Acta Cryst. (1998). D**54**, 32–46

Structure and Proposed Amino-Acid Sequence of a Pepsin from Atlantic Cod (*Gadus morhua*)

SOLVEIG KARLSEN,[a] EDWARD HOUGH[a]* AND RAGNAR L. OLSEN[b]

[a]*Protein Crystallography Group, Department of Chemistry, Faculty of Science, University of Tromsø, N-9037 Tromsø, Norway, and* [b]*Norwegian Institute of Fisheries and Aquaculture, N-9005 Tromsø, Norway.*
E-mail: edward.hough@chem.uit.no

(*Received 28 March 1997; accepted 29 May 1997*)

Abstract

The crystal structure of a pepsin from the gastric mucosa of Atlantic cod has been determined to 2.16 Å resolution. Data were collected on orthorhombic crystals with cell dimensions a = 35.98, b = 75.40 and c = 108.10 Å, on a FAST area-detector system. The phase problem was solved by the molecular-replacement method using porcine pepsin (PDB entry 5PEP) as a search model. The structure has been refined to a crystallographic R factor of 20.8% using all reflections between 8.0 and 2.16 Å, without prior knowedge of the primary sequence. The resulting crystal structure is very similar to the porcine enzyme, consisting of two domains with predominantly β-sheet structure in the same sequential positions as the enzyme from pig. In the course of the model building, 122 residues were substituted and two residues deleted from the starting model to give a polypeptide chain of 324 amino acids and a sequence identity of 57.7% with the pig pepsin. No carbohydrate residues were located. Sequence alignment with available aspartic proteinases, indicates that the fish enzyme seems to be more related to mammalian gastric pepsins than to the mammalian gastricsins and chymosins, lysosomal cathepsin D's and a pepsin from tuna fish. The amino-acid composition of the cod enzyme, however, is more in accordance with the cathepsin D's.

1. Introduction

Pepsins belong to the aspartic endopeptidase family and are digestive proteases present in the gastric juice of vertebrates. Mammalian pepsins are, in general, very well known and today two main forms are recognized (Foltman, 1981; Davies, 1990; Roberts, Peek, Keen & Taylor, 1995). Pepsin A (E.C. 3.4.23.1), often only denoted pepsin, is the predominant form. Gastricsin (E.C. 3.4.23.3), formerly named pepsin C, is reported to occur in significant amounts in species like man, monkey and rat (Tang & Wong, 1987). The third main aspartic proteinase in the stomach is chymosin (E.C. 3.4.23.4). This enzyme which is found in a number of newborn mammals, has weak general proteolytic activity, but can effectively partially hydrolyze casein and clot milk (Foltman & Axelsen, 1980). Multiple forms of gastric proteinases have also been found in fish species, including sardine (Noda & Murakami, 1981), tuna (Tanji, Kageyama & Takahashi, 1988) and capelin (Gildberg & Raa, 1983). Two aspartic proteinases, designated pepsin I and pepsin II, have been isolated from the gastric mucosa of Atlantic cod. The pI's are approximately pI 6.9 and 4.0, respectively. Pepsin II constitutes 80% of the total activity and consists of two forms, pepsin IIa and IIb (Gildberg, Olsen & Bjarnason, 1990).

Cold adaption of ectothermic organisms such as fish living in Arctic waters, involves compensations in the efficiency of enzyme-catalyzed reactions. This may be achieved by either simply producing more enzyme or increasing the catalytic efficiency (Hochacka & Somero, 1971; Hazel & Prosser, 1974). Pepsins in cold-water fish like Atlantic cod, exhibit a low Arrhenius activation energy, a high apparent Michaelis constant, a low temperature optimum, a low thermal stability and high pH optimum in comparison with pepsins from warm-blooded animals (Martinez & Olsen, 1989; Haard, 1992).

In recent years, a large number of reports have been published on the molecular structure of aspartic proteinases such as bovine chymosin (Newman *et al.*, 1991), human and porcine pepsin (Sielecki, Federov, Boodhoo, Andreeva & James, 1990; Abad-Zapatero, Rydel & Erickson, 1990; Cooper, Khan, Taylor, Tickle & Blundell, 1990; Fujinaga, Chernaia, Tarasova, Mosimann & James, 1995), human cathepsin D, a major intracellular proteinase (Gulnik, Baldwin, Tarasova & Erickson, 1992) and renin, a highly specific aspartic proteinase involved in the regulation of blood pressure (Sielecki *et al.*, 1989). Crystal structures have also been reported for fungal aspartic proteinases such as endothiapepsin (Blundell, Jenkins, Pearl & Sewell, 1985) and rhizopus-pepsin (Suguna *et al.*, 1987) and retroviral aspartic proteinases (retropepsin, E.C. 3.4.23.16) including enzymes from Rous sarcoma virus (Miller, Jaskolski, Mohana, Leis & Wlodawer, 1989) and human immunodeficiency virus (Navia *et al.*, 1989; Wlodawer *et al.*, 1989). Early structural studies of aspartic proteinases were carried out with the objective of understanding the relationship between structure and mechanism of the

Table 1. *Survey of the data-collection statistics for pepsin IIb from cod*

$R_{sym} = \sum_h \sum_i |I_{hi} - \langle I_h \rangle| / \sum_h \sum_i |\langle I_h \rangle|$, where I_{hi} is the intensity of the ith measurement of the same reflection and $\langle I_h \rangle$ is the mean observed intensity for that reflection.

d_{min} (Å)	No. of observations	No. of independent observations	Multiplicity	Completeness (%)	R_{sym}	$\langle I/sd \rangle$
6.73	1613	586	2.8	96.4	0.027	20.0
4.80	2983	986	3.0	98.9	0.032	19.4
3.93	3812	1248	3.1	99.3	0.037	18.2
3.41	4450	1465	3.0	100.0	0.046	15.3
3.05	4626	1580	2.9	96.4	0.074	10.1
2.79	4717	1655	2.9	92.0	0.112	6.8
2.58	4751	1735	2.7	89.6	0.149	5.1
2.42	4684	1756	2.7	84.6	0.182	4.2
2.28	4150	1728	2.4	78.5	0.215	3.4
2.16	1322	948	1.4	41.4	0.221	3.5
Totals	37108	13687	2.7	83.7	0.066	10.8

enzymes. In the last decade however, both pharmaceutical and commercial interests have boosted this work and made the aspartic endopeptidase family one of the most studied group of enzymes (Guruprasad, Dhanaraj, Groves & Blundell, 1994).

The relatively high pH optimum and high molecular activity at low reaction temperatures make fish pepsins excellent biocatalysts in processing of certain raw materials. A large amount of the fish waste produced by the industry in Norway is preserved by addition of formic acid to pH 3.9–4.3. At ambient temperatures, the minced waste is converted by endogenous pepsins to a hydrolyzed product which subsequently is defatted and concentrated. This hydrolyzed fish protein concentrate is succesfully used as a feed ingredient by the aquaculture industry. Pepsins are also isolated on a large scale from viscera of Atlantic cod (Almas, 1990) and apparently utilized in processing of certain fish products (Strøm & Raa, 1993; Haard, 1992).

In this paper we present the crystal structure and proposed amino-acid sequence of the pepsin IIb from the gastric mucosa of Atlantic cod. As far as we know, this is the first determined structure of an aspartic proteinase from fish.

2. Experimental

2.1. *Purification and crystalization*

Pepsin IIb was isolated from the gastric mucosa of Atlantic cod (*Gadus morhua*) and purified using ion-exchange chromatography as described by Gildberg, Olsen & Bjarnason (1990).

Orthorhombic crystals of pepsin IIb suitable for X-ray analysis, were grown at 277 K by vapour diffusion using the hanging-drop method (McPherson, 1982). Screening lead to the definition of conditions which consistently gave well defined microcrystals after a few days but a small number of crystals of diffraction quality and size (0.5 × 0.2 × 0.1 mm) were found two years later in several drops which had been prepared by mixing 5 µl protein (20 mg ml^{-1}) with 5 µl of a 1 ml reservoir solution consisting of 7.5% 2-propanol buffered at pH 5.4 with 100 m*M* sodium acetate.

2.2. *Data collection and processing*

Diffraction data were collected from the largest crystal at 291 K on an Enraf–Nonius FAST area detector using graphite-monochromated Cu $K\alpha$ radiation supplied by a FR-571 rotating-anode generator operated at 40 kV and 70 mA. Data frames were collected in 90° scans at $\chi = 0°$ and 45° with an exposure time of 50 s, 0.1° rotation in φ, crystal-to-detector distance 70 mm and a θ angle of −20°. The crystal remained stable during the data collection and diffracted to 2.16 Å.

The crystal form was orthorhombic with unit-cell parameters of 35.98, 75.40 and 108.10 Å. The space group was identified as $P2_12_12_1$ based upon symmetry and systematic absences. Data were collected and intially processed using the *MADNES* program package (Messerschmidt & Pflugrath, 1987), and integrations and profile fitting of the reflections were carried out with *PROCOR* in the *XDS* package (Kabsch, 1988). Merging and scaling to a unique data set were performed with the program *SCALA* and structure factors were calculated by *TRUNCATE* in the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994). 37 108 usable observations were merged to 13 687 unique reflections (R_{merge} 6.6%) and a completeness of 84% from 32.50 to 2.16 Å resolution. Selected data-collection statistics are given in Table 1. A Wilson plot (Wilson, 1949) of the final data gave an overall temperature factor of 22.3 Å^2.

2.3. *Molecular-replacement studies*

The structure was solved by molecular replacement using the *X-PLOR* program package (Brünger, 1992). The search model, the 2.3 Å refined structure of porcine pepsin (Cooper *et al.*, 1990, PDB entry 5PEP), was placed in an artificial *P*1 unit cell with cell dimensions

Table 2. *Summary of PROLSQ refinement statistics for the final set of coordinates*

Space group	$P2_12_12_1$
Resolution range (Å)	8.00–2.16
R factor†	0.208
R_{free} value‡	0.224
No. of atoms	2551
No. of reflections	13315
Min. B value (Å^2)	5.00
Max. B value (Å^2)	62.51
Mean B value (Å^2) (all atoms)	17.22
Model	
No. of protein residues	324
No. of water molecules	161
Mean positional error (Å) from Luzzati plot (Luzzati, 1952)	0.20–0.30
R.m.s. deviations from ideal values	
Bond distances (Å)	0.020
Angle distances (Å)	0.071
Planar 1–4 distances (Å)	0.091
Chiral centers (Å^3)	0.196
Planar groups (Å)	0.018

† $R = \sum ||F_{obs}(hkl)| - |F_{calc}(hkl)|| / \sum |F_{obs}(hkl)|$. ‡ R factor calculated for 10% of data (Brünger, 1993).

