992). Several cycles of positional refinement with non-crystallographic symmetry (NCS) restraints and group B factors were carried out, gradually extending the resolution to 2.18 Å; the R and $R_{\rm free}$ values converged to 32.3 and 43.6%, respectively. 50 cycles of model building and simulated-annealing refinements were then carried out with the gradual removal of NCS restraints and inclusion of individual atomic temperature factors. In the last ten cycles, solvent water molecules based on higher than 2σ peaks in the σ_A-weighted $F_o - F_c$ maps were added gradually and conservatively with regard for their environment, including potential interactions with hydrogen-bond partners. The solvent model was further comprehensively checked several times during the refinement by omitting all water molecules that had high B values (> 60 Å^2) or made either too close contacts with each other or with protein atoms, or too sharp an angle with potential hydrogen-bonding partners. During the refinement, eight electron-density peaks higher than 6σ consistently appeared in the $2F_o - F_c$ map as well as in the $F_o - F_c$ map. Initial attempts to refine them as water molecules failed, as the B factors became zero and the distances to the ligands became too short for a water molecule at these positions. These peaks were then interpreted as zinc ions that were essential for crystallization and could be refined resulting in lower R and $R_{\rm free}$ values and reasonable B factors

(18.5–38.1 Å^2) with decent coordination geometry. In both molecules in the asymmetric unit, extra electron densities extending from Ser60 and Ser235 were found, into each of which a mannose residue could be built and refined. The backbone geometry of several β-hairpins with relatively weak electron density was regularly checked against the structural database of β-hairpins of Sibanda *et al.* (1989). The value of $R_{\rm free}$, calculated with a randomly chosen 5% of the observed data for every refinement cycle, was consistently monitored to validate the refined structure. The crystallographic R factor at the end of the refinement was 21.5% for 25 869 reflections ($F > 2\sigma$) at 2.18 Å resolution; the $R_{\rm free}$ value in the second-last refinement cycle was 26.0% for 1243 reflections. The final coordinates have been deposited in the Protein Data Bank with accession code 1ibq.

3. Results and discussion

3.1. Description of the structure

AP crystallizes with two molecules in the asymmetric unit, designated *A* and *B*. Each AP molecule consists of 325 amino acids (2417 non-H atoms) and the previously unidentified two mannose residues. The refined model contains 431 water molecules and eight zinc ions in the asymmetric unit. The final electron-density map is, in general, well defined, as illustrated in the σ_A-weighted $2F_o - F_c$ map of the active-site region shown in Fig. 1. The root-mean-square (r.m.s.) coordinate error was estimated to be 0.30 Å from a Luzzati plot (Luzzati, 1952). The geometry of the final model checked with *PROCHECK* is good and the r.m.s. deviations from ideality for bond lengths and bond angles are 0.01 Å and 1.3°, respectively (Laskowski *et al.*, 1993). In the Ramachandran plot, taking into account only the non-glycine and non-N- or non-C-terminal residues, 475 residues (85.7%) are in the most favoured region, 74 residues (13.4%) are in the additionally allowed region and two residues (0.4%) are in the generously allowed region. Three residues (Asn11 and Ala116 of molecule *A* and Ala116 of molecule *B*) are in the disallowed region, with ψ angles slightly deviating from those for the allowed region. However, the electron density for these residues is well defined. Fig. 2 shows the average *B* factors of the main-chain atoms along the polypeptide chain, none of which exceeds 60 Å^2. The main-chain and side-chain protein atoms and the water molecules have mean *B* values of 25.9, 26.4 and 33.0 Å^2, respectively. There are no disordered regions even for the side chains.

AP has a characteristic two β-barrel domain aspartic proteinase fold with approximate overall dimensions of 35 × 45 × 65 Å (Fig. 3*a*). The mannose residues O-linked to Ser60 and Ser235 are located at the extreme ends of the two domains. The N- and the C-terminal domains both contain three β-sheets, designated 1N, 2N, 3N and 1C, 2C, 3C, respectively. The two domains are connected by a six-stranded β-sheet A that forms the backbone of the molecule. A deep cleft situated between domains is the substrate-binding site and each domain contributes a catalytic aspartic residue. Besides the β-strands forming the β-sheets, there are many small helices in AP as other secondary-structure elements. Molecule *A* contains six α-helices and three 3_{10}-helices, and molecule *B* contains five α-helices and three 3_{10}-helices. There is a disulfide bridge between Cys250 and Cys285 that is conserved in most of the aspartic proteinases. Pro133 and Pro317 of AP are in *cis* conformation, as are the equivalent residues in PP. In this respect, AP and PP differ from most mammalian and some fungal aspartic proteinases such as endothiapepsin (EP; Blundell *et al.*, 1990), rhizopuspepsin (RP; Suguna *et al.*, 1987), mucorpepsin (MP; Newman *et al.*, 1993) and yeast saccharopepsin (SP; Aguilar *et al.*, 1997), in

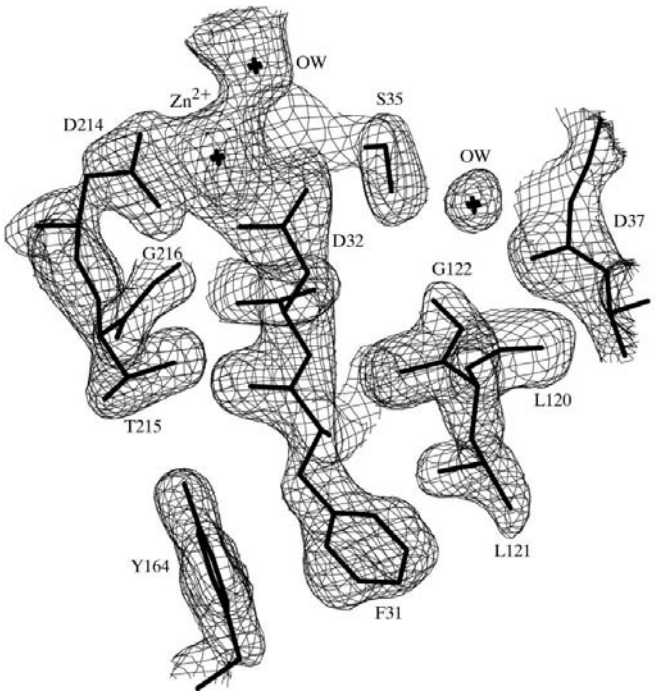

Figure 1
Final electron-density map superimposed with the final model showing a group of residues, water molecules and a zinc ion around the active site of molecule *A*. The σ_A-weighted $2F_o - F_c$ map calculated with the program *SIGMAA* (Read, 1986) was contoured at 2σ.

