

X-ray Analyses of Aspartic Proteinases

II.† Three-dimensional Structure of the Hexagonal Crystal Form of Porcine Pepsin at 2.3 Å Resolution‡

J. B. Cooper, G. Khan, G. Taylor§, I. J. Tickle and T. L. Blundell

Laboratory of Molecular Biology
Department of Crystallography
Birkbeck College
Malet Street, London, WC1E 7HX, U.K.

(Received 27 October 1989; accepted 21 February 1990)

The molecular structure of the hexagonal crystal form of porcine pepsin (EC 3.4.23.1), an aspartic proteinase from the gastric mucosa, has been determined by molecular replacement using the fungal enzyme, penicillopepsin (EC 3.4.23.6), as the search model. This defined the space group as $P6_322$ and refinement led to an R -factor of 0.190 at 2.3 Å resolution. The positions of 2425 non-hydrogen protein atoms in 326 residues have been determined and the model contains 371 water molecules. The structure is bilobal, consisting of two predominantly β -sheet lobes which, as in other aspartic proteinases, are related by a pseudo 2-fold axis. The strands of the mixed β -sheets (1_N and 1_C) of each lobe are related by an intra-lobe topological 2-fold symmetry. Two further β -sheets, 2_N and 2_C , are each composed of two topologically related β -hairpins folded below the 1_N and 1_C sheets. A further six-stranded sheet (3) spans the two lobes and forms a structure resembling an arch upon which the four other sheets reside. The interface between sheets 1_N and 1_C forms the catalytic centre consisting of absolutely conserved aspartate residues 32 and 215, which are shielded from solvent by a β -hairpin loop (75 to 78).

The crystal structure of a mammalian aspartic proteinase indicates that interactions with substrate may be more extensive on the prime side of the active site cleft than in the fungal enzymes and involve Tyr189 and the loop 290 to 295, perhaps contributing to the transpeptidase activity of pepsin and the specificity of the renins. Comparison with the high-resolution structure of pepsinogen gives a root-mean-square deviation of 0.9 Å and reveals that, in addition to local rearrangement at the active site, there appears to be a rigid group movement of part of the C-terminal lobe of pepsin towards the cleft on activation.

A large proportion of the absolutely conserved residues in aspartic proteinases are polar and buried. An examination of the pepsin structure reveals that these side-chains are involved in hydrogen-bond interactions with either the main chain of the protein or other conserved side-chains of the enzyme or propeptide.

1. Introduction

Pepsins are the major digestive enzymes of the alimentary canal and operate optimally at the low pH of the stomach. Pepsin A (EC 3.4.23.1), the major component of gastric juice in pigs, is produced by the mucosal cells of the fundus region of the stomach as an inactive precursor, pepsinogen.

On secretion into the acidic lumen of the stomach, cleavage of the propeptide occurs, yielding active pepsin as a single-chain molecule. Determination of the molecular structure of this readily available aspartic proteinase should give some insight into the structures of the more scarce mammalian enzymes, for example, renin and cathepsin D.

Porcine pepsin A has been sequenced chemically (Sepulveda *et al.*, 1975) and has a calculated molecular mass of 34 kDa, which agrees well with the value of 33 kDa obtained by sedimentation equilibrium studies (Williams & Rajogopalan, 1966). The enzyme has a very low pI in the range of two to three pH units, which must be due to the high proportion of carboxyl residues (43 in total). The protein is phosphorylated at Ser68 and has

† Paper I in this series is Blundell *et al.*, 1990.
‡ Dedicated to J. D. Bernal, Dorothy Hodgkin and Max Perutz, who all contributed to this structure analysis in days when large crystal cells were a much greater challenge to the X-ray crystallographer.
§ Currently at Department of Biochemistry, University of Bath, U.K.

three disulphide bridges, two of which enclose three and four residue loops (for an excellent review, see Fruton, 1987). The essential catalytic residues were identified as aspartate residues 32 and 215 by chemical modification (Tang *et al.*, 1973).

Single crystals of porcine pepsin were first obtained by Northrop (1930) and those grown by Dr J. Philpot in Tiselius's laboratory in Uppsala were the first protein crystals to be studied by X-ray diffraction using the now standard method of capillary mounting to prevent water loss (Bernal & Crowfoot, 1934; Hodgkin, 1970). The crystals were hexagonal; $a=b=67.4$ Å, $c=290.1$ Å (1 Å = 0.1 nm; 1 molecule/asymmetric unit), although the c -dimension was not determined correctly until the advent of monochromatic X-ray sources (Perutz, 1949). Although the diffraction extended to high resolution, the crystal structure of this form was never determined. However, a monoclinic form of pepsin can be obtained from a protein solution of 20 mg/ml at pH 2.0 made with 20% (v/v) ethanol and the structure of this form has been determined at 2.0 Å resolution (Andreeva *et al.*, 1984). An independently refined set of monoclinic co-ordinates from Drs C. Adad-Zapatero and J. Erickson (Abbott Laboratories, Chicago) has been deposited in the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

The structure of porcine pepsinogen has been solved to 1.8 Å resolution (James & Sielecki, 1986) and 1.6 Å (Hartsuck & Remington, 1988) and elsewhere a number of renin inhibitors have been co-crystallized non-isomorphously with pepsin using ethanol as precipitant (C. Abad-Zapatero & J. Erickson, personal communication). Chymosin, another mammalian aspartic proteinase has been solved by X-ray analysis at 2.2 Å (N. Andreeva, M. Saffro, M. Newman, I. J. Tickle & T. L. Blundell, unpublished results; Gilliland *et al.*, 1988) and human renin has been solved at 2.5 Å (Sielecki *et al.*, 1989).

A number of fungal aspartic proteinase structures have been determined by X-ray diffraction at high resolution; for example, endothiapepsin (Blundell *et al.*, 1985), rhizopuspepsin (Suguna *et al.*, 1987a) and penicillopepsin (James & Sielecki, 1983).

In this paper, we describe the hexagonal crystal form of porcine pepsin solved using synchrotron film data that extended to 2.34 Å resolution. Molecular replacement using penicillopepsin (EC 3.4.23.6) as the search model gave a solution that defined the space group as $P6_322$. Remodelling and refinement were greatly aided in the later stages by the pepsinogen structure (Hartsuck & Remington, 1988). We provide a detailed comparison of the enzyme and proenzyme structures of an aspartic proteinase.

2. X-ray Analysis

(a) Crystallization and data collection

The crystals were grown from a 280 mg pepsin/ml solution, the pH of which was lowered to 3.6 by addition of 0.5 M-sulphuric acid (Jenkins, 1979). Turbidity was

cleared by warming to 35°C and crystals were grown from this solution by slow cooling in a vacuum flask. The largest reached a size of 0.3 mm $\times$ 0.3 mm $\times$ 0.5 mm after 4 days. X-ray data were collected with 1.4 Å radiation at LURE using V-shaped cassettes and processed using the methods described by Arndt & Wonacott (1977). In all, 51,501 native measurements consisting of 48,056 fully recorded reflections were merged to give 15,613 uniques to 2.34 Å with an R_{merge} of 9.6%; 69% of these data are more significant than $3\sigma(I)$. Hexagonal pepsin crystals were found to have an ambiguity of the hand ($P6_322$ or $P6_322$) that was not resolved by efforts to determine the structure using multiple isomorphous replacement.

(b) Rotation function studies

Several attempts at molecular replacement using the partially refined fungal structures proved unsuccessful with this crystal form so, more recently, another effort was made using all the highly refined aspartic proteinases available in the Brookhaven Protein Data Bank at that time.

The 2.0 Å co-ordinates of monoclinic pepsin (kindly provided by Dr N. Andreeva) were placed in a hypothetical P1 cell with the molecules separated by 20 Å for calculation of structure factors and E values in readiness for the fast rotation function (Crowther, 1972). E values were also calculated for the hexagonal reflection data. The cross-rotation function was calculated using 3.0 Å data with various Patterson radii limits between 4 and 25 Å but did not produce a peak except for an extremely broad plateau. Inspection revealed that this structure contained some deviations from ideal stereochemistry, indicating that further careful restrained refinement was necessary. Therefore, the other 3 high-resolution aspartic proteinase structures (endothiapepsin, penicillopepsin and rhizopuspepsin) were fitted to the pepsin co-ordinates using FITZ (Taylor, 1983), and structure factors and E values were calculated for each in the same P1 cell as before. The results from the Crowther rotation function calculations using 3.0 Å data and radii limits of 4 and 15 Å are shown in Fig. 1. None of the rotation functions, which are notably inconsistent, contains peaks corresponding to that obtained with pepsin, although each contains sharper features. At this stage, efforts were made to find consistent peaks with the fungal enzymes by adjusting the resolution and radii limits. These enzymes have sequence identities with porcine pepsin of about 30% on average, which is far from ideal for molecular replacement. The results are given in Table 1.