$a = b = 150$, $c = 200$ Å and $a = b = c = 90°$. The rotation search was carried out using reflections with $F > 2\sigma(F)$ in the resolution range from 15 to 4 Å and the Patterson map was calculated with a 0.25 Å grid. The Euler angles (θ_1, θ_2 and θ_3) were rotated in steps of 1.1° in the asymmetric unit as desribed by Rao, Jih & Hartsuck (1980).

For the 38 highest rotation peaks, Patterson correlation (PC) refinements gave one clear solution, $\theta_1 = 165.1$, $\theta_2 = 25.6$ and $\theta_3 = 156.8°$. A translation search was initiated on this peak using the same resolution range as for the rotation function. The translation distances in fractional coordinates in the x, y and z directions for the model molecule in its initial position equaled 0.481, 0.385 and 0.442, respectively. The crystallographic R factor based on 9687 reflections [$F > 2\sigma(F)$] in the resolution range from 8 to 2.5 Å was estimated at 51.3%.

Rigid-body refinement in the *X-PLOR* package using data between 15 and 3.5 Å and 200 cycles of positional refinement in the resolution range of 8–2.5 Å (9710 reflections), lowered the R value to 34%. Finally, the refinement was accomplished with the simulated-annealing procedure (Brünger, Kuriyan & Karplus, 1987) using a temperature of 3000 K, followed by 200 cycles of traditional refinement on the atomic positions. The final crystallographic R factor for the molecular-replacement solution was 29.7% for $F > 1\sigma(F)$ between 8 and 2.5 Å resolution. Only ten intermolecular contacts shorter than 3.5 Å were found between symmetry-equivalent molecules in the orthorhombic unit cell. These contacts were between residues Phe64-Glu65 and Glu240, Glu65 and Ser242, Pro293 and residues Asn158 and Asp160, and finally Ser295 and Asp160.

2.4. *Model building and refinement*

Refinement of the structure of cod pepsin was carried out by simulated annealing (Brünger *et al.*, 1987) using the programs in the *X-PLOR* package (Brünger, 1992) combined with restrained least-squares refinement using *PROLSQ* (Hendrickson, 1985) in the *CCP*4 package (Collaborative Computational Project, Number 4, 1994). The model was systematically improved through iterative cycles of crystallographic refinement and manual rebuilding, with a stepwise increased resolution from 2.50 to 2.16 Å. Model fitting was carried out by inspection of both $F_o - F_c$ and $2F_o - F_c$ electron-density maps, supported by data from aligned primary sequences of other acid proteases as shown in Fig. 6. For uncertain regions in the structure omit maps were calculated and the model was built into omit density after being omitted from the refinement and map calculations. After each round of model building and refinement a real-space correlation plot (Jones & Kjeldgaard, 1993) was calculated to evaluate how well each residue fitted the electron density. The program *PROCHECK* (Laskowski, MacArthur, Moss & Thornton, 1993) was used throughout the model building to check the stereochemistry of the structure and to indicate regions needing further investigation, followed by manual adjustment on the graphics screen.

Before initial inspection of the electron-density maps and manual adjustments of the model, one round of simulated annealing (SA) was carried out in the *X-PLOR* package (Brünger *et al.*, 1987). This procedure lowered the crystallographic value from 34.0 to 29.7%. Further refinement steps were carried out by the convential least-squares method by alternating *X-PLOR* and *PROLSQ* refinements, resulting in a final R factor of 20.8% and an R_{free} value of 22.4 (10% of the data) for data between 8.0 and 2.16 Å (no σ cutoff) as seen in Table 2.

Water molecules were located using the *PEAKMAX* routine in the *CCP*4 package (Collaborative Computational Project, Number 4, 1994). Peaks above the 3σ level in both $F_o - F_c$ and in the corresponding $2F_o - F_c$ maps where one or more hydrogen bonds (3.3 Å distance) could be identified to other atoms, were accepted as water molecules. Each potential water molecule was manually checked on the graphics for reasonable hydrogen bonding and well defined electron density. Water molecules with temperature factors greater than 60 Å^2 were omitted from the structure.

The pepsin model makes only nine intermolecular contacts shorter than 3.5 Å to the four symmetry-equivalent molecules in the unit cell, and only two of these, between NH_2 in Arg59 and O in Ala237 and OG in Ser238, are shorter than 3.0 Å. Fig. 1 illustrates the packing of the pepsin molecules in the orthorhombic unit cell. The water content in the unit cell is estimated to be 43% according to the method of Matthews (1968).

The coordinates and structure factors for the structure of the cod pepsin have been deposited with the Protein Data Bank.†

3. Results and discussion

3.1. *Quality of the model*

The model of the pepsin structure contains 324 amino acids and 161 water molecules making a total of 2551 atoms. The observed number of residues is not in accordance with earlier analyses of amino-acid composition (Gildberg, Olsen & Bjarnason, 1990) which indicated one residue less. The crystallographic R factor is 20.8% for 13 315 reflections in the resolution range from 8.0 to 2.16 Å, and r.m.s. deviations from ideal geometry for bond lengths, angle distances and planar groups are 0.020, 0.071 and 0.018 Å, respectively (see Table 2). The estimated coordinate error according to Luzzati (Luzzati, 1952) is in the the range 0.20–0.30 Å.

PROCHECK (Laskowski *et al.*, 1993) shows that 84.7% of the residues, excluding glycines and prolines, are in the most energetically favoured regions of the Ramachandran plot (see Fig. 2). 13.0% (34 residues) are found in additional allowed areas, whereas six amino acids (Thr77, Ala204, Ala205, Ala208, Ala250 and Thr295) localized in generously allowed regions, are all in flexible loops with diffuse density and high temperature factors. The number of residues in the pepsin model showing unfavourable φ and ψ angles, is higher than for most structures determined to medium resolution where the amino-acid sequence is known.

3.2. *The correlation coefficients and temperature factors*

As described by Jones & Kjeldgaard (1993), the real-space correlation plot, given in Fig. 3(*a*), allows one to evaluate how well each residue fits the density. Side chains with well defined density have, in most cases, correlation coefficients greater than 0.8, while the values for those in poorly defined density are lower. In the pepsin structure 62 of the 324 residues (19.2%) have a correlation value lower than 0.8, higher than normally observed for structures with known amino-acid sequences. Many of the residues with poor density fit are arginines, glutamines and lysines located in loop regions on the surface of the pepsin molecule with a great deal of conformational freedom.

The same trend is seen in Fig. 3(*b*) which illustrates a temperature-factor plot for the pepsin structure. The overall mean temperature factors for main-chain and side-chain atoms are 15.8 and 17.0 Å^2, respectively. The B factor and the real-space correlation plots illustrate a close correlation between high B factor and poor density fit for surface residues. This feature is seen for residues 73–79 and residues 107–113 in loops between β-strands e and f, and g and h, respectively (see Fig. 10*b*), and residues 202–210 and 277–281 in loops between l_2 and m, and n_5 and n_6, respectively (see Fig. 10*c*). The first of these (73–79) is described by Davies (1990) as a flexible 'flap' which projects over the active-site cleft. Another region in the structure at residues 225–255, which also includes several β-strands and one α-helix, is more mobile than the rest of the pepsin molecule. The temperature-factor plot illustrates that the N- and C-terminal ends of the enzyme are not particularly flexible.

3.3. *Interpretation of electron-density maps*

The interpretation of electron-density maps has been carried out without prior knowledge of the primary structure of the enzyme. However, available sequence information from different aspartic proteinases as seen in Fig. 6, in combination with experimental data to 2.16 Å, has made it possible to deduce the structure with reasonable accuracy.

Exept for three short loops (74–79, 108–112 and 203–206), the entire main chain is well defined in the electron-density maps, and most of the side chains can be determined from shape and size of the electron density. As it is difficult at this resolution to distinguish leucine, aspartic acid and aspargine from each other, and glutamic acid from its corresponding amide there will always be some element of uncertainty in the crystallographic sequence determination. Unique identification of some aromatic amino acids has also been difficult. In these cases, the choice of residue was always based on sequence information from other aspartic proteinases.

In the early stages of the structure determination of the cod pepsin, no complete sequence of a fish pepsin was reported, but N-terminal sequences of pepsin I and IIa from cod and and a pepsin from tuna fish were available (Gildberg *et al.*, 1990; Tanji *et al.*, 1988). From earlier characterizations of the pepsin from atlantic cod in catalytic properties and chemical compostion, the IIa and IIb enzymes were expected to be very similar, both in primary and three-dimensional structure. Therefore, it was natural to look at the N-terminal sequence of pepsin IIa in the building of this part of pepsin IIb. In the latter stages of the model building, the complete primary sequence of pepsin 2 from tuna fish became available (Tanji, Yakabe, Kageyama & Takahashi, 1996) and was useful in the final choices of amino acids in the cod pepsin sequence.

All in all, we have substituted 122 and deleted two of the 326 residues in the initial pig pepsin model. The deleted residues are Ser207 and Ser241 in two loops (see Figs. 4*a* and 4*b*) which gives a cod pepsin with a polypeptide containing 324 residues. The number of

† Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory (Reference: 1AM5, R1AM5SF). Free copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England (Reference: SE0214).

residues is one more than determined earlier in an analysis of amino-acid composition of pepsin IIa (Gildberg *et al.*, 1990), but it is possible that an extra residue should be deleted in one of the flexible loops with poorly defined density. The proposed sequence for cod pepsin IIb, seen in Fig. 6, shows 57.7% sequence identity with the pig enzyme. As there is, however, always a risk of reproducing the starting model (porcine pepsin) in the electron-density maps, the number of identical residues could be slightly too high.