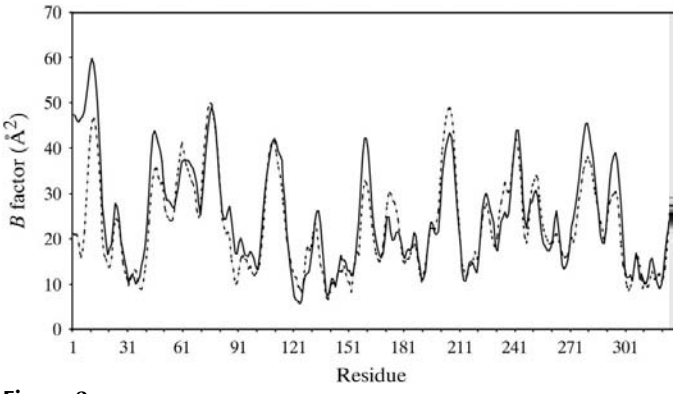

Figure 2
The average main-chain *B* factor along the polypeptide chain. The *B* factors for molecules *A* and *B* are denoted in solid and dashed lines, respectively.

which Pro23 (in pepsin numbering) is invariably in the *cis* conformation. EP and RP have an additional *cis* proline at positions 133 and 316, respectively. The fungal candidapepsin (CP) has no *cis* proline (Cutfield *et al.*, 1995).

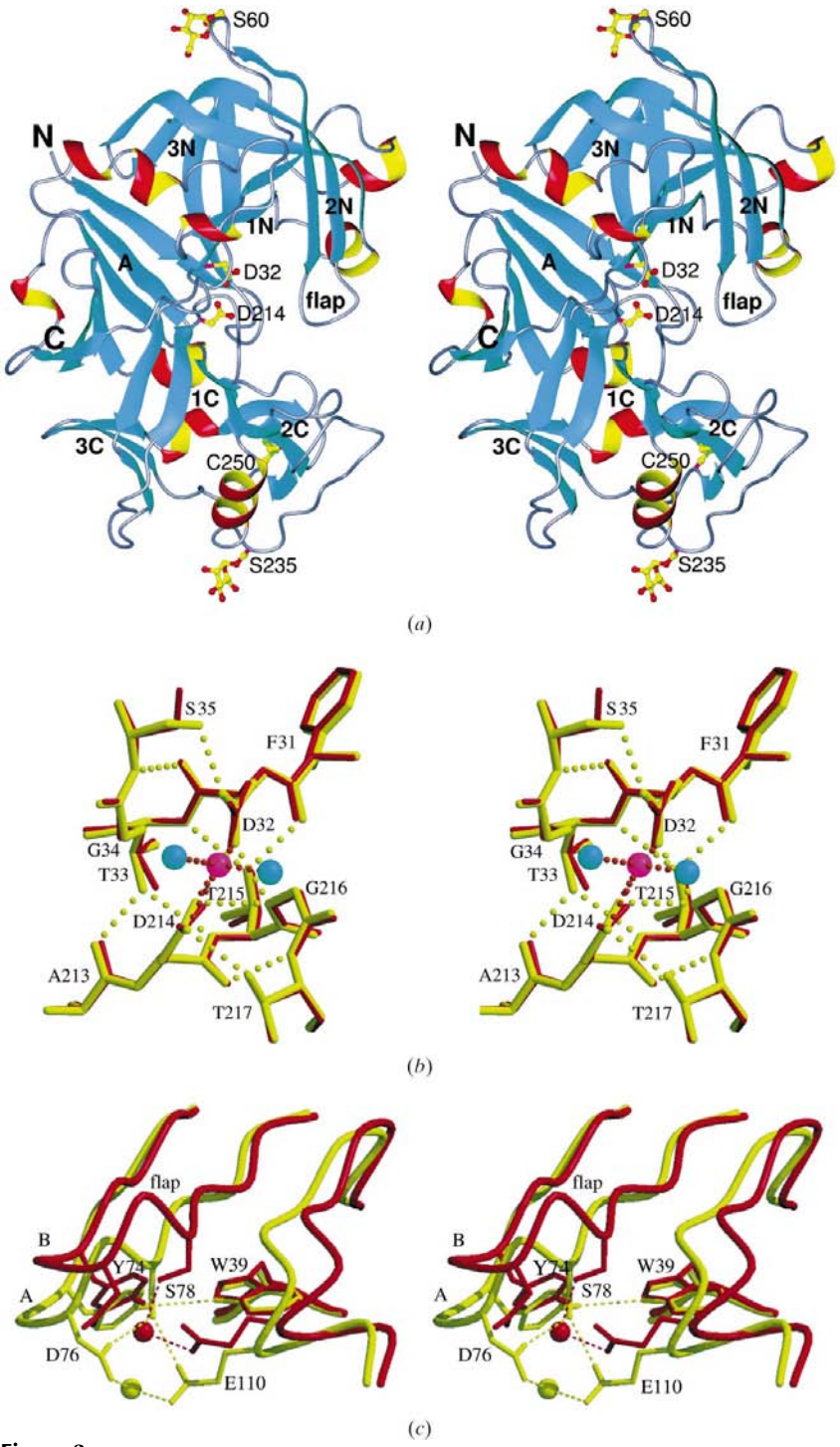

Figure 3
(*a*) A ribbon diagram of molecule *A*. The catalytic aspartates, the cysteine residues forming a disulfide bridge and the mannoses are indicated as ball-and-stick representations. (*b*) Superimposed active-site structures of the two independent AP molecules. The zinc ion is shown in magenta and the water molecules in cyan. (*c*) Superimposed flap structures showing different side-chain interactions. Molecules *A* and *B* are shown as thick lines in yellow and red, respectively, in both (*b*) and (*c*).