From Fig. 1 and Table 1, it can be seen that the sharp feature was obtained with penicillopepsin. This peak has the largest height to r.m.s. ratio (where r.m.s.† is the root-mean-square value of the rotation function) of all the enzymes and reappears consistently when all but the lowest resolution data are used. The fact that this peak appears only at low resolution with rhizopuspepsin is alarming, as this protein has the highest homology to pepsin of the three fungal enzymes. Endothiapepsin has the lowest homology and fails to produce any peaks that agree with the other rotation functions. The aspartic proteinase fold can be divided into 2 rigid bodies that differ slightly in orientation in different enzymes (Sali *et al.*, 1989). Hence the penicillopepsin structure may fortuitously resemble porcine pepsin slightly more than one or both of the other search models used in the rotation function. To check that no mistakes had been

† Abbreviation used: r.m.s., root-mean-square.

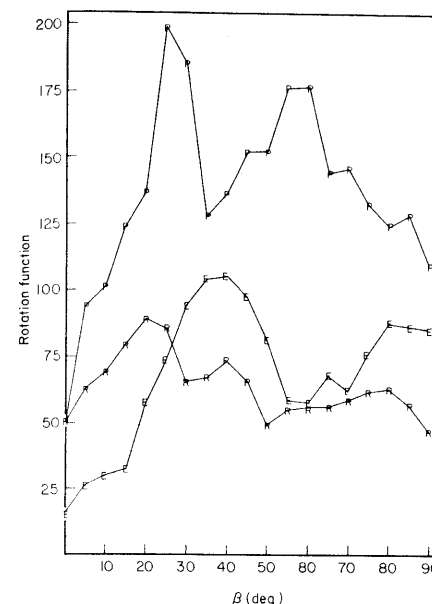


Figure 1. The cross-rotation function results using endothiapepsin (E), rhizopuspepsin (R) and penicillopepsin (P) versus porcine pepsin. The maximum value in each β -section is shown. E values to 3 Å were used with Patterson radii limits of 4 and 15 Å.

Table 1

The fast rotation function results for the fungal enzymes versus porcine pepsin (α , β and γ angles in degrees for the 3 highest peaks in each run are shown)

A. <i>Endothia</i> pepsin			
Resolution limits, 20 to 6 Å; radii limits, 6 to 20 Å			
α	β	γ	Height/r.m.s.
52.5	45.0	5.0	3.0
32.5	35.0	97.5	2.7
22.5	55.0	90.0	2.7
Resolution limits, 20 to 5 Å; radii limits, 6 to 20 Å			
40.0	85.0	127.5	2.6
20.0	60.0	95.0	2.5
47.5	45.0	7.5	2.2
Resolution limits, 20 to 3 Å; radii limits, 4 to 15 Å			
30.0	40.0	-175.0	3.3
35.0	80.0	127.5	2.7
52.5	65.0	147.5	2.1

Table 1 (continued)

α	β	γ	Height/r.m.s.
Resolution limits, 20 to 3 Å; radii limits, 4 to 20 Å			
25.0	35.0	-172.5	3.7
47.5	60.0	165.0	2.9
20.0	90.0	117.5	2.7
B. <i>Rhizopuspepsin</i>			
Resolution limits, 20 to 6 Å; radii limits, 6 to 20 Å			
α	β	γ	Height/r.m.s.
15.0	45.0	95.0	3.4
55.0	80.0	140.0	2.6
47.5	20.0	5.0	2.5†
Resolution limits, 20 to 5 Å; radii limits, 6 to 20 Å			
47.5	20.0	5.0	3.5†
0.0	85.0	145.0	2.4
35.0	60.0	170.0	2.2
Resolution limits, 20 to 4 Å; radii limits, 6 to 20 Å			
12.5	75.0	-57.5	2.6
5.0	20.0	82.5	2.3
17.5	25.0	-160.0	2.3
Resolution limits, 20 to 3 Å; radii limits, 4 to 15 Å			
0.0	20.0	120.0	2.9
12.5	40.0	-157.5	2.4
57.5	80.0	-75.0	2.0
Resolution limits, 20 to 3 Å; radii limits, 4 to 20 Å			
50.0	20.0	125.0	2.9
17.5	35.0	-157.5	2.5
57.5	80.0	-75.0	2.2
C. <i>Penicillopepsin</i>			
Resolution limits, 20 to 6 Å; radii limits, 6 to 20 Å			
α	β	γ	Height/r.m.s.
55.0	75.0	95.0	3.0
25.0	35.0	162.5	2.7
22.5	25.0	17.5	2.4†
Resolution limits, 20 to 5 Å; radii limits, 6 to 20 Å			
45.0	25.0	20.0	3.3†
50.0	80.0	105.0	3.2
30.0	30.0	17.5	3.1
Resolution limits, 20 to 3 Å; radii limits, 4 to 15 Å			
45.0	25.0	17.5	3.6†
35.0	60.0	65.0	3.1
45.0	60.0	-175.0	3.0
Resolution limits, 20 to 3 Å; radii limits, 4 to 20 Å			
42.5	30.0	22.5	3.9†
32.5	60.0	170.0	3.6
47.5	50.0	65.0	3.5
Resolution limits, 20 to 3 Å; radii limits, 4 to 25 Å			
47.5	25.0	12.5	4.2†
50.0	55.0	67.5	3.3
12.5	70.0	-177.5	3.1

† Peak consistent with penicillopepsin.

‡ A consistent peak.

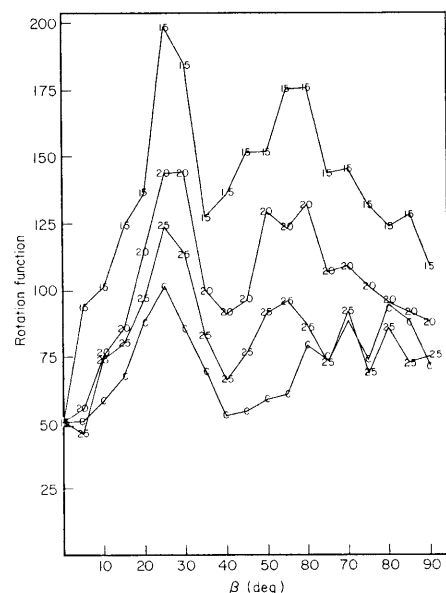


Figure 2. The penicillopepsin rotation function calculated using various Patterson radii limits (shown in Å) and the structurally conserved regions (C).

made in preparing the E values for the oriented enzymes, cross-rotation functions were calculated against chymosin (M. Newman, unpublished results). All 3 enzymes produced 1 sharp peak in the same place to within a few degrees. We suggest that the difficulties encountered with hexagonal pepsin may be due to the compactness of the rotation function arising from high symmetry of the target Patterson function (6/nmm) which is expected to give a high noise level.

It was thought that removing the parts of the search molecules that are not structurally conserved (i.e. C^* separation ≥ 3.5 Å) might reduce the noise in the rotation functions. The MNYFIT program (Sutcliffe *et al.*, 1987) was used to produce co-ordinate files for the structurally conserved regions, approx. 70% of each enzyme. These co-ordinates were fitted to the monoclinic pepsin molecule and their E values calculated. The rotation functions for the endotheiapepsin and rhizopuspepsin cores are almost completely featureless. In contrast, the 1 major peak obtained with penicillopepsin is unaltered (Fig. 2) and, significantly, the second highest peak obtained with the whole enzyme has vanished. Therefore, the peak obtained with penicillopepsin was considered the most likely due to its height, consistency and sharpness (Fig. 3). The improvement in height/r.m.s. ratio as the resolution and radius of integration were increased (Table 1), and the insensitivity of the peak to removal of the loops from the model also supported this conclusion. The rotation function was repeated using E values for the penicillopepsin model transformed by the Eulerian angles of the highest peak. The strongest peak reappeared close to the origin, the corresponding angles were applied and the process repeated, producing a peak within 2.5° of the origin. These angles were applied.