As mentioned above, two N-terminal sequence segments of pepsin IIa have been determined earlier by the Edman method (Gildberg *et al.*, 1990). The second part of these, residues 16–32, is well conserved and agrees well with the observed electron density in this region. In the first segment however, some differences between pepsin IIa and the density map for IIb are seen. As in the N-terminal sequence analysis, residues 3, 4, 6 and 8 are clearly Thr, Glu, Met and Asn, respectively. The non-determined residue in position 5 seems to be Gln as seen for the cod pepsin I, and a lysine as seen for porcine cathepsin D, rather than the threonine found in pepsin IIa, fits the electron density in position 7. Observed electron density for residues 3–5 and 6–8 is shown in Figs. 5(*a*) and 5(*b*), respectively. The remainder of this sequence fragment (residues 9–12), show some sequence identity with the primary structure of the fish pepsins (D-A-D-L) but the cod pepsin IIb seems to have a larger acidic residue, glutamic acid, in postition 9 and a threonine in position 12 as found in the porcine pepsin.

3.4. *The amino-acid composition*

The composition and number of amino acids interpreted from the electron-density map and seen in Table 3, agrees well with earlier analysis of pepsins IIa and IIb (Martinez & Olsen, 1989; Gildberg *et al.*, 1990), although differences are observed for some amino acids. However, the amino-acid composition varies from analysis to analysis and can, therefore, only be a guide in the course of model building.

In our model, the number of histidines, which is extremely low in the pig pepsin (only one), lysines (four in porcine pepsin) and arginines (two in porcine pepsin), are found to be 4, 11 and 5, respectively, in the cod pepsin IIb, making the fish enzyme more basic than the mammalian one. Table 3 shows that tuna fish pepsin 2 and porcine cathepsin D contain 13 and 34 basic amino acids, respectively, and have a closer resemblance in amino-acid composition with the two cod pepsins than with the porcine pepsin. These differences in acidicity are also reflected in the p*I* values for the aspartic proteinases. The content of glycines in the cod pepsins IIa and IIb is higher (39 and 38, respectively) than in the porcine pepsin (35) and cathepsin D (31), which could effect the flexiblility of the cold-adapted fish enzymes. The number of S atoms determined from the X-ray structure of pepsin IIb from cod is six, but earlier analyis of the amino-acid composition of pepsin IIa from cod (Gildberg *et al.*, 1990) indicated eight cysteines for this pepsin. The tuna fish pepsin 2 and porcine pepsin both contain six cysteines but the cathepsin D from pig has eight like cod pepsin IIa.

Table 3. *Amino-acid composition*

Evaluated from the X-ray structure of cod pepsin IIb, analysis of cod pepsin IIa (Gildberg *et al.*, 1990), and sequence analysis of tuna fish (Tanji *et al.*, 1996) and pepsin (Tsukagoshi *et al.*, 1988) and cathepsin D from pig (Cunningham & Tang, 1976). The basic amino acids (arginines, histidines and lysines) are shown in bold and the glycines and cysteines are in italic and bold italic, respectively.

Amino acid	Cod pepsin IIb	Cod pepsin IIa	Tuna fish pepsin 2	Porcine pepsin	Porcine cathepsin D
Ala	27	23	19	17	25
Arg	**5**	**6**	**5**	**2**	**11**
Asn	17	†	17	13	†
Asp	16	30	16	28	29
Cys	***6***	***8***	***6***	***6***	***8***
Gln	15	†	30	13	†
Glu	16	33	6	13	32
Gly	*39*	*38*	*34*	*35*	*31*
His	**4**	**4**	**3**	**1**	**5**
Ile	18	18	23	25	15
Leu	17	14	15	26	27
Lys	**11**	**11**	**5**	**1**	**18**
Met	7	8	8	4	7
Phe	13	13	15	14	10
Pro	17	21	13	15	21
Ser	33	33	38	44	32
Thr	23	21	23	26	18
Trp	4	4	6	5	4
Tyr	12	10	15	16	13
Val	24	28	26	22	23
No of residues	324	323	323	326	329
No of basic residues	20	21	13	4	34
No of S–S linkages	3	4	3	3	4

† No amino-acid determination.

Earlier analysis of the carbohydrate content of the cod pepsins revealed that they all are glycoproteins with an approximate carbohydrate content of 1, 2 and 5% for pepsin I, IIa and IIb, respectively (Gildberg *et al.*, 1990). The latter was found to be similar to the carbohydrate content of cathepsin D from porcine spleen (Huang, Huang & Tang, 1979), indicating that the cod pepsins are closer to the mammalian cathepsins than the mammalian pepsins. Electron density which could be attributed to bound carbohydrate residues was not found in the 2.16 Å maps but this could be thermal movement or disorder. Sequence analysis of pepsin 2 from the tuna fish showed that this fish enzyme does not contain any *N* glycosylation sites (Tanji *et al.*, 1996).

3.5. *The amino-acid sequence*

In order to search for conserved amino acids among the aspartic proteinases which could guide us in the

building of the pepsin model, the sequences (see Fig. 6) of 12 pepsins, chymosins and cathepsins D from different species were aligned using the program *MegAlign* (DNAstar Inc.). The alignment suggests that, although earlier N-terminal sequence analysis implied that the cod pepsins may be equally closely related to mammalian cathepsin D as to mammalian pepsins (Gildberg *et al.*, 1990), cod pepsin IIb (the last aligned sequence) has higher sequence identity with the pepsin A's (51–58%) than with the pepsins C's (44–45%), chymosines (46–47%) and lysosomal cathepsin D's (43%). The sequence identity between the cod enzyme and tuna fish pepsin 2 (50.5%) are lower than for the pepsin A's but higher than for the cathepsins D's. On the other hand, the comparison of the amino-acid content seen in Table 3, reveals that the porcine cathepsin D and tuna fish pepsin 2 are closer related to the cod pepsin IIb than the porcine pepsin.

A comparison of the amino-acid sequences of the regions in the catalytic sites, shows that many postions are well conserved among the aspartic proteinases. There are however, some subsitutions that are unique for the fish pepsins; Pro110, Ala113 and Gly287. The tetrapeptide sequence of residues 110–113 appears to be one of the most variable regions among the subsites of the aspartic proteinases, but is almost unique to each related group of enzymes. This feature is also seen for the fish enzymes where the sequences for the cod and tuna fish pepsins are PFQA and PFMA, respectively, and only differ in position 112.

Two segments of four amino acids in the aspartic proteinases are completely conserved. The first is Asp-Thr-Gly-Ser at residues 32–35 in the N-terminal moiety, and the other at residues 215–218 (Asp-Thr-Gly-Thr) in the C-terminal lobe. The aspartic acids in these two segments, Asp32 and Asp215, are believed to play an important role in the catalytic mechanism of aspartic proteinases (in James & Sielecki, 1987). This extraordinary sequence homology suggests strongly that the two domains in aspartic proteinases have envolved by gene dublication. Further support for this hypothesis is seen in the highly conserved regions of residues 117–122 (Phe-Asp-Gly-**Ile-Leu-Gly**) in the N-terminal part and residues 300–305 (**Ile-Leu-Gly**-Asp-Val-Phe) in the C terminus, where the common Ile-Leu-Gly segment is found in almost every analysed sequence. In the cod pepsin, however, an **Ile-Phe-Gly** sequence is found at residues 300–302.

Apart from the cathepsin D's which have eight cysteines, all gastric and fish enzymes in the sequence alignment have six conserved cysteins which form three disulfide bridges. Only two cysteines, 249 and 282, are completely conserved in all enzymes. Another typical feature which can be seen from the alignment, is that the

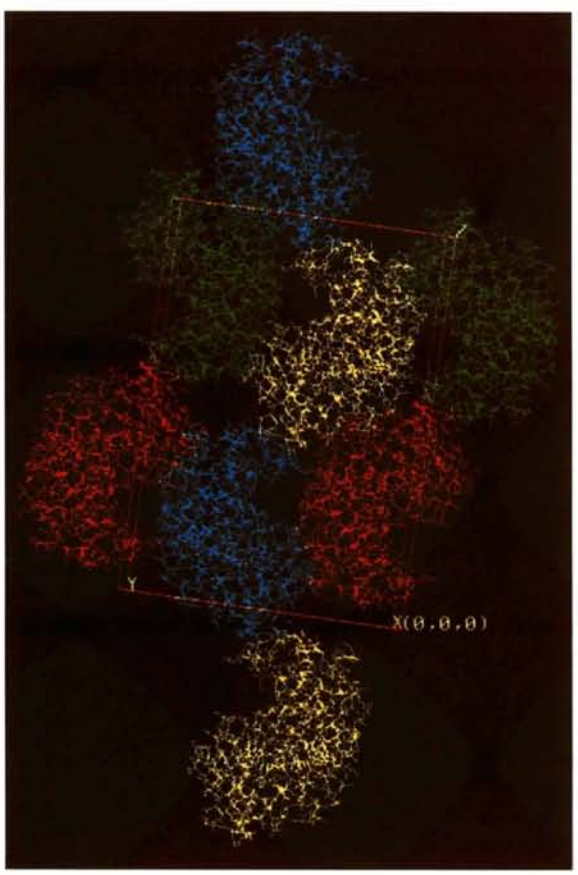

Fig. 1. The packing of pepsin molecules in the orthorhombic unit cell. Symmetry-equivalent molecules are coloured as following: x, y, z (red), $\frac{1}{2}-x$, $-y$, $\frac{1}{2}+z$ (green), $\frac{1}{2}+x$, $\frac{1}{2}-y$, $-z$ (yellow), $-x$, $\frac{1}{2}+y$, $\frac{1}{2}-z$ (blue).