The asymmetric unit contains eight zinc ions that can be classified as four pairs of ions, each ion in a pair occupying the same site in each independent molecule. Two pairs of zinc ions are bound within a molecule and the other pairs mediate crystal contacts. The first pair is coordinated to His28, Asp54, Asp118 and a water molecule and the second to the catalytic Asp32 and Asp214 and two water molecules. The third pair is coordinated to Asp141 and Asp148 of molecule *A* (or *B*) and Ser1 of its symmetry-related molecule and the fourth to Asp256*A*, Glu232*B* (or Asp256*B*, Glu232*A*) and two water molecules. The coordination number and the distances to the ligands are compatible with those found for the zinc ions in the proteins (Alberts *et al.*, 1998). The present structure is interesting in that a zinc ion, though certainly not functional, is bound between the two catalytic aspartic residues while a catalytic water molecule is usually bound in the other aspartic proteinases (Fig. 3*b*). The water molecule is displaced from the plane formed by the two carboxyl groups, but the Zn^{2+} ion in AP is situated on the plane. The Zn^{2+} ion is coordinated to the two O^{δ} atoms of Asp214 (both 2.4 Å in *A* and 2.3 Å in *B*) but only to $O^{\delta1}$ of Asp32 (2.1 Å in *A* and 2.3 Å in *B*). These distances to the ligands are considerably short for a water molecule. The two O^{δ} atoms of Asp32 and Asp214 in the coordination sphere are separated by 2.5 Å as in the other aspartic proteinases. This close contact is usually regarded as a strong hydrogen bond, but in AP the possibility of it being a hydrogen bond is eliminated since both O atoms are coordinated to the metal ion. The coordination pattern and the hydrogen bonds involving these two aspartic residues are consistent with the previous proposal that Asp32 is protonated and the unprotonated Asp214 serves as a general base (Davies, 1990). It should be noted that all of the zinc ions affected neither the structures of the active site, the loops nor the flap.

3.2. Comparison of the two aspergillopepsin molecules

A stereoview showing the superposed C^{α} structures of the two AP molecules and other aspartic proteinases with known structure as well as the highly conserved residues is presented in Fig. 4. Two AP molecules are related by pseudo-twofold screw-axis symmetry and form a dimer mediated by several hydrogen bonds, including those

between the side-chain and main-chain atoms (Thr51*A* $O^{\gamma1}\cdots$ Gly236*B* O, 2.6 Å; Val112*A* O$\cdots$Gln238*B* $N^{\varepsilon2}$, 3.0 Å) and those between the side-chain atoms (Glu242*B* $O^{\varepsilon1}\cdots$ Asn11*A* $O^{\gamma1}$, 2.6 Å; Glu242*B* $O^{\varepsilon2}\cdots$Gln10*A* $N^{\varepsilon2}$, 2.9 Å). Superposition of the two molecules revealed that the 321 C^{α} atoms of 325 overlay with an r.m.s. deviation of 0.61 Å. However, they show some differences in detailed structures, exemplifying the flexible nature present in the aspartic proteinases.

The most significant difference is in the relative position of the three flanking loops at the surface of the N-terminal domain with respect to the remaining part of the molecule. These include a loop (Ser42–Gly52) at the extreme end of the N-domain, a β-hairpin (Tyr70–Gly82) called the 'flap' and a loop (Ala103–Asp114) containing a β-strand and a short α-helix. Three β-strands in the last two form β-sheet 2N. The backbone structures of these loop regions are nearly identical in the two molecules. However, the loops in molecule *A* are closer to the active site than those in molecule *B*, as if a hinged group motion has occurred. The largest separation of 4.0 Å occurs between the C^{α} atoms of Gly75 at the flap as shown in Fig. 3(*c*). This difference in the structure is related to different crystal contacts. It seems that the two surface loops in molecule *A* move inward owing to close contacts with a β-strand in sheet 3C of molecule *B*, further inducing a concerted movement of the flap. In contrast, these loops in molecule *B* are exposed to solvent. It has been found that the whole C-terminal domain as well as the flap can undergo a rigid-body movement when the substrates bind to EP, inducing a change in the shape of the active-site cleft (Šali *et al.*, 1992; Bailey & Cooper, 1994). The present structure shows that a similar movement is possible in the loop regions even without binding a substrate. As far as the flap is concerned, molecule *A* is more similar to the complexed aspartic proteinases, while molecule *B* is more similar to the native proteins in the open form.

The two independent molecules also show differences in the local structures; for instance, the interaction modes between Asp76 and Ser78 (Fig. 3*c*). These two residues on the active-site flap are conserved in the fungal proteins and play important roles in the recognition of a basic amino-acid residue at the P1 position (Shintani *et al.*, 1996). In all mammalian proteins that favour a hydrophobic P1 residue, Asp76 is substituted by Thr or Ser and Ser78 is deleted. It has been suggested that Asp76 interacts with P1 lysine, contributing to the transition-state stabilization, and that Ser78 is important for the proper orientation of the side chain of Asp76 (Shintani *et al.*, 1997). The hydrogen bond between Asp76 and Ser78 is indeed present in molecule *A* but not in *B*. In molecule *A*, these two residues as well as Glu110 at the adjacent α-helix and a water molecule form a cyclic hydrogen-bonding network such as Asp76$\cdots$Ser78$\cdots$Glu110$\cdots$water$\cdots$Asp76. In molecule *B*, there is no inter-residue hydrogen bond, but a water molecule is hydrogen bonded to Ser78 and Glu110.