(c) The translation function

A hemisphere of reflections to 4.0 Å was calculated using GENSEC (Dodson, 1984) for the oriented penicillopepsin molecule in the hexagonal unit cell, assuming $P1$ symmetry. The partial F_c terms for each asymmetric unit were obtained from this data set using the PREPARE/SORTLICE/COLLATE series of programs (P. R. Evans, unpublished results). The collated partial F_c terms were combined with native F_o data to 4.0 Å. The TFGGEN program (Tickle, 1985) was used to calculate the Fourier coefficients of the T -function ($T(t)$; function of Crowther & Blow (1967) with intra-molecular vectors subtracted) for both possible space groups ($P6_322$ and $P6_322$) using 6.0 and 4.0 Å data. The T -function maps were generated using the FFT program (Ten Eyck, 1973) for half the unit cell along c . The results are shown in Table 2, together with the packing tests performed using FRODO (Jones, 1978).

There are no peaks that appear in both the $P6_322$ runs, and inspection reveals that the 5 highest peaks in each

Table 2

Translation functions of penicillopepsin with hexagonal pepsin data

Resolution = 6.0 Å				Resolution = 4.0 Å					
Grid	co-ord.	Height/r.m.s.	Packing	Grid	co-ord.	Height/r.m.s.	Packing		
A. Translation function for $P6_322$									
14	0	36	4.6	Crashes	0	4	11	4.7	Crashes
4	15	8	4.6	Crashes	13	1	46	4.4	Crashes
24	60	129	4.4	Crashes	11	30	93	4.4	Crashes
14	0	108	4.4	Crashes	12	0	46	4.3	Crashes
24	37	119	4.3	Crashes	8	50	102	4.3	Crashes
B. Translation function for $P6_322$									
19	29	58	5.2	Crashes	63	17	79	5.7	*PACKS
19	29	122	5.0	Crashes	63	17	93	5.1	Crashes
63	17	80	4.6	*PACKS	63	17	127	4.9	Crashes
39	49	117	4.5	Crashes	63	17	138	4.8	Crashes
63	16	96	4.5	Crashes	63	17	43	4.8	Crashes

One grid unit is approx. 1 Å along real axes.

* Indicates same peak in both runs.

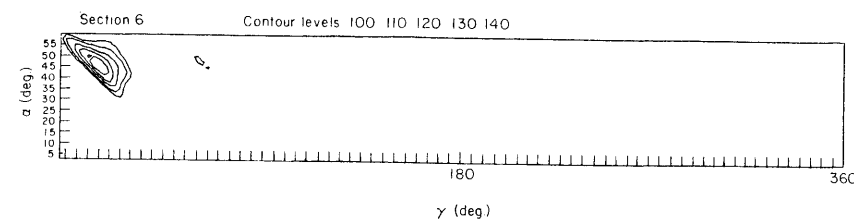


Figure 3. The $\beta=25^\circ$ section of the penicillopepsin rotation function showing a single sharp peak ($3.9 \times$ r.m.s.). Patterson radii limits of 4 and 20 Å with 20 to 3 Å resolution E values.

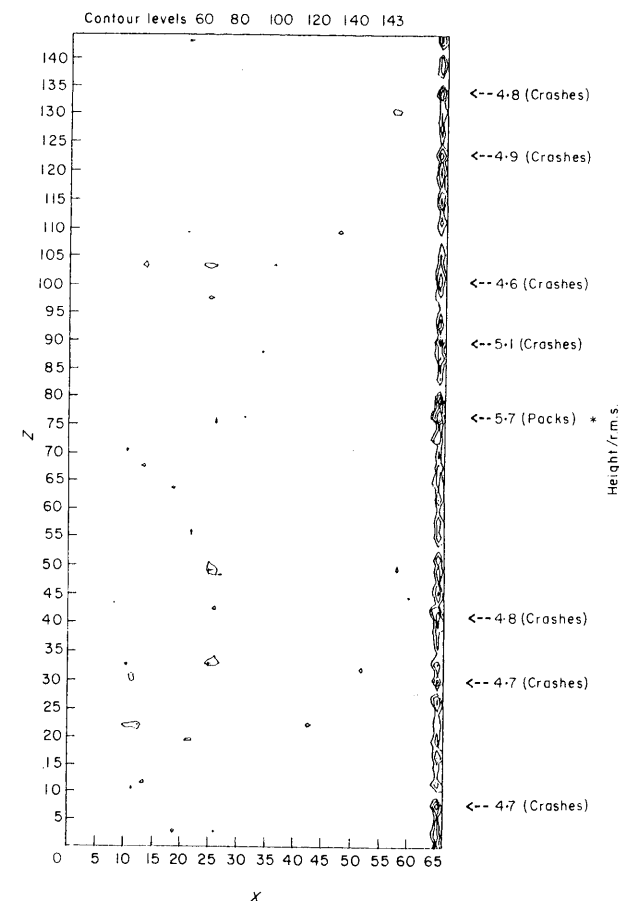


Figure 4. One section of the translation (T) function for hexagonal pepsin calculated with resolution limits of 20 and 4 Å assuming $P6_322$. The largest features in the whole map lie along the streak parallel to z , although only the highest (*) allows reasonable packing of the penicillopepsin model.

run cause the C^α skeletons of symmetry-related molecules to interpenetrate. For $P6_522$, the 3rd highest peak for 6.0 Å data allows the molecules to pack sensibly and when the calculation is performed using 4.0 Å data, the same peak reappears as the strongest in the map. The strongest features in this map lie along a streak parallel to c (Fig. 4). All these peaks, except the highest, cause the symmetry-related molecules to crash. Hence, there appeared to be a plausible molecular replacement solution that defined the space group as $P6_522$.

The translation functions at 4.0 Å corresponding to the highest rotation function peaks obtained with endotheiapepsin and rhizopuspepsin were also calculated for both space groups. From analyses similar to that in Table 2, it was obvious that none of the maps contained outstanding T -function peaks that allowed the molecules to pack reasonably.

(d) Hexagonal pepsin refinement and rebuilding

In order to assess the validity of the solution, a control refinement with the molecule in an incorrect position was necessary. Therefore, the penicillopepsin molecule was refined in positions corresponding to the highest and the 2nd highest T -function peaks that differ only in their z positions by 5% of the cell edge, although only the highest peak allows reasonable packing. Exactly the same refinement strategy was used for each position. The 1st 10 cycles consisted of rigid body refinement using RESTRAIN (Haneef *et al.*, 1985) of the 2 domains with medium-resolution data. For the following 8 cycles of restrained refinement, the molecule was split into 35 rigid

bodies corresponding to the segments of secondary structure as defined in the Brookhaven database entry for penicillopepsin (2APP). The last 4 cycles, performed using 2.3 Å data, included unrestrained refinement of isotropic u^2 values, subsequently referred to as U or U_{iso} atomic values ($B_{iso}/8\pi^2$). Since the molecules had the same orientation and position except for a 5% difference in z -position, it is not surprising that they behaved similarly during refinement, although the highest T -function peak has a correlation coefficient (F_o versus F_c) of 57% which is 10% higher than the wrong solution. This difference should be exaggerated by using the reflections that are most sensitive to the z -position of the molecule, i.e. those close to c^* . Accordingly, the 300 reflections closest to c^* gave a correlation coefficient for the correct peak of 37%, which is 20% higher than the value obtained for the incorrect peak. The U values obtained for the main chain after this refinement were examined for a meaningful variation along the polypeptide. As can be seen in Fig. 5 the largest U values for the highest T -function peak are obtained in regions that are not structurally conserved in the 3 fungal enzymes; i.e. those regions where penicillopepsin would be expected to differ most from porcine pepsin. The wrong peak causes the average main chain U values to fluctuate much less, as might be expected if the individual U_{iso} values are random. The R -factors for the correct and incorrect peaks were 47% and 51% respectively.