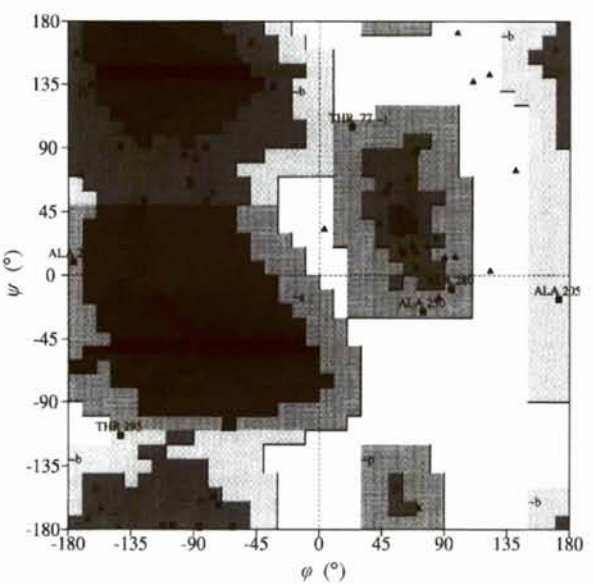

Fig. 2. The ramachandran plot (Ramachandran & Sasisekharan, 1968) of the main-chain torsional angles for the model of pepsin IIb from cod. Glycine residues are shown as triangles and all other residues as squares. The plot was produced with *PROCHECK* (Laskowski *et al.*, 1993).

conserved amino acids are mostly aromatic residues and glycines, such as Tyr14, Gly21, Phe27, Phe31, Trp39, Tyr75, Gly76, Gly78, Gly82, Gly119, Gly122, Phe151, Gly168, Tyr189, Trp190, Gly243, Gly285, Trp299, Phe305, Tyr309 and Gly321. The stacking interactions combined with hydrogen bonds make these aromatic residues very important for protein stability and probably for substrate binding, whereas glycines permit greater conformation freedom of the polypeptide chain, a feature which is often associated with the psychrophilic nature of many fish enzymes (Haard, 1992).

In Fig. 7 the amino-acid sequence of the pepsin IIb from cod is compared with the pepsin sequences from pig and tuna fish and areas with secondary-structure elements are marked on the top. The alignment shows that areas with β-sheet and α-helices are, on average,

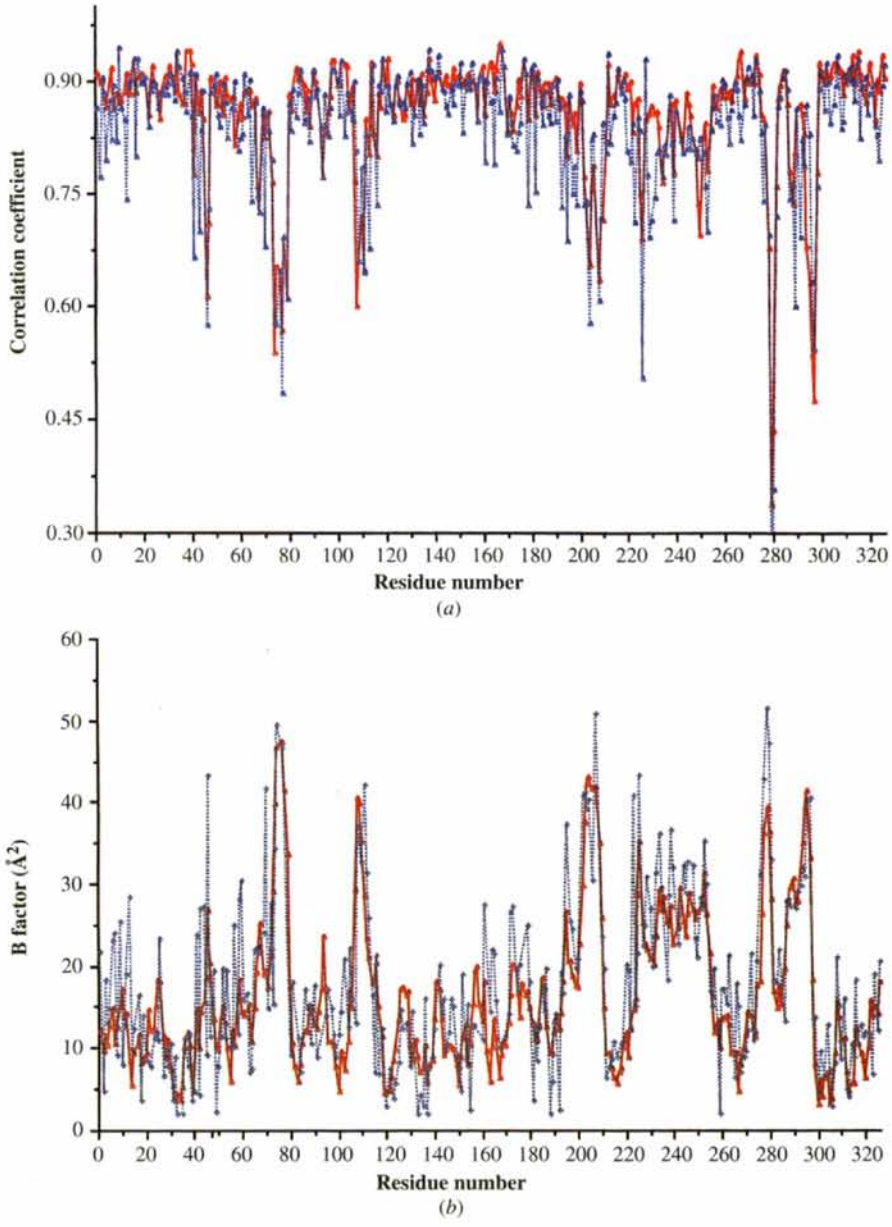

Fig. 3. Variation of correlation coefficients along the polypeptide chain of the cod pepsin. (*a*) Real-space correlation coefficient. The main-chain values are shown as a red solid line and the side-chain values are represented with a blue dashed line. The values for A_o and c (Jones & Kjeldgaard, 1993) were 0.90 and 1.04, respectively. (*b*) Mean crystallographic temperature factors averaged over main-chain (red solid line) and side-chain atoms (blue dashed line) respectively, for each residue in the pepsin structure.

more conserved than the loops. However, short conserved loops can also be seen in residues 21–25, 33–37, 61–64, 75–78 in the 'flap', 91–95, 123–125, 154–156 and 216–218. In the figure the residues common to the cod pepsin and the pig and tuna fish pepsins are marked in green and blue, respectively.

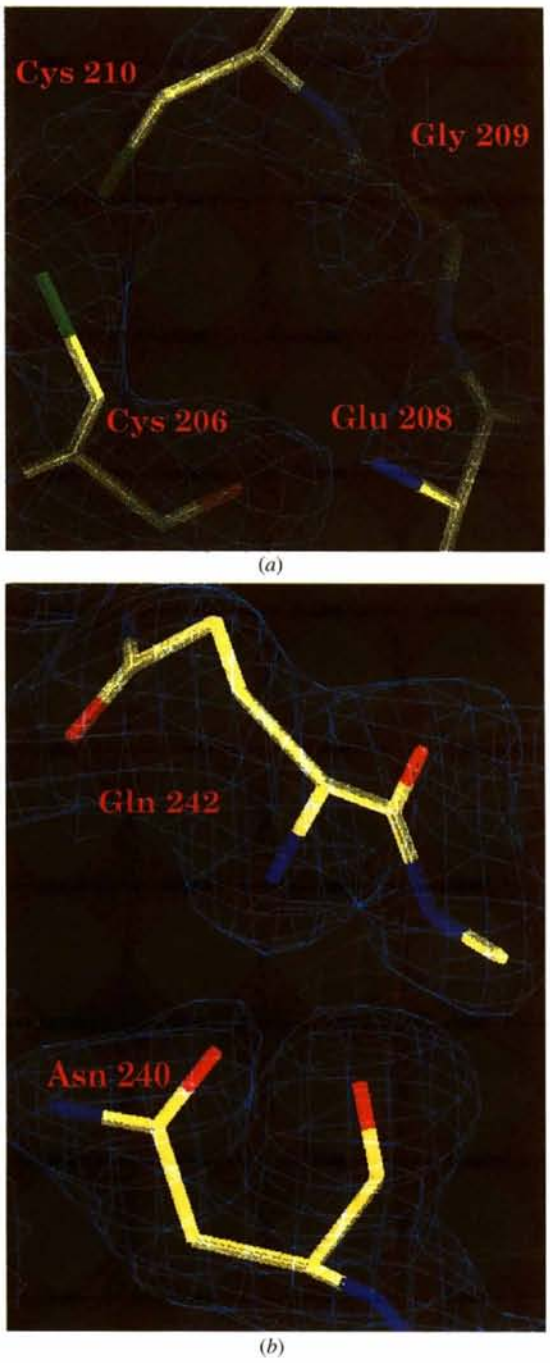

Fig. 4. Observed electon density (blue) contoured at 1σ level, for the cod pepsin showing that (*a*) residue 207 and (*b*) residue 241 are absent in the cod pepsin.

3.6. *The overall fold and secondary structure*

The three-dimensional structure of pepsin from cod, shown in Figs. 8 and 9, is very similar to previously reported structures of pepsin-like single-chain proteinases (in Davies, 1990). The enzyme has bilobal general shape of a croissant with approximate dimensions of 65 × 50 × 40 Å with two domains (N and C) of similar structure which are related by a twofold rotation axis and separated by a deep substrate-binding cleft. A less precise twofold symmetry has been noted within the N- and C-terminal domains.