3.3. Conserved residues in aspartic proteinases

Structure-based sequence alignment revealed that there are 23 strictly conserved residues in the 14 aspartic proteinases of known structure. They are depicted as large spheres in Fig. 4. Most of them are concentrated in the N-terminal domain and only four residues are in the C-terminal domain. Each of the conserved residues occupies a similar spatial position in all proteins. Ten residues of the 23 are glycines, occurring at the positions 23, 34, 77, 82, 119, 122, 167, 176, 216 and 301. These glycine residues invariably locate at the turn and serve as conformational determinants specifying the uniqueness of the fold. Gly75 in the flap, conserved with the exception of one serine, also has the same role.

At the active site, the Asp32, Thr33, Ser35 and Asp214 as well as Gly34 and Gly216 are strictly conserved. Thr215, which also constitutes the active site, is conserved with the exception of two serines, making the active-site motif denoted DT(S)G. Thr33 and Thr215 play an important role in the formation of the active site. Their hydroxyl groups connect the two ψ loops, each containing the catalytic Asp32 and Asp214, by forming hydrogen bonds with the main-chain N and O atoms in the opposite lobe beneath the catalytic dyad (Fig. 3*b*). This hydrogen-bonding motif is often called the 'fireman's grip' (Cooper *et al.*, 1990). It is generally accepted that the Ser35 residue assists the maintenance of the catalytic machinery by forming a hydrogen bond with Asp32. In molecule *A*, Ser35 O^{γ} is hydrogen bonded (2.8 Å) to Asp32 $O^{\delta2}$, which is not coordinated (2.8 Å) to the zinc ion. However, in molecule *B*, Ser35 has a different side-chain conformation and Ser35 O^{γ} is far (4.4 Å) from Asp32 $O^{\delta2}$, showing that the hydrogen bond may not always be present. The two O^{δ} atoms of Asp214, both coordinated to Zn^{2+}, make hydrogen bonds with Thr217 O^{γ} (2.8 Å in *A* and in *B*) and the amide Gly216 N (3.2 Å in *A* and 3.1 Å in *B*), as observed in other aspartic proteinases.

The other strictly conserved Tyr16, Trp39, Ser59, Tyr74, Asp87, Gln99 and Ala322 help maintain the stability of the fold. All of them, except for Ala322 in a β-strand near the C-terminal end, are involved in mediating two separate loops or strands through hydrogen bonds. Tyr16 O^{η} in a β-strand in the N-terminal domain forms a hydrogen bond (2.6 Å) with the main-chain Lys157 O in a β-strand in the central sheet *A*. The bulky Trp39 residue fills a large void in the N-terminal domain. Tyr74, which locates at the flap and makes a contact with the substrate P1 residue, is juxtaposed with Trp39. The O^{η}–N^{ε} distance is 3.1 Å in molecule *A* but 3.7 Å in molecule *B*, reflecting the difference in the relative positions of the flap (Fig. 3*c*). The role of Ser59, at the extreme tip of the N-terminal domain, is not obvious. Its O^{γ} atom is hydrogen bonded (2.6 Å) to Asp87 $O^{\delta2}$ in molecule *B*, but does not form a hydrogen bond in molecule *A*. Instead, Asp87 $O^{\delta2}$ forms a hydrogen bond with Tyr56 in molecule *A*. Asp87 $O^{\delta1}$ forms a hydrogen bond (2.6 Å) with the main-chain N of Ser59 in both molecules. Gln99 is hydrogen bonded to Thr134 O^{γ} and main-chain Thr137 O, holding two separate chains. There are 44 more residues that are highly conserved (depicted as small spheres in Fig. 4), the major variations being either Val/Leu/Ile or Phe/Tyr. Most of them seem to help maintain the stability of the fold by packing the hydrophobic core.

3.4. Comparison with other aspartic proteinase structures

The overall fold of AP is very similar to those of other aspartic proteinases as shown in Fig. 4. The three-dimensional structures are generally aligned better in the N-terminal domain than in the C-terminal domain, in accordance with the fact that the sequence variation in the former is smaller. The β-sheet structures which are important for the overall architecture overlay very well. Superposition of the equivalent C^{α} positions for molecule *A* gives r.m.s. values of 0.76, 0.96, 1.08, 1.20, 1.33 and 1.35 Å with PP (298 of 323 residues, sequence identity of 67.6%), EP (307/327, 55.7%), RP (284/325, 34.5%), SP (240/327, 25.8%), MP (242/325, 34.0%) and CP (244/342, 23.4%), respectively. The r.m.s. deviations with the mammalian aspartic proteinases are 1.25 Å for human pepsin (252/326, 30.1%; Fujinaga *et al.*, 1995), 1.26 Å for porcine pepsin (252/326, 30.1%; Sielecki *et al.*, 1990), 1.32 Å for human cathepsin D (251/326, 30.1%; Baldwin *et al.*, 1993), 1.36 Å for bovine chymosin B (259/326, 30.1%; Newman *et al.*, 1991), 1.32 Å for human renin (233/326, 30.1%; Dhanaraj *et al.*, 1992) and 1.26 Å for mouse renin (222/326, 30.1%; Dealwis *et al.*, 1994). The r.m.s. deviation is 1.28 Å for cardosin (254/325, 25.8%; Frazão *et al.*, 1999), the only plant aspartic proteinase of known structure.