After this refinement, Sim weighted $2F_o - F_c$ and $F_o - F_c$ maps were calculated for the correct peak using 2.3 Å data. These maps were very encouraging in certain regions, e.g. the active site and the β -sheet beneath the

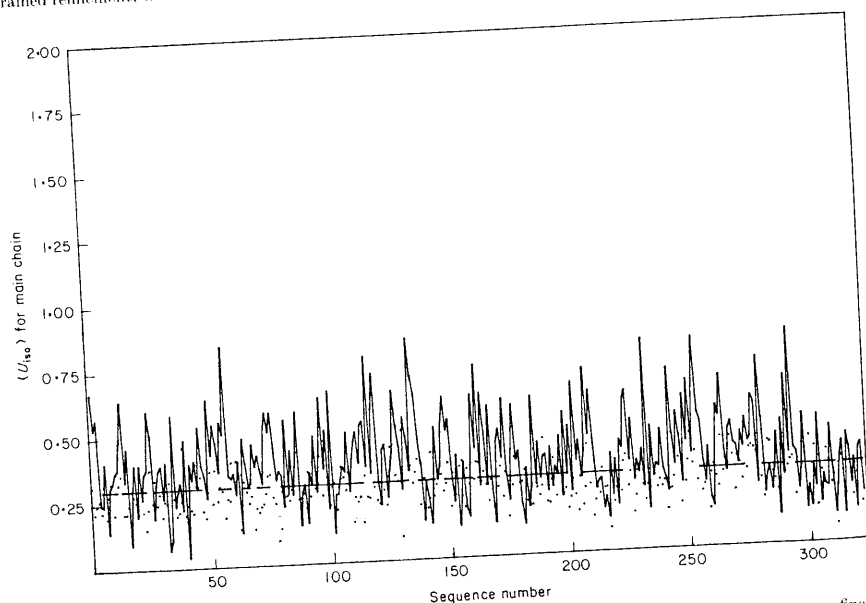


Figure 5. The average U_{iso} values for main-chain atoms of penicillopepsin after 18 cycles of least-squares refinement using the hexagonal pepsin F_o data. The oriented molecule was refined in 2 positions, corresponding to the highest and 2nd highest translation function peaks; lines and dots, respectively. The structurally conserved regions are shown by horizontal lines.

substrate binding cleft (sheet 3). The electron density, which was mainly contiguous throughout the protein, was used to guide the necessary sequence changes, insertions and deletions. Residues in the vicinity of the disulphides (45 to 50, 206 to 210 and 250 to 283) proved particularly problematic, e.g. the 206-210 cystine necessitated inserting a 4 residue loop, for which there was little density, just before a structurally conserved region including Asp215. The electron density for the conserved residues was generally good and there were usually sufficient indications to allow insertion of aromatic residues where necessary. All the insertions and deletions were made at this stage, although some residues were left as Ala or Gly. To check that the solution was self-consistent in the rotation function, the rebuilt molecule was placed in the original orientation. E values were calculated and the cross-rotation function re-run. This produced a large peak in the same position as was obtained with penicillopepsin (height/r.m.s. = 4.0 for radii = 4 to 20 Å, resolution, 20 to 3 Å). The co-ordinates were subjected to another round of refinement and rebuilding followed by more refinement, which reduced the R -factor to 38% (correlation coefficient, 68%). Despite big improvements for inserted aromatic residues, the absence of $2F_o - F_c$ density in some regions after 41 cycles, especially the disulphides, indicated that some major re-interpretations were needed in these parts and temporarily cast doubt on the validity of the solution.

(e) Real space rotation function tests

In view of the inconsistency of the fast rotation functions, it was decided to repeat these using a real space method (Huber, 1965) encoded in the program PROTEIN by Steigemann (1974). F_o terms to 3.0 Å were used for calculation of the target Patterson over all a and b , and enough of c to cover the sphere of integration (20 Å). All 3.0 Å F_c terms for the 3 fitted fungal search models and the structurally conserved regions for rhizopuspepsin were used to calculate Patterson maps, which were searched for grid points above a chosen density level and within the radii limits of 4 and 15 Å. Since F_{000} was not used in the map calculation, the lowest density value was added to all grid points to remove negative density. Hence, all grid points of the search Patterson could be included for correlation with the target map. The resulting rotation functions, calculated at 6° intervals on the Rossmann & Blow angles ($\theta_1, \theta_2, \theta_3$ where $\theta_1 = \alpha + 90^\circ$, $\theta_2 = \beta$ and $\theta_3 = \gamma - 90^\circ$) were remarkably similar, with the strongest peak occurring at the same grid point in each map (Fig. 6), rhizopuspepsin being the best. However, this peak does not correspond to any of the peaks obtained in the fast rotation function, indicating that the previous solution, obtained with penicillopepsin, may have been incorrect. To test the latest solution, the search was repeated for rhizopuspepsin using a fine grid (0.5°). The refined angles were applied to the co-ordinates and the rotation function re-run confirming that the largest peak in the whole map was within 2° of the origin. After slight re-orientation of the model, structure factors were calculated in the hexagonal pepsin cell and the translation function was calculated for both possible space groups. Fortunately, neither contained outstanding peaks, although the highest 5 in each map were tested for packing. Only the 3rd highest peak for $P6_522$ allowed reasonable packing, but applying the corresponding translation and calculating the correlation coefficient gave an unconvincing figure of 7%.

The real space rotation function was recalculated for

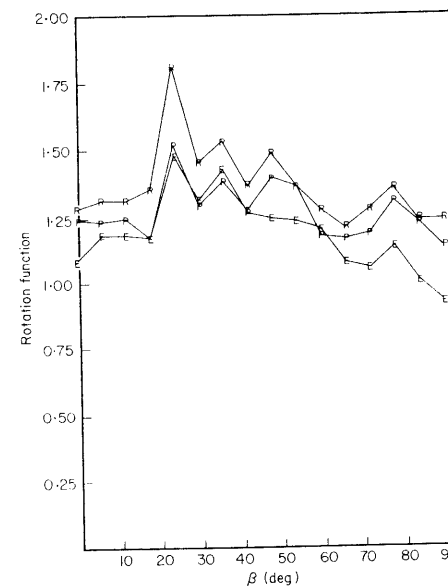


Figure 6. The real space rotation functions calculated with penicillopepsin (P), rhizopuspepsin (R) and endotheiapepsin (E) versus the pepsin target. The maximum values in each β -section are shown. Radii limits of 4 and 15 Å were used with all the reflections to 3 Å.

rhizopuspepsin using E values, rather than F terms, to 3.0 Å and found to be almost completely flat. The failure of the real space rotation search may be due to the low level of homology of the enzymes and the interpolation involved in obtaining the density values between the grid points of the target map, which had to be sampled quite coarsely (1.3 Å) for computational expedience. An encouraging result was obtained by calculating the cross-rotation function for the refined pepsin model from cycle 41, which had the largest peak on the origin, indicating that further refinement of this model was worthy of pursuit.

(f) Further refinement and use of pepsinogen co-ordinates

The model was rebuilt twice and care was taken to improve a number of poor surface regions that were penetrating neighbouring molecules, thereby lowering the R -factor to 36% after refinement. Another complete rebuild followed by refinement gave an R -factor of 35% and correlation coefficient of 75%.

At this point, the co-ordinates for the proenzyme, porcine pepsinogen ($R=17\%$ at 1.65 Å resolution), were kindly provided by Dr J. Remington (University of Oregon). These, minus the propept, were fitted to our pepsin model and refined with hexagonal reflection data, giving an R -factor of 36% (correlation coefficient, 74% at 2.3 Å). The resulting map was a vast improvement on the

previous one, e.g. excellent density for the 1st 2 disulphides and the enclosed loops. This was used to rebuild the pepsin model. The pepsinogen structure revealed that 2 strands and 1 helix in the pepsin model were 1 or 2 residues out of alignment. Fortunately, the 1st 14 residues (where enzyme and zymogen have different folds; James & Sielecki, 1986) and residues 156 to 161, which were absent from the pepsinogen co-ordinates, had good density in our pepsin model. The pepsinogen structure also allowed correction of a helix (225 to 236) where an extra residue had been included due to an error in the published sequence (Sepulveda *et al.*, 1975). Two refinement/rebuilding rounds rapidly lowered the *R*-factor to 23.5% (correlation coefficient, 90.1%) for data from 8 Å to 2.34 Å.