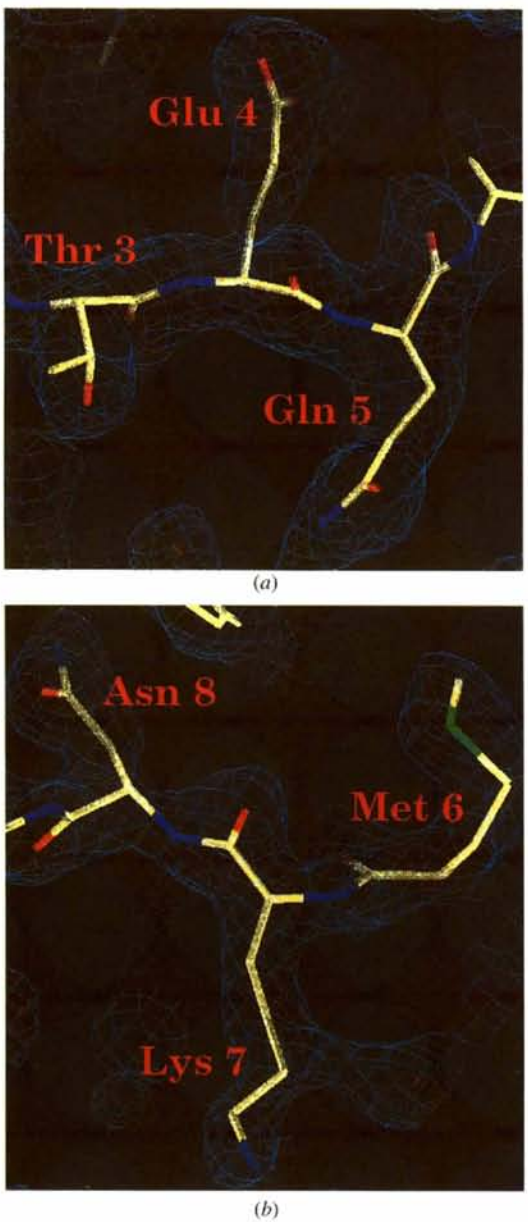

Fig. 5. Electron density (blue) contoured at 1σ for residues (*a*) 3–5 and (*b*) 6–8 in the structure of the cod pepsin.

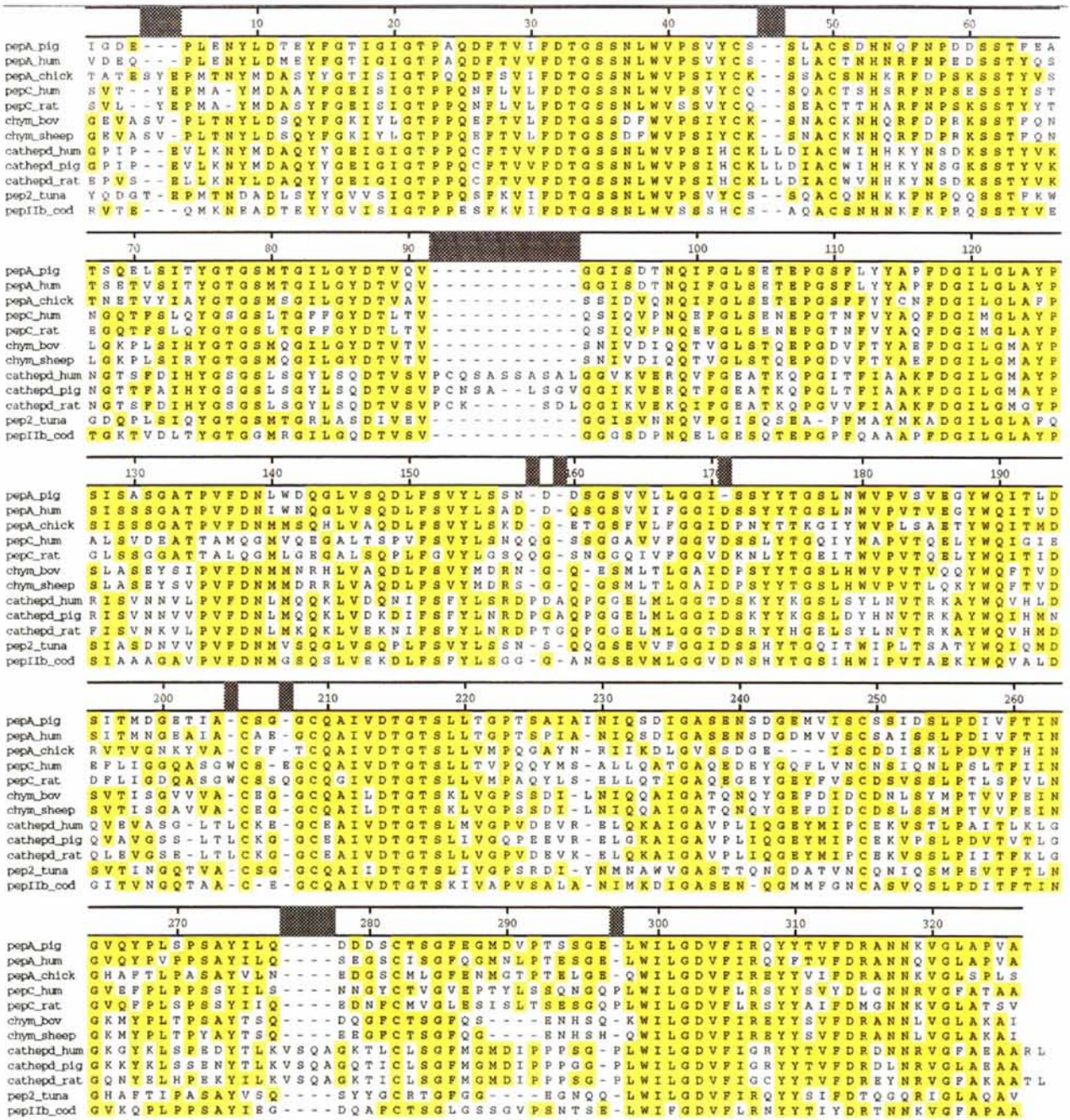

Fig. 6. Alignment of amino-acid sequences from acid proteinases. The alignment is based on amino-acid data of pepsin A from pig (pepa_pig, Tsukagoshi *et al.*, 1988), human (pepa_hum, Evers *et al.*, 1989), chicken (pepa_chick, Baudys & Kostka, 1983), and pepsin C from human (pepc_hum, Hayano, Sogawa, Ichihara, Fuji-Kuriyama & Takahashi, 1988) and rat (pepc_rat, Ishihara *et al.*, 1989), chymosin from sheep (chym_sheep, Pungercar, Strukelj, Gubensek, Turk & Kregar, 1990) and bovine (chym_bov, Foltmann, Pedersen, Jakobsen, Kauffman & Wybrandt, 1977), cathepsin D from human (cathepd_hum, Faust, Kornfeld & Chirgwin, 1985), pig (cathepd_pig, Shewale & Tang, 1984) and rat (cathepd_rat, Fujita *et al.*, 1991) and pepsin 2 from tuna fish (pep2_tuna, Tanji *et al.*, 1996). PepIIb_cod is the proposed sequence of pepsin IIb in this work. Residues representing the majority at each position are marked with boxes and the numbering system is based on the pig pepsin. The alignment was carried out by the Clustal V method (Higgins & Sharp, 1989) using the *MegAlign* program (DNAstar Inc.).

The cod pepsin consists almost entirely of β-sheet structure with very few helical regions. A *PROCHECK* (Laskowski *et al.*, 1993) analysis of the secondary structure of the pepsin from cod, the pig pepsin (Cooper *et al.*, 1990) and the endothiapepsin (Barkholt, 1987) shows that the β-strands are found almost in the same sequential positions in all pepsins, whereas the α-helices have greater variation.

The six-stranded antiparallel β-sheet (a, j, i, q, r and k), seen in Figs. 10(a) beneath the N and C domain, forms the base of the molecule with each domain contributing with three strands. In addition to strands a, j and i, the N-terminal domain is built up of seven β-strands (b–h) as seen in Fig. 10(b), forming one eight-stranded central sheet (together with strand a) and a smaller sheet the start of strands c and g and the end of strands b and f.

In the the C-terminal domain (Fig. 10c) the q, r and k strands form the second part of the central β-sheet. In addition, one five-stranded sheet (l, m, p, n and o) and two four-stranded sheets (n_1, n_2, n_5 and n_6 and l_1, l_2, n_3 and n_4) are found packed orthogonally in two sheets.

3.7. *The active site*

The structurally and sequentially well conserved core of the active site of the proteinases is built up of four loop regions (in Davies, 1990): 30–35 and 120–124 in the N domain and 213–218 and 301–305 in the C domains. The