These data indicate that the structural similarity generally parallels the sequence identity. AP is most structurally homologous to PP, EP and RP in that order and differs from other fungal enzymes as much as it does from the mammalian enzymes. Although the overall structures are similar to each other, the local structures are different in some loop regions. For instance, AP and PP show significant differences at the six loop regions Tyr56–Thr63, Lys115–Asn117, Gly196–Gly208, Gln238–Tyr246, Tyr275–Gly287 and Ser291–Ser298. The structures of AP and EP are different at the five loop regions Thr21–Thr26, Ser42–Gly52, Leu55–Ser60, Thr280–Cys285 and Ser314–Pro317, the third and the fourth overlapping with the regions where the structures of AP and PP are different. Most of these regions are far from the active site and thus the structural variations do not affect the essential features in catalysis. However, variations in the segment between Ser291 and Ser298 are important in that the loop structure is directly related to substrate specificity. This segment is a part of the so-called ψ loop and the structure of this loop determines the shape and the size of the S1′ and S2 subsites. Comparison of this part of the various aspartic proteinase structures revealed interesting correlations between the amino-acid sequence, the loop structure and the pocket size.

3.5. Variations in the ψ-loop structure and the subsite size in aspartic proteinases

In the C-terminal domain, there are two interleaved ψ-loops each connecting two β-strands in sheet 1C. One, designated ψ_{in}, contains a DTG motif at the base of the active-site cleft. The other, designated ψ_{out}, protrudes to the surface, partly covering the cleft. For the two ψ loops in the 14 aspartic proteinases of known structure, the aligned sequences based on the structure are shown in Table 1. In the ψ_{in} loop there is no insertion or deletion for all proteins regardless of their origins. The amino acids at the β-strands show variations while those at the turn constituting an active-site motif do not. The structures of this segment are nearly identical in all aspartic proteinases, being one of the most structurally conserved regions (Fig. 4).

A 14-residue segment between Gly288 and Gly301 comprises the ψ_{out} loop. In contrast to the ψ_{in} loop, the number of residues in this loop varies widely depending on the origin of the protein. Mammalian and plant aspartic proteinases have three or four more residues and the fungal CP has three fewer residues than AP. Conservation of the amino-acid sequence is noticeable only within the subfamily. AP is also most homologous to PP and then to EP in the sequence of this loop as well as in overall sequence. AP and PP differ by three residues, while AP and EP differ in six of the 14 residues. Surprisingly, however, the ψ_{out}-loop structure of AP is quite different from that of PP, but is rather similar to that of EP. These interesting structural variations are shown in Fig. 5(*a*), which shows the structures of the ψ_{out} loop of AP superposed with those of the PP–macrocycle (PDB code 1bxo) and EP–pepstatin (PDB code 4er2) complexes. In both complexed crystals, the ψ_{out}-loop structure is essentially the same as that in the native. The C^{α} positions of Ile289 and Ile299 are nearly in register in the three proteins, but the backbone structures of the nine residues in between show considerable variation.

The ψ_{out} loops of AP and EP have similar backbone structures but are slightly shifted from each other as a group. The maximum displacement between the C^{α} atoms of the equivalent residues is 2.0 Å. The turn is folded toward the binding site, limiting the accessible space for the P1′ and P2 residues. The side chains of Leu295 (AP) and Ile293 (EP), occupying the same site, locate at the tip of the turn and define the hydrophobic surface for the S1′ and S2 subsites. In contrast, the ψ_{out} loops of AP and PP have quite different backbone structures, especially for the five-residue segment at the turn (Ser293–Leu297 in AP and Ser291–Phe295 in PP). Ser291 is directed toward the binding site in PP, while the equivalent Ser293 is flipped outside in AP. The next four residues of PP make a turn receding from the active site. The largest separation between the equivalent C^{α} atoms is 5.0 Å for Gly294 (AP) and Gly292 (PP). Equivalent Ile293 (PP) and Leu295 (AP) protruding toward the active site are the hydrophobic residues that directly determine the shape of the binding site. Ile293 (PP) recedes from the active site further than Leu295 (AP), their C^{α} atoms being displaced by 3.7 Å. Ile293 (PP) is directed toward the P1′ residue, while Leu295 (AP) is directed toward the border between the P1′ and P2 residues. The side chains of Leu297 (AP) and the equivalent Phe295 (PP) occupy approximately the same site, although their C^{α} positions are displaced by 3.0 Å. The S1′ and S2 subsites are well defined and small in AP, while there is no definite border between S1′ and S2 and the open space for the S2 subsite is larger in PP. The superposed structures suggest that AP would not efficiently bind the rigid macrocyclic inhibitor that forms a complex with PP, since Leu295 $C^{\delta1}$ (AP) makes close contacts (less than 3.0 Å) with the macrocycle. In

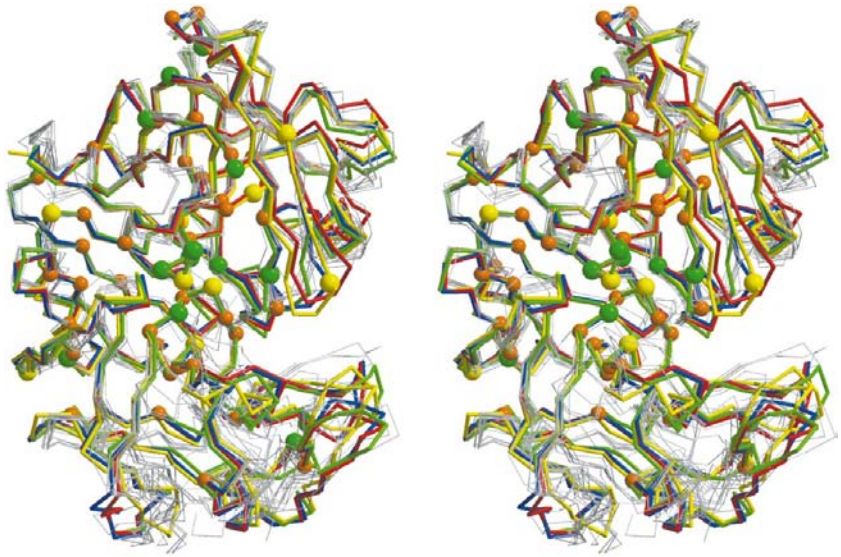

Figure 4
A stereoview of the superimposed C^{α} chains of aspartic proteinases with known structure. Molecules *A* and *B* of AP, PP and EP are shown as thick lines in blue, red, yellow and green, respectively, and those of the other aspartic proteinases listed in Table 1 are shown as thin lines in grey. 23 strictly conserved residues are shown as large spheres in yellow for ten glycines and in green for other residues. 44 highly conserved residues are shown as small spheres in orange.