At this stage, the hexagonal pepsin model was least-squares fitted to the monoclinic pepsin co-ordinates and refined against the 2.0 Å reflection data, kindly supplied by Drs N. Andreeva and A. Fedorov (Moscow) and typed into the computer by the Birkbeck College Data Preparation Service. The *R*-factor after the 1st round of refinement was 27% and with 4 rounds of complete rebuilding and refinement this dropped to 21% (8.0 to 2.0 Å). Particular attention was needed in the vicinity of the active site loop (75 to 79) and 2 loops in the C-terminal domain that reside close to the active site; 240 to 245 and 294 to 299. The lower symmetry of the monoclinic pepsin crystals gave a significant reduction in the computer time needed for refinement, allowing rebuilds in difficult regions to be tested more quickly than in the hexagonal form. Electron density for the phosphate group attached to Ser68 became apparent and a phosphoserine residue was built in: 165 water molecule were placed in discrete features of density, lowering the *R*-factor to 19% (correlation coefficient, 92%) for data between 10.0 Å and 2.0 Å. Another 83 water molecules were added (total, 248) during a complete rebuild and another 10 cycles of refinement lowered the *R*-factor to 0.186 (correlation coefficient, 93%) for 13,644 reflections between 10.0 Å and 2.0 Å. The estimated r.m.s. co-ordinate error (Read, 1986) is 0.21 Å.

Rebuilding of the hexagonal form was continued with particular attention to some loop regions that were borrowed from the monoclinic structure, e.g. 240 to 245. Refinement lowered the *R*-factor to 0.224 for reflections between 7.0 and 2.34 Å. Gradual inclusion of water molecules to a total of 371 and refinement/rebuilding of the difficult regions lowered the *R*-factor to 0.190 for 12,429 reflections ($F \geq 3\sigma(F)$) between 8.0 Å and 2.34 Å (correlation coefficient, 91%). A lower estimate of the r.m.s. error in the co-ordinates is 0.11 Å (Read, 1986); refinement statistics are given in Table 3. All residues were built according to the chemical and gene sequences of porcine pepsin A (Sepulveda *et al.*, 1975; Lin *et al.*, 1989) with the exception of residue 254, which should be Asp but was left as Ala due to the paucity of the map in this surface region. The electron density indicates that residue 243 is Asp, in accordance with the chemical sequence, and not Tyr as found in the gene sequence. These co-ordinates have been deposited in the Brookhaven Protein Data-bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973, U.S.A., from where copies are available.

Least-squares fitting of the hexagonal pepsin co-ordinates to a set of independently refined monoclinic co-ordinates, kindly donated by Drs C. Abad-Zapatero, J. Erickson & T. Rydel, reveals that 85% of C α atoms are within 3.5 Å. The r.m.s. deviation for these regions is 0.8 Å. For at least half of the regions that differ, our map is good.

Table 3
Refinement statistics for P6₅22 pepsin after 11 complete rebuilds and 136 cycles

r.m.s. deviation from target bond distances	0.008 Å
r.m.s. deviation from target angle distances	0.011 Å
r.m.s. deviation from minimum non-bonded contact distance	0.010 Å
r.m.s. distance from least-squares main-chain planes	0.001 Å
r.m.s. distance from least-squares side-chain planes	0.001 Å

Resolution range (Å)	<i>R</i> -factor	Number of reflections
8.0 to 4.0	0.23	2868
4.0 to 3.3	0.17	2669
3.3 to 2.9	0.16	2432
2.9 to 2.6	0.18	1981
2.6 to 2.4	0.16	1574
2.4 to 2.3	0.16	905

Mean U_{iso} value for 2425 enzyme atoms = 0.37 Å² ($\sigma = 0.26$).
Mean U_{iso} value for 371 solvent atoms = 0.67 Å² ($\sigma = 0.20$).

3. Results and Discussion

The structure of the hexagonal form of porcine pepsin, refined to an *R*-value of 0.19 at 2.34 Å, defines the positions of 2425 enzyme atoms with reasonable thermal parameters. The polypeptide consists of 326 residues forming two topologically similar lobes characteristic of the aspartic proteinase fold observed in the fungal structures (Blundell *et al.*, 1985, 1990). Residues 1 to 175 form the N-terminal lobe and residues 176 to 327 constitute the C-terminal lobe. Least-squares fitting reveals that 33 residues per lobe overlap with a r.m.s. deviation of 1.6 Å. The sequence homology between the two lobes is 14%. The aspartic proteinase fold is

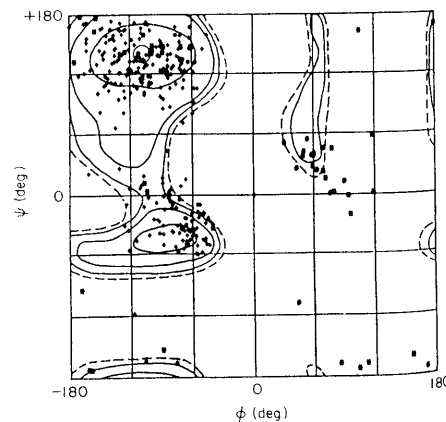


Figure 7. A Ramachandran plot for pepsin: (*) Gly; (●) Pro; (#) non-glycyl residue with a positive ϕ -angle; (+) all other residues.

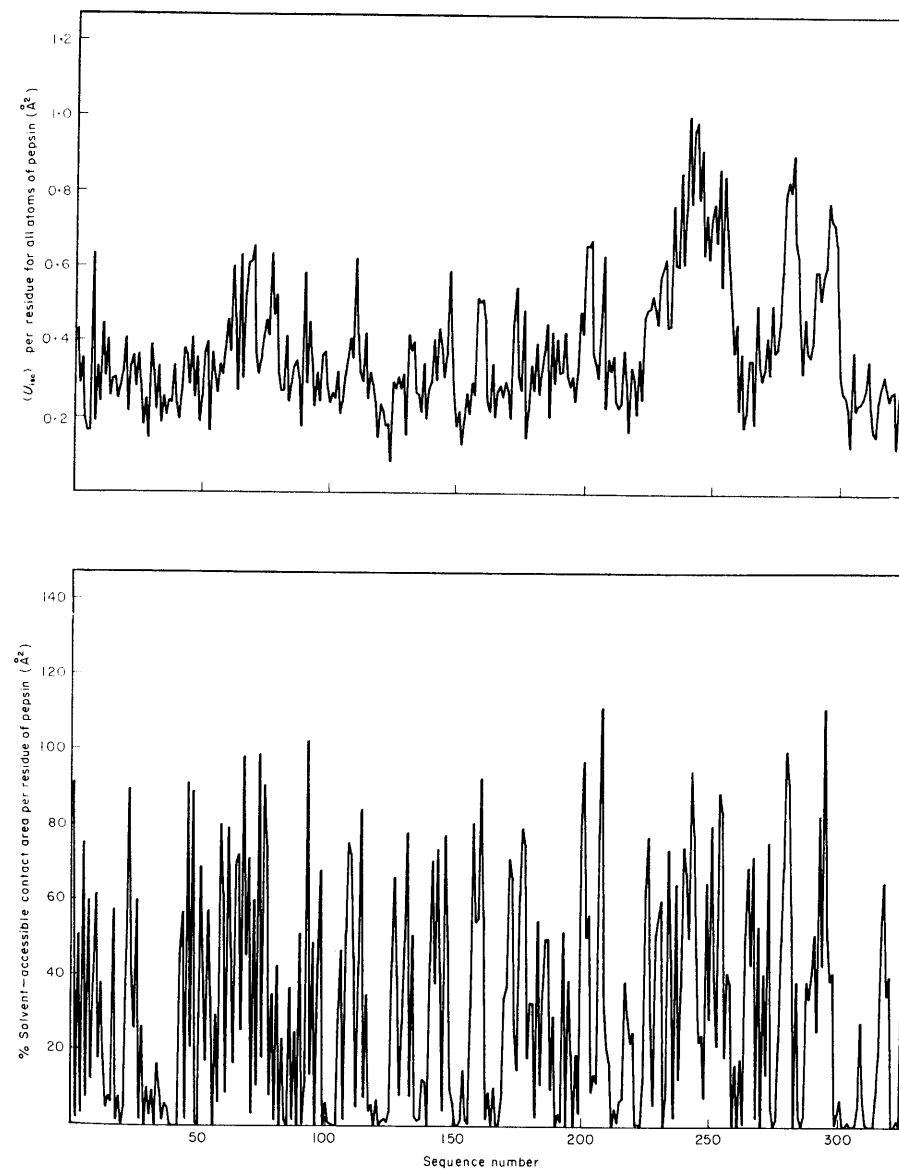


Figure 8. The average U_{iso} value (top) and solvent-accessible contact (bottom) are per residue of pepsin. The latter is for a 1.4 Å probe and is expressed as a percentage of the value for the extended peptide (Richmond & Richards, 1978).