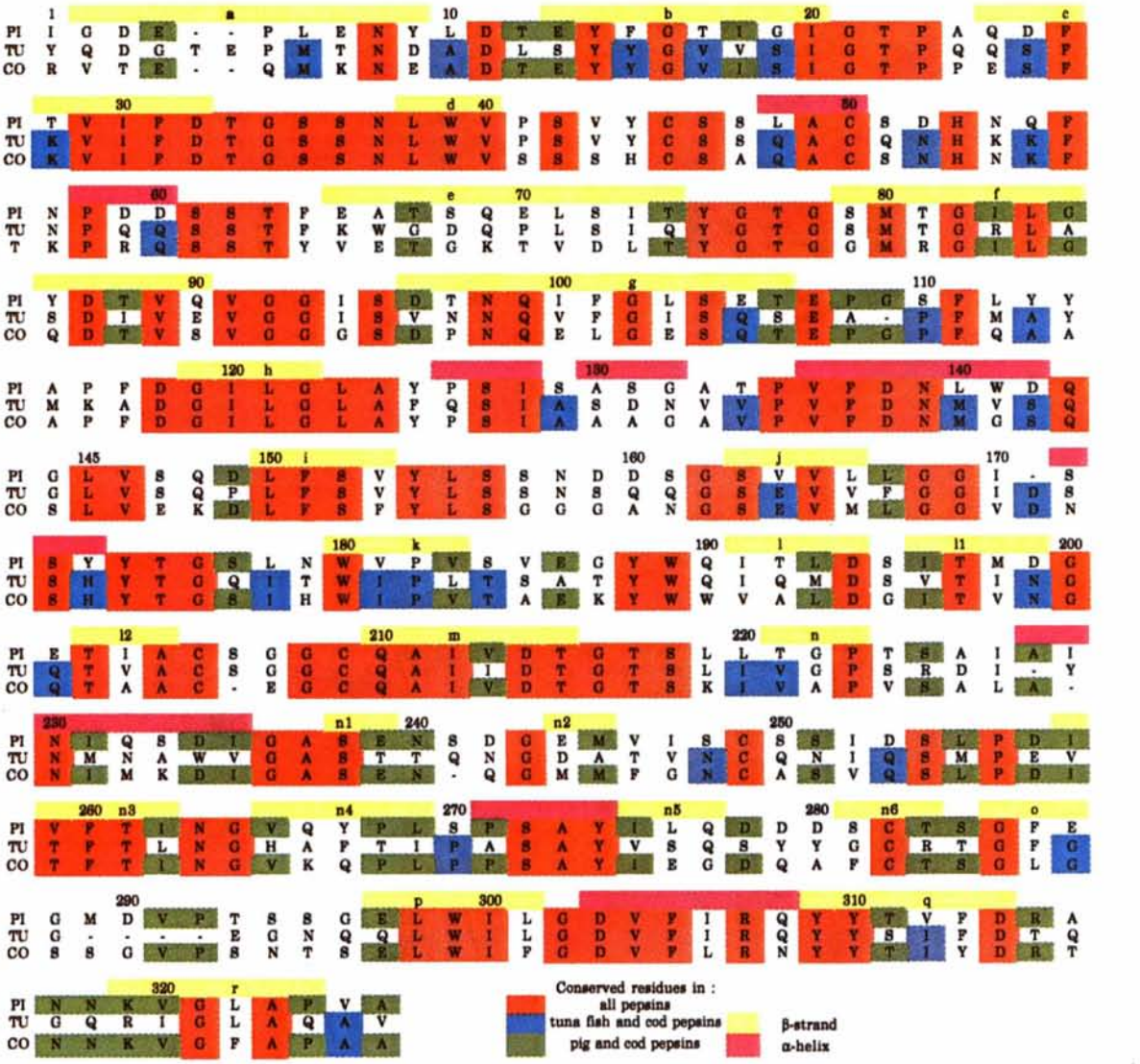

Fig. 7. The same alignment as in Fig. 6 for the pig pepsin (PI), the tuna fish pepsin 2 (TU) and the cod pepsin IIb (CO). Conserved residues in all the pepsins, the tuna fish and cod pepsins and the pig and cod pepsins are shown in red, blue and green, respectively. The β-strands and α-helices marked with yellow and violet boxes, respectively, were defined using the program *PROCHECK* (Laskowski *et al.*, 1993). The β-strands are numbered from a to r.

catalytic aspartic acid residues, Asp32 and Asp215, are both found in Asp-Thr-Gly-Thr/Ser sequences and are connected through the complex network of hydrogen bonds shown in Fig. 11. Threonines, Thr216 and Thr33, play a crucial role in the hydrogen bonding with each hydroxyl group forming a hydrogen bond to the other threonine's amide N atom. Hydrogen bonds are also seen from OD2 of Asp32 and the amide groups of Gly217 and

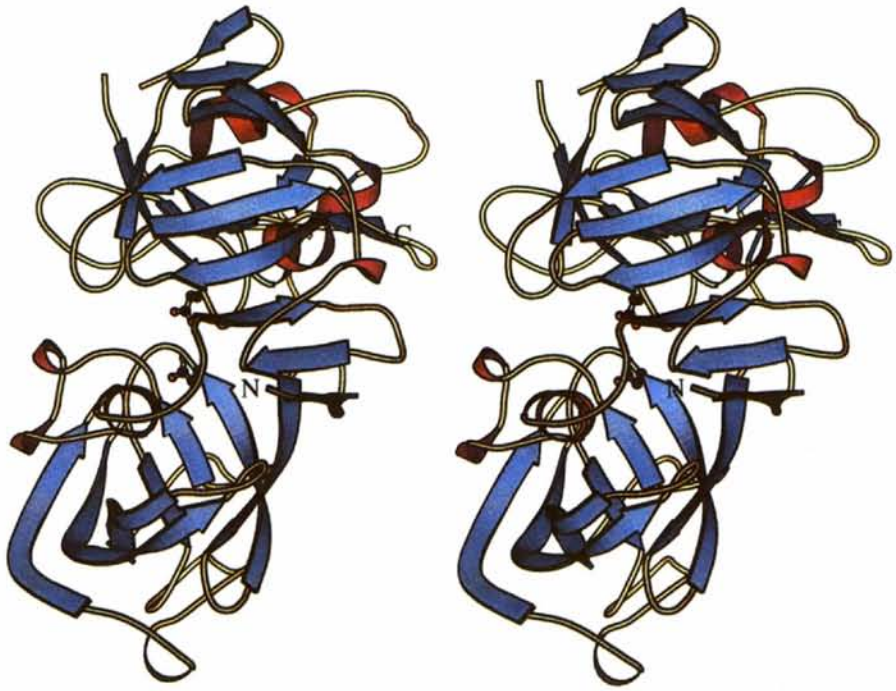

Fig. 8. A ribbon drawing of the cod pepsin made by the program *MOLSCRIPT* (Kraulis, 1991). α-Helices are coloured red, the β-strands are light blue and the loops are yellow. The catalytic Asp32 and Asp215 are illustrated in ball-and-stick models.

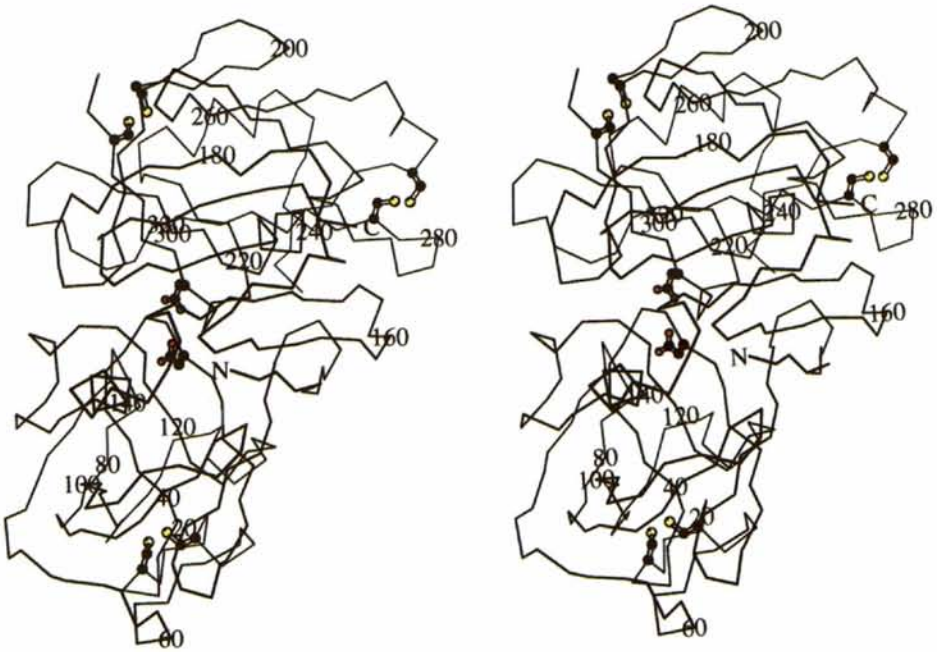

Fig. 9. A Cα trace of the pepsin in the same orientation as the previous figure. The three disulfide bridges with yellow S atoms are emphisized and the catalytic residues. The picture was generated by *MOLSCRIPT* (Kraulis, 1991).

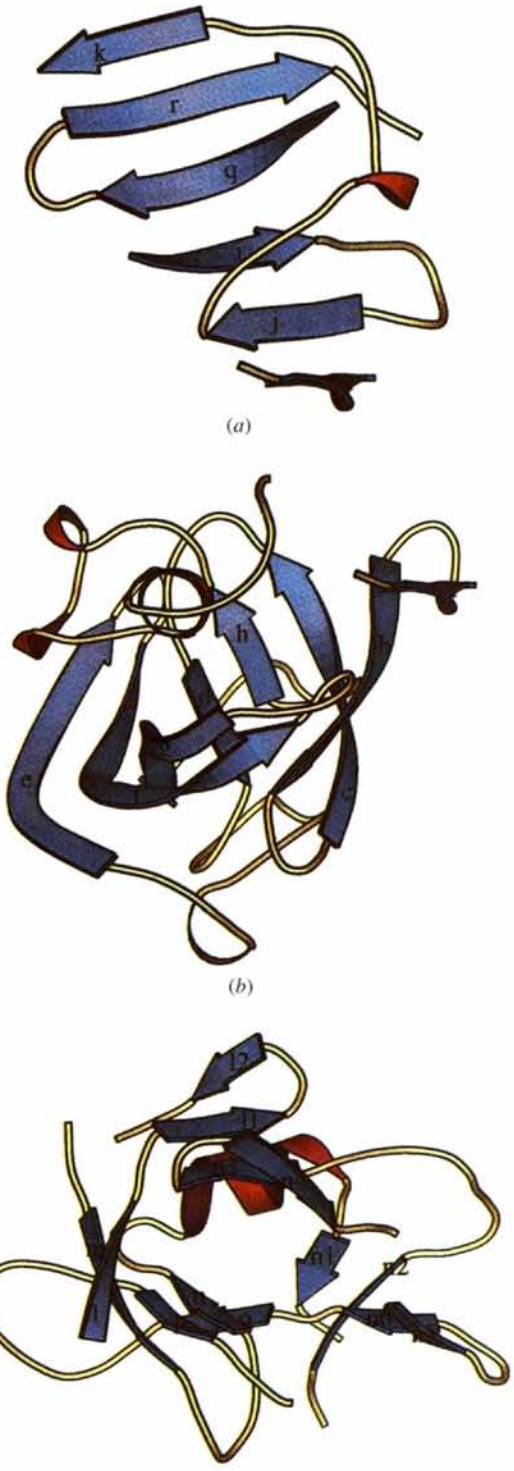

Fig. 10. (*a*) The β-central sheet in the cod pepsin structure which consists of strands *a*, *j* and *i* from the N-terminal domain and *k*, *q* and *r* from the C-terminal part of cod pepsin. (*b*) Part of the N-terminal domain where strands *a*–*h* contribute to the main β-sheet. Another, but smaller sheet, is formed from the beginning of strands *c* and *g* and the end of strands *b* and *f*. (*c*) Part of the C-terminal domain showing the five-stranded (l_1, *m*, *p*, *n* and *o*) and two four-stranded (n_1, n_2, n_5 and n_6 and l_2, l_3, n_3 and n_4) β-sheets.

Gly34. The OD1 of Asp215 is also hydrogen bonded to the amide group of Gly217 and the same atom is connected to the amide group of Thr216. The OG1 atom of Thr218 is also bound to OD2 of Asp215 so that this residue is rigidly positioned in the structure. As observed for the structure of rhizopuspepsin (Suguna *et al.*, 1987), a water molecule (Wat621) is located between the aspartic acids in the cod pepsin but is not hydrogen bonded to any residues in the structure.