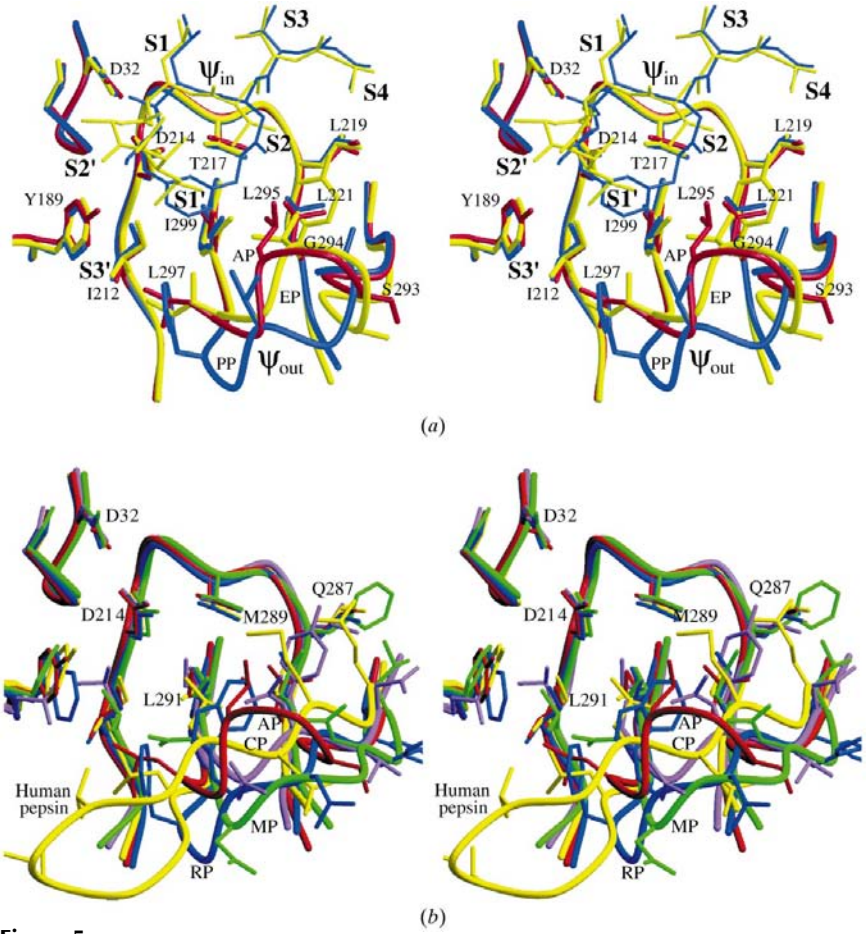

Figure 5
(*a*) Superimposed ψ-loop structures of AP, PP and EP: AP is in red, the PP–macrocycle complex is in blue and the EP–pepstatin complex is in yellow. The two ligands are drawn in thin lines. The labelled amino acids are those of AP. (*b*) Superimposed ψ-loop structures of AP, RP, MP, CP and human pepsin: AP in red, RP in blue, MP in green, CP in pink and human pepsin in yellow. The labelled amino acids are those of human pepsin.

order for AP to bind the macrocycle, the backbone structure of the ψ_{out} loop needs to be changed, which is not very likely to occur as the concerted rearrangements of many side chains costs considerable energy. It remains to be tested whether AP has the same binding ability towards the macrocyclic inhibitor as in PP. Although AP has a small S1′–S2 pocket, it can still accommodate a substrate containing a large hydrophobic P1′ or P2 residue, because the side chains can rotate toward the large S3′ or S4 site. It is not yet clear whether the structural differences in the S1′ and S2 subsites are related to the difference in milk-clotting properties of AP and PP. To our best knowledge, studies on the interactions between AP and various oligopeptides have not been reported; therefore, further biochemical data are needed to discuss the relationship between the subsite structure and substrate specificity in detail.

It is intriguing why the ψ_{out}-loop structure of PP differs from those of AP and EP contrary to the trend in sequence identity. It seems that Phe295 (PP), which is bulkier than the equivalent Leu297 (AP) and Ile299 (EP), plays an important role in this discrepancy. In AP, two methyl groups of Leu297 make van der Waals contacts (∼4.0 Å) with the side chains of Tyr189 at the loop connecting sheets A and 1C and Ile212 at the β-strand in the ψ_{in} loop (Fig. 5*a*). EP has similar contacts. However, in PP Phe295 lies between Phe190 and Ile293. If the ψ_{out} loop of PP had assumed the same backbone structure as that of AP, the side chain of Phe295 (PP) could not have the same orientation as that of Leu297 (AP) owing to collision with Phe190 and Ile211, which occupy essentially the same positions as the equivalent Tyr189 and Ile212 of AP, respectively. Thus, the phenyl ring had to rotate toward the solvent. This would result in the loss of the hydrophobic interactions. In order to avoid such a situation, PP seems to adopt a different conformation for the flexible loop segment containing two glycine residues. The integrity of the overall fold and the favourable hydrophobic interactions can thereby be maintained simultaneously. Validity of this explanation may easily be confirmed if the three-dimensional structure of aspergillopepsin F, a virulence factor in invasive aspergillosis, is deter-

Table 1
Structure-based sequence alignment of the two ψ loops.

The abbreviations are HP, human pepsin; pP, porcine pepsin; CD, cathepsin D; CB, chymosin B; HR, human renin; MR, mouse renin; Ca, cardosin. The strictly conserved residues are in bold. The bold PDB codes denote the structures of the protein complexed with a ligand.