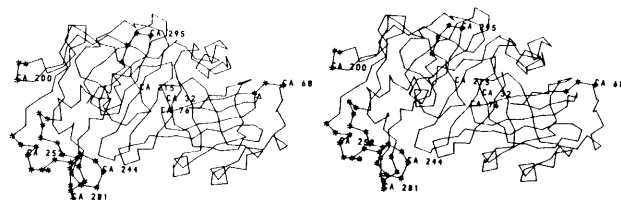


Figure 9. A stereo view of the $C\alpha$ backbone of pepsin with the most mobile regions (68 to 70, 200 to 202, 234 to 255, 278 to 283 and 295 to 298). Dotted spheres indicate where $\langle U_{iso} \rangle \geq 0.6 \text{ \AA}^2$ for more than 1 residue.

strongly conserved in the family, despite the low sequence homology of some members. The r.m.s. $C\alpha$ deviation of pepsin from the fungal enzymes is around $0.8 \text{ \AA}$ for about two-thirds of each enzyme, yet the level of sequence identities with pepsin are comparatively low, around 30%.

The fold is stabilized by a large number of hydrogen bonds. Calculating theoretical hydrogen positions with COMPARE (Dr A. M. Hemmings) and using a maximal acceptor-hydrogen distance of $3.5 \text{ \AA}$ and minimal angles at the hydrogen and acceptor atoms of 120° and 90° , respectively, a total of 188 main-chain-main-chain hydrogen bonds were found. Based on a simple donor-acceptor distance cutoff of $3.4 \text{ \AA}$, 179 main-chain-side-chain and 32 side-chain-side-chain possible hydrogen-bond interactions were found. The enzyme has a high proportion of serine and threonine residues, the total number being 60. Of these residues, 32 are involved in side-chain hydrogen-bond interactions with other residues of the enzyme.

The ϕ and ψ torsion angles of the molecule (Fig. 7) mostly lie within the allowed β -sheet and α -helix regions with the exception of glycyl residues (marked *) and 13 residues, mostly Asp and Asn, with positive ϕ angles (marked #) which lie close to or within the left-handed α -helix region. The unfavourable $C^{\beta}-O_{i-1}$ interaction can be compensated for by a hydrogen bond from the side-chain to the main chain with these residues. The structure is dominated by sheet, with 44% of residues lying in β -strands as defined by Kabsch & Sander (1983). The mean ϕ, ψ angles for these regions are $-112^\circ, 126^\circ$.

The average isotropic temperature factors for all the protein atoms (Fig. 8) lie within reasonable limits, the mean value for main-chain atoms being $0.36 \text{ \AA}^2$. Figure 8 indicates that the region from the end of helix b_c to around the second cystine residue (234-256) is somewhat more disordered or mobile than the remainder of the structure, although the electron density is reasonable with the exception of the 240 to 245 loop. The loops 278 to 283 and 294 to 298 also appear to be more flexible than the buried regions. Overall, the C-terminal lobe of the molecule appears to be more disordered than the N-terminal lobe. The most mobile regions are clustered (Fig. 9) on the surface of a distinct lobe (190 to 302), which corresponds to the flexible domain of endothermopepsin (see later and Šali *et al.*, 1989). As can be seen

in Figure 8, the regions of highest thermal mobility correspond to those with large solvent-accessible contact areas.

Residue 23, which resides in a highly conserved -Gly-Thr-Pro- sequence, has good density for a *cis*-proline, as suggested by Pearl for endothermopepsin (Pearl, 1983; Blundell *et al.*, 1985). Figure 10 shows the χ_1 distribution for pepsin excluding proline residues. There is a marked preference for the sterically most favourable g^+ conformation ($\chi_1 \approx 300^\circ$), in which the side-chain is *trans* to the carbonyl group as viewed along the $C^{\beta}-C^{\alpha}$ bond. This was observed by Janin *et al.* (1978). The least favourable g^- conformation, where the side-chain is *trans* to the $C^{\alpha}-H$ bond ($\chi_1 \approx 80^\circ$), is populated mainly by threonine, isoleucine and buried aromatic residues. The dominance of the g^+ conformation is reflected in the mean χ_1 value of 270.1° for the whole molecule (Table 4).

Pepsin possesses three disulphide bridges at positions 45-50, 206-210 and 250-283. These represent some of the regions that differ most from the penicillopepsin starting model. The electron density for these regions is shown in Figure 11. The disulphide link at 206-210 appears to replace the main-chain of the conserved strand preceding the second catalytic aspartate residue.

Residue Ser68 of pepsin is phosphorylated, although there is no electron density for the phosphate group in the hexagonal crystal form. The chain near the phosphorylation site is exposed both

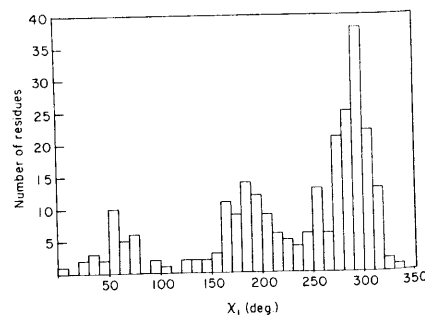


Figure 10. The χ_1 distribution for all residue types except Gly, Ala and Pro.

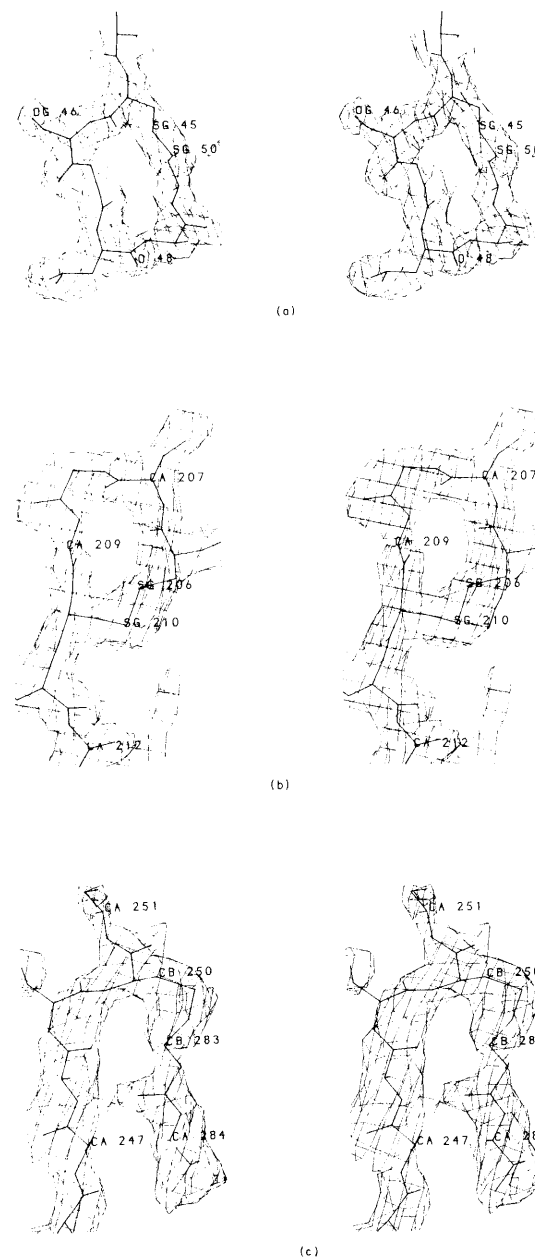


Figure 11. The disulphide links of pepsin at 45-50, 206-210 and 250-283 shown in stereo (downwards order). Thin lines show the $2F_o - F_c$ electron density.

Table 4
Internal co-ordinate analysis of pepsin at 2.3 Å

Parameter	ϕ (°)	ψ (°)	ω (°)	χ_1 (°)	χ_2 (°)	τ (°)	N-C ^a (Å)	C ^a -C (Å)
Mean	-97.8	115.9	179.1	-89.9	162.3	110.9	1.47	1.52
s.d.	30.8	45.8	13.6	43.3	48.6	1.6	0.01	0.01
Sample	325	325	325	273	176	326	326	326

in the pepsin and proenzyme crystal structures. Chicken pepsin is glycosylated at an asparagine residue equivalent to residue 68 (Baudys *et al.*, 1982). Asparagine 67 is *N*-glycosylated in yeast proteinase A (Dreyer *et al.*, 1986), cathepsin D (Shewale *et al.*, 1985) and human renin (Imai *et al.*, 1983). This region may be involved in the targeting of aspartic proteinases, since attachment of a mannose 6-phosphate to lysosomal hydrolases, e.g. cathepsin D, may be essential for targeting to the lysosomes (Sly & Fischer, 1982).