All residues building up the core of the active site are well defined in the electron-density maps (see Fig. 12) and have particularly low temperature factors compared with the rest of the pepsin structure. This is also the seen for Wat621, indicating that this region is stable and conserved in all acid proteinases. It seems likely, therefore, that all aspartic proteinases catalyze the breaking of peptide bonds in the same way and have the same catalytic mechanism.

3.8. *Comparison with the pig pepsin*

The overall folding of the cod pepsin seems to be very similar to porcine pepsin. However, differences are seen in loop regions, particularly on the surface of the enzyme.

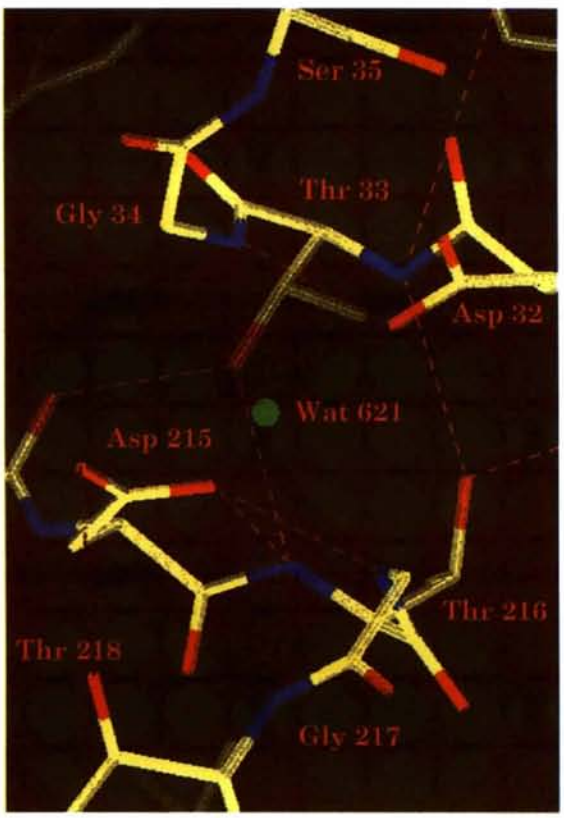

Fig. 11. Part of the active site of cod pepsin showing the catalytically active residues (Asp32 and Asp215) and their neighbouring amino acids. The atoms are coloured red, blue and yellow for O, N and C, respectively. Hydrogen bonds are in green. The figure was generated using *O* (Jones & Kjelgaard, 1993).

A superposition of the two pepsins is shown in Fig. 13(*a*). Least-squares fitting of 1292 common main-chain atoms gives an overall r.m.s. difference of 1.29 Å. The largest differences are found in loops between β-strands *e* and *f*, *f* and *g*, *g* and *h*, *n* and *n*, *o* and *p*. In addition, two regions consisting of strands l_2 and l_3 with connecting loops and the long helix from 225 to 235, strands n_1 and n_2 with connecting loops, also show large coordinate differences. Fig. 13(*a*) illustrates that the r.m.s. deviations are larger in the C-terminal lobe of the pepsin molecule than in the N-terminal lobe. That the greatest differences are found in loops is also seen from Fig. 13(*b*), where the r.m.s. differences for main-chain atoms are plotted for each residue. The overall r.m.s. displacements are calculated to 1.22 and 1.50 Å for backbone atoms and all atoms, respectively. Comparisons with the temperature-factor and correlation-coefficient plots in Figs. 3(*a*) and 3(*b*), indicate that the loops with high flexibility and badly defined electron density, are also in the regions that show greatest deviations from the structure of porcine pepsin. These deviations between the two structures therefore may not be significant, and conclusions must be drawn with caution.

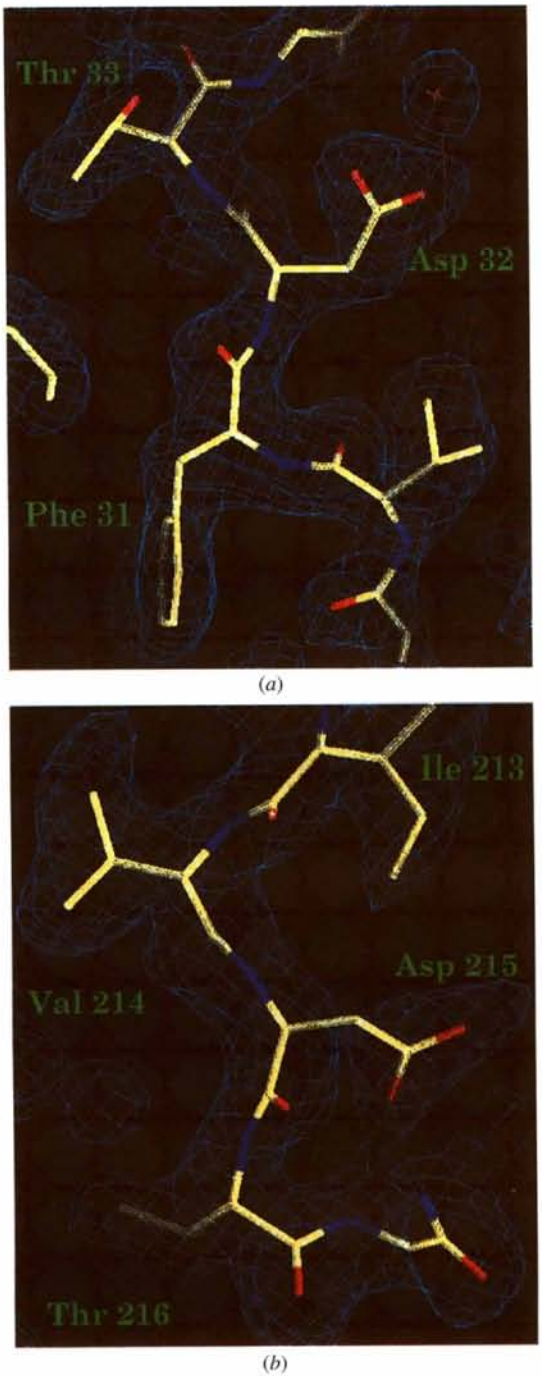

Fig. 12. Electron density ($2F_o-F_c$ map contoured at 1σ) for the catalytic amino acids, (*a*) Asp32 and (*b*) Asp215 and neighbouring residues.

4. Concluding remarks

The size, overall fold and secondary structure of cod pepsin is approximately similar to the mammalian pepsins. The 324 amino-acid sequence, interpreted from the 2.16 Å electron-density map, resembles the primary structure of mammalian pepsins more than that of mammalian cathepsin D's. However, since the present sequence is deduced from a medium-resolution map with diffuse density in some regions, conclusions must be drawn with care. As mentioned earlier, none of the glycosyl moieties which are believed to be in present cod pepsin IIb (Gildberg *et al.*, 1990), as well in the porcine cathepsin D structure (Huang *et al.*, 1979), have been detected.

Ealier biochemical studies have shown that porcine pepsin is thermostable up to 333 K (Gildberg, 1988) whereas the cod pepsins are denatured at temperatures above 313 K (Martìnez & Olsen, 1989). In addition, the pepsins from polar cod showed 40% lower temperature coefficients and 50% of maximal activity at 275 K and lower activation energies than the mammalian pepsins (Arunchalam & Haard, 1985; Brewer, Helbig & Haard, 1984), indicating that these enzymes are more flexible and have higher turnover rates at low temperatures.

The differences in thermostability and flexibility may arise from differences in the number of disulfide bridges, the average hydrophobicity and the amount of intramolecular bonds. As the number and positions of the disulfide bonds are the same in porcine pepsin as cod pepsin IIb and the average hydrophobicities are almost equal (Gildberg *et al.*, 1990), the reason for the higher stablility of the former enzyme may be a greater number of hydrogen bonds, aromatic and electrostatic interactions. Furthermore, it is possible that carbohydrates in the fish pepsins may cause higher flexibility and reduced stability by interfering with the intramolecular hydrogen-bonding network.