PDB code	ψ_{in} loop	ψ_{out} loop
AP	210SAIA**DTG**TTLILL222	288GIQSNSGL....GLSIL**G**301
PP (3app)	209SGIA**DTG**TTLLL221	286GIQSNSGI....GFSIF**G**299
EP (4ape)	211DGIA**DTG**TTLLYL223	286GIQSSAGI....GINIF**G**299
RP (2apr)	214DAIL**DTG**TTLLIL226	288GFGYGNW.....GFAII**G**300
MP (1mpp)	211AFTI**DTG**TNFFIA223	285IVLPDGG.....NQFIV**G**297
CP (**1eag**)	214DVLL**DSG**TTITYL226	297LFDVN.......DANIL**G**307
HP (1psn)	211QAIV**DTG**TSLLTG223	285GFQGMNLPTESGELWIL**G**302
pP (4pep)	211QAIV**DTG**TSLLTG223	285GFEGMDVPTSSGELWIL**G**302
CD (1lya)	227EAIV**DTG**TSLMVG239	305GFMGMDIPPPSGPLWIL**G**312
CB (4cms)	211QAIL**DTG**TSKLVG223	285GFQSENH....SQKWIL**G**298
HR (1bbs)	211LALV**DTG**ASYISG223	285AIHAMDIPPPTGPTWAL**G**302
MR (**1smr**)	211EVVV**DTG**SSFISA223	285ALHAMDIPPPTGPVWVL**G**302
Ca (1b5f)	211QAFA**DSG**TSLLSG223	285GFTAMDAPL.LGPLWIL**G**301
SP (**2jxr**)	211GAAI**DTG**TSLITL223	285AITPMDFPEPVGPLAIV**G**302

mined. This protein shares an overall sequence identity of 67% with AP and differs from AP at four residues of 14 in the ψ_{out} loop (Lee & Kolattukudy, 1995). These include Arg355, His356, Phe364 and Phe367, which are equivalent to Gly288, Ile289, Leu297 and Leu300, respectively, in AP. We propose that the loop structure of aspergillopepsin F and the S1′–S2 subsite will resemble that of PP rather than AP, mainly as a consequence of the presence of the Phe364 residue at position 297 in AP. The other three residues of aspergillopepsin F would not affect the loop structure, since it is very likely that Arg355 points to the open surface and both His356 and Phe367 fill the void inside the C-terminal domain.

Comparison of the ψ_{out} loops with other aspartic proteinases revealed further interesting features (Fig. 5*b*). CP has three fewer amino acids than AP and its ψ_{out} loop makes a turn that is much shorter than the others. However, its binding pocket is similar in shape to, albeit slightly larger than, that of AP. In CP, the side chain of an Asn residue is located near the position of Leu295 (AP). RP contains one less amino acid than PP and a Phe residue equivalent to Phe295 (PP). The loop structure of RP most closely resembles that of PP. However, its S1′ and S2 subsites are very similar to those of AP, owing to Trp294 (RP) which is positioned at the site where Leu295 (AP) is located. MP has the same number of residues as RP and a similar, but shifted as a group, backbone structure to RP. In MP, a flexible Gln298 residue is positioned near the site where Trp294 (RP) is located but is directed oppositely. MP has a smaller subsite than PP, but a larger subsite than AP and RP. Although MP has a Phe299 residue in this loop, it is out of register by one residue from Phe295 (PP). These structural data indicate that the two amino acids equivalent to Leu295 and Leu297 of AP are the major determining factors in shaping the S1′ and S2 subsites in the fungal aspartic proteinases.

As far as the ψ_{out} loop is concerned, the aspartic proteinases from higher organisms differ from the fungal proteins in two aspects. Firstly, the former proteins have at least three more residues and thus longer loops (see Table 1) and are endowed with additional substrate specificities extended to the S3′ subsite and beyond compared with the fungal proteins. Secondly, the characteristics of the S1′ and S2 subsites are defined by three residues in the former proteins, while they are mostly defined by one residue in fungal proteins. A representative example, shown in Fig. 5(*b*), is human pepsin, in which Gln287, Met289 and Leu291 on the ψ_{out} loop are directed toward the active site shaping the substrate-binding subsites. Equivalent residues are Glu287, Met289 and Val291 in porcine pepsin, Met307, Met309 and Leu311 in human cathepsin D, His287, Met289 and Ile291 in human and mouse renins, Thr287, Met289 and Ala291 in plant cardosin, and Thr287, Met289 and Phe291 in yeast SP. In bovine chymosin B, Gln287 and Glu289 are in register but the third residue was not found in the electron density. The residues equivalent to Leu295 (AP) that determines the S1′ and S2 subsites in the fungal proteins are situated between the second and the third residues.

It is interesting to note that PP retains a larger pocket size by incorporating a single bulky residue that eventually induces a major structural change in the proximal flexible turn. This economic way to confer variability on substrate specificity seems unique to the fungal aspartic proteinases. It seems that in the higher organisms the insertion was instead exploited for elaborate tuning of the substrate specificity during evolution. The observations made in this study may illustrate that the knowledge of the precise three-dimensional structure is important in describing the substrate specificity of an aspartic proteinase. Homology modelling of the AP structure based on the structure of PP did not produce a correct structure as far as the ψ_{out} loop is concerned. The present study indicates that detailed analysis of structural cause and effect may be necessary for the high-resolution homology modelling of the loop regions in an aspartic proteinase.

This work was supported by a grant from the Korea Science and Engineering Foundation through the Center for Molecular Catalysis at Seoul National University. SWC was a recipient of a fellowship from the Brain Korea 21 Program.