The crystal packing of the hexagonal form appears to be looser than in the monoclinic form (Andreeva *et al.*, 1984) where, for example, the active site flap forms hydrogen bonds with a neighbouring molecule. This may account for the failure to obtain the monoclinic form with several crystalline pepsin inhibitor complexes (J. Remington, personal communication). It may also account for the greater apparent disorder of the carboxyl-rich loop at 240 to 245 in the hexagonal form, which is reasonably well defined in the monoclinic structure.

Pepsin contains several salt-bridges, two of which involve interactions between a residue from one of the dyad-related helices (h'_N and h'_C) and a residue from the opposite lobe, e.g. Asp138 (h'_N), interacts with Arg316 (C-lobe). The electron density for this

salt-bridge is shown in Figure 12. Although this is one of the least common modes of arginine-carboxyl interaction according to Singh *et al.* (1987), the electron density defines the interaction well. This ion-pair may be important for the stability of the pepsin fold, since chemical modification of Arg316 with butane-2,3-dione partially inactivates the enzyme (Huang & Tang, 1972) as does treatment with phenylglyoxal (Kitson & Knowles, 1971), an arginine-specific reagent. The other arginine of pepsin (308 in helix h'_C) appears to form a salt-bridge with the highly conserved Asp11. Accordingly, residue 308 is highly conserved as a basic residue, being mainly arginine in the mammalian aspartic proteinases and lysine in some of the fungal enzymes. There are only two other basic residues in pepsin, His53 and Lys320. The imidazole of His53 is sandwiched between the carboxyl groups of Asp118 and Asp52, and makes a hydrogen bond to the carbonyl of residue 117. Lys320 is exposed to solvent, although it has good side-chain density.

(a) Secondary structure

Pepsin consists largely of β -sheet, with six small, right-handed α -helical segments, the longest being helix h_C , which spans residues 225 to 236. The

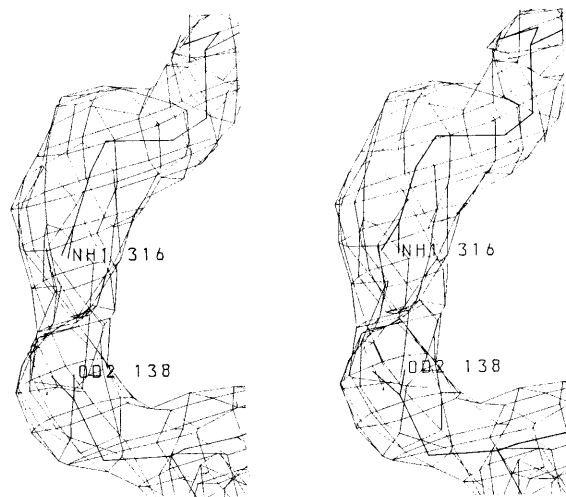


Figure 12. A stereo view of the electron density (thin lines) for the salt-bridge between Asp138 and Arg316.

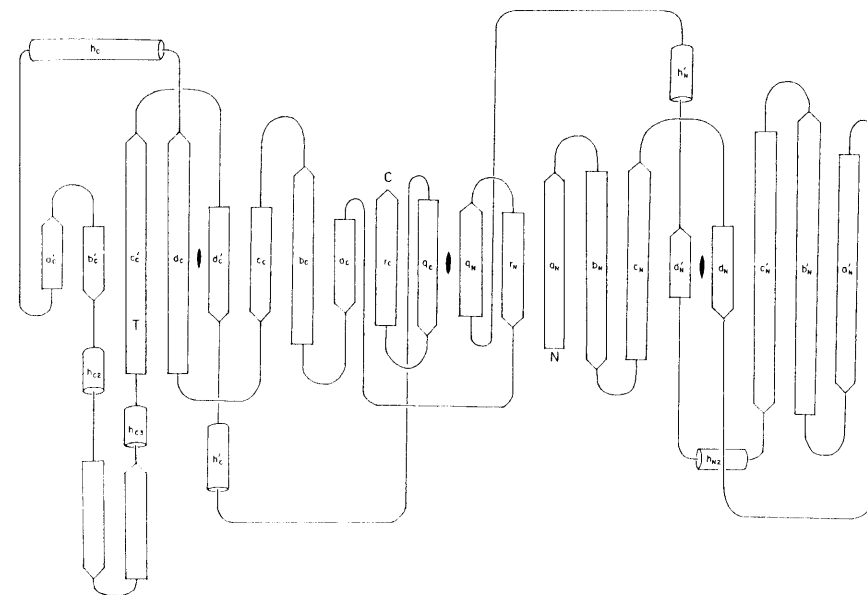


Figure 13. A diagram of the pepsin molecule showing the topological intra- and inter-lobe dyad axes (only β -strand lengths are to scale). Note, some of the strands we have defined are interrupted by bends, short helices and, in one case, a β -hairpin turn (T).

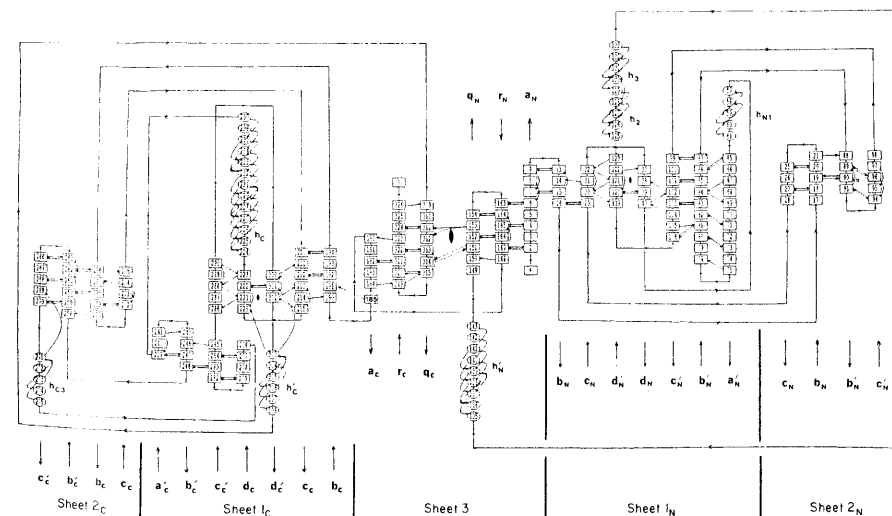


Figure 14. A schematic view of the hydrogen bonds that stabilize the pepsin structure; prepared with the assistance of the program HBOND (G. Hutchinson & J. M. Thornton), which shows hydrogen bonds in the Kabsch & Sander (1983)-defined sheets and helices. A large dyad marker shows the topological interlobe 2-fold axis and smaller markers show the intra-lobe axes.

pattern of secondary structure elements forming the two domains is illustrated by Figure 13. The sheets and helices are named by analogy with the scheme adopted for endothepepsin (Blundell *et al.*, 1985, 1990), which emphasizes the intra and inter-lobe 2-fold symmetry. In each lobe, strands labelled a, b, c and d are related to a', b', c' and d' by the intra-lobe dyad, and these strands are related to their equivalents in the opposite lobe by the inter-lobe dyad, e.g. a_N and a_C.

The Kabsch & Sander (1983) criteria were used to define regions of secondary structure. Helices are identified from consecutive hydrogen bonds to residues three, four or five places along the chain and β -sheets are defined by the parallel or anti-parallel hydrogen-bonding pattern of neighbouring

Table 5

The β -strands forming sheets 3, 1_N, 1_C, 2_N and 2_C defined by the hydrogen-bonding criteria of Kabsch & Sander (1983)

Initial residue	Final residue	Strand
A. Sheet 3		
Gly2	Leu6	a _N
Ser163	Leu167	r _N
Leu150	Tyr154	q _N
Tyr310	Asp315	q _C
Lys320	Val326	r _C
Asn180	Pro183	a _C
B. Sheet 1_N		
Glu7	Tyr9	a _N
Glu13	Gly16	b _N
Val29	Asp32	c _N †
Gly119	Gly122	d _N †
Leu38	Val40	d _N †
Gln99	Thr106	c _N
Ser79	Asp87	b _N
Glu65	Thr74	a _N
C. Sheet 1_C		
Gln191	Leu194	b _C †
Cys210	Val214	c _C †
Tyr300	Leu302	d _C †
Gly221	Gly223	d _C †
Ser286	Gly290	c _C §
Met246	Ile248	b _C
Ser239	Glu240	a _C
D. Sheet 2_N		
Gln25	Thr28	c _N
Thr17	Ile20	b _N
Thr88	Val91	b _N
Ile94	Asn98	c _N
Sheet 2_C		
Glu202	Ala205	c _C
Ser196	Met199	b _C
Ile259	Ile263	b _C
Val266	Leu270	c _C

† These strands form ψ_1 .