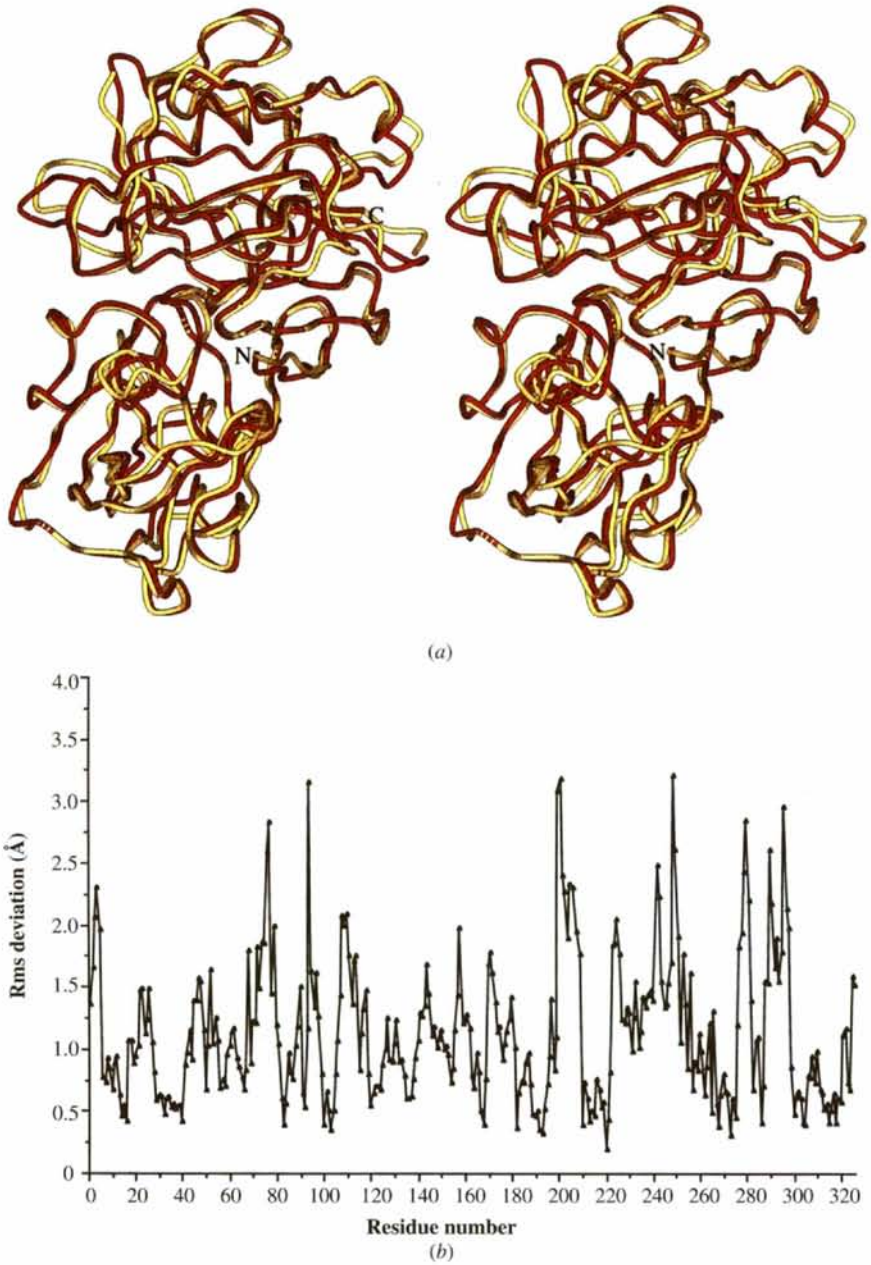

Fig. 13. (*a*) Superimposition of the cod pepsin (yellow) on the pig pepsin (red). The picture was generated using *MOLSCRIPT* (Kraulis, 1991). (*b*) Root-mean-square differences (Å) in the main-chain coordinates between the pig and cod enzyme.

As the structure of the cod pepsin IIb is built without knowledge of the primary sequence of the enzyme, a detailed comparison of the hydrogen-bonding pattern with that in porcine pepsin is of doubtful value. In order to investigate this further, the complete amino-acid sequence of the cod pepsins must be determined and X-ray data should be collected to a higher resolution.

This work has been supported by by the Norwegian Research Council.

References

Abad-Zapatero, C., Rydel, T. & Erickson, J. (1990). *Proteins Struct. Funct. Genet.* **8**, 62–81.

Almas, K. A. (1990). *Advances in Fisheries Technology and Biotechnology for Increased Profitability*, edited by M. N. Voigt & J. R. Botta, pp. 361–372. Lancaster: PA Publishing Co.

Arunchalam, K. & Haard, N. F. (1985). *Comp. Biochem. Physiol. B*, **80**, 476–473.

Barkholt, V. (1987). *Eur. J. Biochem.* **167**, 327–338.

Baudys, M. & Kostka, V. (1983). *Eur. J. Biochem.* **136**, 89–99.

Blundell, T., Jenkins, J., Pearl, L. & Sewell, T. (1985). *Aspartic Proteinases and Their Inhibitors*, edited by V. Kostka, pp. 151–161. Berlin: Walter de Gruyter.

Brewer, P., Helbig, N. & Haard, N.F. (1984). *Can. Inst. Food Sci. Technol. J.* **17**, 38–43.

Brünger, A. T. (1992). *X-PLOR, Version 3.1. A System for X-ray Crystallography and NMR*. Yale University Press, CT, USA.

Brünger, A.T. (1993). *Acta Cryst.* D**49**, 24–36.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

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

Cooper, J. B., Khan, G., Taylor, G., Tickle, J. & Blundell, T. L. (1990). *J. Mol. Biol.* **214**, 199–222.

Cunningham, M. & Tang, J. (1976). *J. Biol. Chem.* **251**, 4528–4536.

Davies, D. R. (1990). *Annu. Rev. Biophys. Chem.* **19**, 189–215.

Evers, M. P. J., Zelle, B., Bebelman, J. P., Van Beusechem, V., Kraakman, L., Hoffer, M. J. V., Pronk, J. C., Mager, W. H., Planta, R. J., Eriksson, A. W. & Frants, R. R. (1989). *Genomics*, **4**, 232–239.

Faust, P. L., Kornfeld, S. & Chirgwin, J. M. (1985). *Proc. Natl Acad. Sci. USA*, **82**, 4910–4914.

Foltman, B. (1981). *Essays Biochem.* **17**, 52–84.

Foltman, B. & Axelsen (1980). *Enzyme Regulation and Mechanism of Action*, edited by P. Mildner & B. Ries, pp. 271–280. Oxford: Pergamon Press.

Foltmann, B., Pedersen, V. B., Jakobsen, H., Kauffman, D. & Wybrandt, G. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2321–2324.

Fujinaga, M., Chernaia, M. M., Tarasova, N. I., Mosimann, S. C. & James, M. N. G. (1995). *Protein Sci.* **4**, 960–972.

Fujita, H., Tanaka, Y., Noguchi, Y., Kono, A., Himeno, M. & Kato, K. (1991). *Biochem. Biophys. Res. Commun.* **179**, 190–196.

Gildberg, A. (1988). *Comp. Biochem. Physiol. B***91**, 425–435.

Gildberg, A., Olsen, R. L. & Bjarnason, J. B. (1990). *Comp. Biochem. Physiol. B*, **96**, 323–330.

Gildberg, A. & Raa, J. (1983). *Comp. Biochem. Physiol. A*, **75**, 337–342.

Gulnik, S., Baldwin, E, T., Tarasova, N. & Erickson, J. (1992). *J. Mol. Biol.* **227**, 265–270.

Guruprasad, K., Dhanaraj, V., Groves, M. & Blundell, T. L. (1994). *Perspec. Drug Discov. Design*, **2**, 329–341.

Haard, N. F. (1992). *J. Aquat. Food Prod. Technol.* **1**, 17–35.

Hayano, T., Sogawa, K., Ichihara, Y., Fujii-Kuriyama, Y. & Takahashi, K. (1988). *J. Mol. Biol. Chem.* **263**, 1382–1385.

Hazel, J. R. & Prosser, C. L. (1974). *Physiol. Rev.* **54**, 620–677.

Hendrickson, W. (1985). *Methods Enzymol.* **115**, 252–270.

Higgins, D. G. & Sharp, P. M. (1989). *Comput. Appl. Biosci., 5, 151–153.*

Hochacka, P. W. & Somero, G. N. (1971). *Fish Physiology*, edited by W. S. Hoar & J. D. Randall, Vol. 6, pp. 100–156. New York: Academic Press.

Huang, J. S., Huang, S. S. & Tang, J. (1979). *J. Biol. Chem.* **254**, 504–417.

Ishihara, T., Ichihara, Y., Hayano, T., Katsura, I., Sogawa, K., Fujii-Kuriyama, Y. & Takahashi, K. (1989). *J. Biol. Chem.* **264**, 10193–10199.

James, M. N. G. & Sielecki, A. R. (1987). *Biological Macromolecules and Assemblies*, edited by J. Jurnak, A. McPherson, Vol. 3, pp. 413–482. New York: John Wiley.

Jones, T. A. & Kjeldgaard, M. (1993). *O The Manual, Version* 5.9.1, Uppsala, Sweden.

Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.

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

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

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

McPherson, A. (1982). *Preparation and Analysis of Protein Crystals*, pp. 96–97. New York: John Wiley.

Martinez, A. & Olsen, R. L. (1989). *US Biochem. Corp.* **16**, 22–23 (20).

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

Messerschmidt, A. R. & Pflugrath, J. W. (1987). *J. Appl. Cryst.* **20**, 306–315.

Miller, M., Jaskolski, M., Mohana, R., Leis, J. & Wlodawer, A. (1989). *Nature (London)*, **337**, 576–579.

Navia, M. A., Fitzgerald, P. M. D., McKeever, B. M., Leu, C-T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L. & Springer, J. P. (1989). *Nature (London)*, **337**, 615–620.

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.

Noda, M. & Murakami, K. (1981). *Biochim. Biophys. Acta*, **658**, 27–34.

Pungecar, J., Strukelj, B., Gubensek, F., Turk, V. & Kregar, I. (1990). *Nucleic Acids Res.* **18**, 4602.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Adv. Protein Chem.* **23**, 283–437.

Rao, S. N., Jih, J.-H. & Hartsuck, J. A. (1980). *Acta Cryst.* A**36**, 878–884.

Roberts, N. B., Peek, K., Keen, J. N. & Taylor, W. H. (1995). *Int. J. Biochem. Cell Biol.* **27**, 133–137.

Shewale, J. G. & Tang, J. (1984). *Proc. Natl Acad. Sci. USA*, **81**, 3703–3707.

Sielecki, A. R., Fedorov, A. A., Boodhoo, A., Andreeva, N. S. & James, M. N. G. (1990). *J. Mol. Biol.* **214**, 143–170.

Sielecki, A. R., Hayakawa, K., Fujinaga, M., Murphy, M. E. P., Fraser, M., Muir, A. K., Carilli, C. T., Lewicki, J. A., Baxter, J. D. & James, M. N. G. (1989). *Science*, **243**, 1346–1351.

Strøm, T. & Raa, J. (1993). *J. Mar. Biotechnol.* **1**, 3–7.

Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E. & Sheriff, S., Cohen, G. H. & Davies, D. (1987). *J. Mol. Biol.* **196**, 877–900.

Tang, J. & Wong, R. N. S. (1987). *J. Cell. Biochem.* **33**, 53–63.

Tanji, M., Kageyama, T. & Takahashi, K. (1988). *Eur. J. Biochem.* **177**, 251–259.

Tanji, M., Yakabe, E., Kageyama, T. & Takahashi, K. (1996). *J. Biochem.* **120**, 647–656.

Tsukagoshi, N., Ando, Y., Tomita, Y., Uchida, R., Takemura, T., Sasaki, T., Yamagata, H., Udaka, S., Ichihara, Y. & Takahashi, K. (1988). *Gene*, **65**, 258–292.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J. & Kent, S. B. H. (1989). *Science*, **245**, 616–621.

Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.