References

Abita, J. P., Delaage, M., Lazdunski, M. & Savrda, J. (1969). *Eur. J. Biochem.* **8**, 314–324.
Aguilar, C. F., Cronin, N. B., Badasso, M., Dreyer, T., Newman, M. P., Cooper, J. B., Hoover, D. J., Wood, S. P., Johnson, M. S. & Blundell, T. L. (1997). *J. Mol. Biol.* **267**, 899–915.
Alberts, I. L., Nadassy, K. & Wodak, S. J. (1998). *Protein Sci.* **7**, 1700–1716.
Bailey, D. & Cooper, J. B. (1994). *Protein Sci.* **3**, 2129–2143.
Baldwin, E. T., Bhat, T. N., Gulnik, S., Hosur, M. V., Sowder, R. C., Cachau, R. E., Collins, J., Silva, A. M. & Erickson, J. W. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 6796–6800.
Barrett, A. J., Rawlings, N. D. & Woessner, J. F. (1998). *Handbook of Proteolytic Enzymes*, edited by A. J. Barrett, N. D. Rawlings & J. F. Woessner, pp. 801–805. San Diego: Academic Press.
Blundell, T. L., Jenkins, J. A., Sewell, B. T., Pearl, L. H., Cooper, J. B., Tickle, I. J., Veerapandian, B. & Wood, S. P. (1990). *J. Mol. Biol.* **211**, 919–941.

Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, I. T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Cooper, J. B., Khan, G., Taylor, G., Tickle, I. J. & Blundell, T. L. (1990). *J. Mol. Biol.* **214**, 199–222.
Cutfield, S. M., Dodson, E. J., Anderson, B. F., Moody, P. C., Marshall, C. J., Sullivan, P. A. & Cutfield, J. F. (1995). *Structure*, **3**, 1261–1271.
Davies, D. R. (1990). *Annu. Rev. Biophys. Biophys. Chem.* **19**, 189–215.
Dealwis, C. G., Frazão, C., Badasso, M., Cooper, J. B., Tickle, I. J., Driessen, H., Blundell, T. L., Murakami, K., Miyazaki, H., Sueiras-Diaz, J., Jones, D. M. & Szelke, M. (1994). *J. Mol. Biol.* **236**, 342–360.
Dhanaraj, V., Dealwis, C. G., Frazão, C., Badasso, M., Sibanda, B. L., Tickle, I. J., Cooper, J. B., Driessen, H. P., Newman, M., Aguilar, C., Wood, S. P., Blundell, T. L., Hobart, P. M., Geoghegan, K. F., Ammirati, M. J., Danley, D. E., O'Connor, B. A. & Hoover, D. J. (1992). *Nature (London)*, **357**, 466–472.
Dunn, B. M. & Hung, S. (2000). *Biochim. Biophys. Acta*, **1477**, 231–240.
Frazão, C., Bento, I., Costa, J., Soares, C. M., Veríssimo, P., Faro, C., Pires, E., Cooper, J. & Carrondo, M. A. (1999). *J. Biol. Chem.* **274**, 27694–27701.
Fujinaga, M., Chernaia, M. M., Tarasova, N. I., Bartlett, P. A., Hanson, J. E. & James, M. N. G. (2000). *Acta Cryst.* D**56**, 272–279.
Fujinaga, M., Chernaia, M. M., Tarasova, N. I., Mosimann, S. C. & James, M. N. (1995). *Protein Sci.* **4**, 960–972.
Ichishima, E. (1998). *Handbook of Proteolytic Enzymes*, edited by A. J. Barrett, N. D. Rawlings & J. F. Woessner, pp. 872–878. San Diego: Academic Press.
Ichishima, E. & Yoshida, F. (1965). *Biochim. Biophys. Acta*, **99**, 360–366.
James, M. N. G. & Sielecki, A. R. (1983). *J. Mol. Biol.* **163**, 299–361.
Khan, A. R., Parrish, J. C., Fraser, M. E., Smith, W. W., Bartlett, P. A. & James, M. N. G. (1998). *Biochemistry*, **37**, 16839–16845.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Lee, J. D. & Kolattukudy, P. E. (1995). *Infect. Immun.* **63**, 3796–3803.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
McRee, D. E. (1992). *J. Mol. Graph.* **10**, 44–46.
Messerschmidt, A. & Pflugrath, J. W. (1987). *J. Appl. Cryst.* **20**, 306–315.
Newman, M., Safro, M., Frazão, C., Khan, G., Zdanov, A., Tickle, I. J., Blundell, T. L. & Andreeva, N. (1991). *J. Mol. Biol.* **221**, 1295–1309.
Newman, M., Watson, F., Roychowdhury, P., Jones, H., Badasso, M., Cleasby, A., Wood, S. P., Tickle, I. J. & Blundell, T. L. (1993). *J. Mol. Biol.* **230**, 260–283.
Pitt, J. I. & Samson, R. A. (1990). *Modern Concepts in Penicillium and Aspergillus Classification*, edited by R. A. Samson & J. I. Pitt, pp. 3–13. New York: Plenum.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Šali, A., Veerapandian, B., Cooper, J. B., Moss, D. S., Hofmann, T. & Blundell, T. L. (1992). *Proteins Struct. Funct. Genet.* **12**, 158–170.
Scarborough, P. E. & Dunn, B. M. (1994). *Protein Eng.* **7**, 495–502.
Shintani, T. & Ichishima, E. (1994). *Biochim. Biophys. Acta*, **1204**, 257–264.
Shintani, T., Kobayashi, M. & Ichishima, E. (1996). *J. Biochem. (Tokyo)*, **120**, 974–981.
Shintani, T., Nomura, K. & Ichishima, E. (1997). *J. Biol. Chem.* **272**, 18855–18861.
Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). *J. Mol. Biol.* **206**, 759–777.
Sielecki, A. R., Fedorov, A. A., Boodhoo, A., Andreeva, N. S. & James, M. N. G. (1990). *J. Mol. Biol.* **214**, 143–170.
Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. & Davies, D. R. (1987). *J. Mol. Biol.* **196**, 877–900.
Weissman, L. (1982). *Computational Crystallography*, edited by D. Sayre, pp. 56–63. Oxford University Press.