‡ These strands form ψ_2 .

§ This strand is preceded by a short β -hairpin: 276-279, 282-285 which forms an anti-parallel β -sheet with a_Cb_C (see Fig. 14). (Note, strand a_N forms part of sheets 3 and 1_N (Fig. 14) whereas the dyad related strand a_C forms part of sheet 3 but only makes one hydrogen bond with sheet 1_C).

residues to another section of the chain. Main-chain ϕ, ψ angles are not used in this definition. The hydrogen-bonding pattern of the molecule illustrated by Figure 14.

(b) β -Sheets

The central underlying sheet of the aspartic proteinase fold is formed by six anti-parallel β -strands: a_N, r_N, q_N, q_C, r_C and a_C (Table 5). These form sheet 3, which resides beneath the strands forming the base of the active site cleft. Upon sheet 3, which consists of strands from both the N and C-terminal lobes, are two β -sheets, one per domain, each formed by seven or eight strands. The residues forming these sheets, 1_N and 1_C, are defined in Table 5. As can be seen in Figures 13 and 14, the pattern of these strands is similar in both domains and is related by a pseudo-2-fold axis. Both of these sheets are highly twisted (Fig. 15) and adopt different three-dimensional structures, although the topologies are identical. The structural differences between these two sheets account for the smaller number of equivalences (33) obtained when the two lobes of the enzyme are least-squares fitted. The b, c, b' and c' strands are the most twisted, since they fold to form extra sheets beneath 1_N and 1_C. The strands named b_C and c_C are the most distorted, each being interrupted by turns, short helices and in one case a β -hairpin (see Fig. 14).

Strands c_N d_N d_N and d_N c_N in sheet 1_N and c_C d_C d_C and d_C c_C in sheet 1_C form structures resembling the Greek letter psi (Fig. 14). These structures are parallel to sheet 3, and the C-termini of the c_C and c_N strands associate to form the catalytic centre. The stabilizing arrangement of

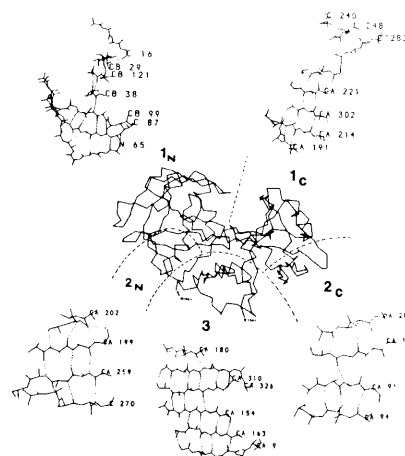


Figure 15. A collage showing the different β -sheets that form the pepsin molecule: 1_C, 2_C, 3, 2_N, 1_N (clockwise from top right).

Table 6
The main chain-main chain hydrogen-bonding geometry of pepsin (calculated using COMPARE by Dr A. M. Hemmings)

	Average value	Standard deviation
A. Overall molecule		
H...O length (Å)	2.18	0.38
O...H-N (deg.)	152.2	14.5
H...O=C angle (deg.)	140.5	19.7
B. β-Sheets (as defined by Kabsch & Sander (1983) criteria)		
H...O length (Å)	2.06	0.30
O...H-N angle (deg.)	154.8	14.9
H...O=C angle (deg.)	151.1	11.6
C. α-Helices (hydrogen bonds formed by residues of the best defined α-helical regions h_N, h_C, h_{C3} and h_{C4})		
H...O length (Å)	2.16	0.27
O...H-N angle (deg.)	150.0	13.9
H...O=C angle (deg.)	135.8	19.1

hydrogen bonds is called the fireman's grip (Pearl & Blundell, 1984) involving threonine residues 33 and 216, and the main chain occurs here. The sequence homology between both domains and with other aspartic proteinases is greatest at the interface, as is the 2-fold structural pseudo-symmetry between the two lobes.

There are two extra anti-parallel sheets, 2_N and 2_C (Table 5), which arise from bends in the b, c, b' and c' strands of each lobe and fold beneath sheets 1_N and 1_C. The arrangement of these β -sheets is shown in Figure 15.

The mean hydrogen-bond parameters for the whole molecule, β -sheets and α -helices are given in Table 6. It can be seen that the β -strands have somewhat better hydrogen-bonding geometry than α -helices and the overall molecule, although the differences are no larger than the standard deviations.

(c) α -Helices

Segments forming right-handed α -helices are shown in Figure 13 (and see Table 7). The positions

Table 7
The helical segments of porcine pepsin

Helix	Initial residue	Final residue	Class	Comment
h _{N2}	Ser110	Tyr114	α_R	Distorted
h _{N1}	Val136	Asp142	α_R	
h _C	Thr225	Ile236	α_R	
h _{C2}	Ser252	Ser255	α_R	Distorted
h _{C3}	Pro272	Tyr275	α_R	
h _{C4}	Asp304	Ile307	α_R	
Others				
h _N	Leu48	Cys50	3 ₁₀	
h _{N1}	Pro58	Asp60	3 ₁₀	
h ₂	Pro126	Ile128	3 ₁₀	
h ₃	Ala130	Gly132	3 ₁₀	

of these helices further emphasize the topological intra and inter-domain 2-fold symmetry. Helices h_N, h_C and h_{C'} are equivalent in that they all occur after d strands (d_N, d_C and d_{C'}, respectively). Although this does not apply to d_N in pepsin, there are two 3₁₀ helices shortly after this strand at 48 to 50 and 58 to 60. Two more 3₁₀ helices occur immediately before h_N. The dyad-related α -helical stretches are shown in Figure 16.

All of the helices with the exception of helix h_{C'} (304 to 307) are partly exposed and have some amphiphilic character especially the longest helix, h_C (225 to 236), which has a solvated surface and a buried hydrophobic side. Of these helices, only h_{N2} (110 to 114) forms part of the substrate-binding cleft.

(d) Loops

Loops are regions of polypeptide between two segments of regular secondary structure. The loops of pepsin vary from relatively short β -hairpins to long stretches of polypeptide containing one or more reverse turns. Such turns are defined by Kabsch & Sander (1983) as regions enclosed by single hydrogen bonds from residues i to $i+3$, $i+4$ or $i+5$ and are shown for pepsin in Table 8.

One class of turn is the β -hairpin, which can be classified using X:Y nomenclature (Sibanda *et al.*, 1989, and references therein). For example, if the residues at the end of a β -ladder enclosing a two residue loop are connected by two hydrogen bonds, the loop is classed as 2:2 but if one of these hydrogen bonds is not formed, the loop becomes a 2:4 hairpin. There are examples of both these types of β -hairpin in pepsin (Table 8, where the loops are named after the strands they connect, e.g. a_Nb_N). The topologically related loops b_Nc_N (Gly, Gly) and b_Cc_C (Asn, Gly) form standard 2:2 β -hairpins with the observed preference for Asn, Asp or Gly at position L1 and Gly at L2. The q_Nr_N hairpin is interesting, since pepsin has a two residue insertion here relative to endothepepsin, which has a 7:7 hairpin at this loop. The insertion of two residues results in a 9:9 turn in pepsin in accordance with the rules of Sibanda *et al.* (1989).

Other reverse turns include the conserved Asp-Thr-Gly sequences at 32 to 34 and 215 to 217, which both form type I turns, these being the most common type for pepsin (Table 8) and for proteins in general (Wilmot & Thornton, 1988). The disulphide bridge from 206 to 210 encloses the only example of a type II turn, which is the second most common type.

(e) The aspartate dyad

The presumed catalytic site consists of the nearly co-planar carboxyl groups of residues 32 and 215, which are held in close proximity by a network of hydrogen bonds involving the hydroxylic side-chains three residues along from each aspartate residue and the main-chain nitrogen atoms of the

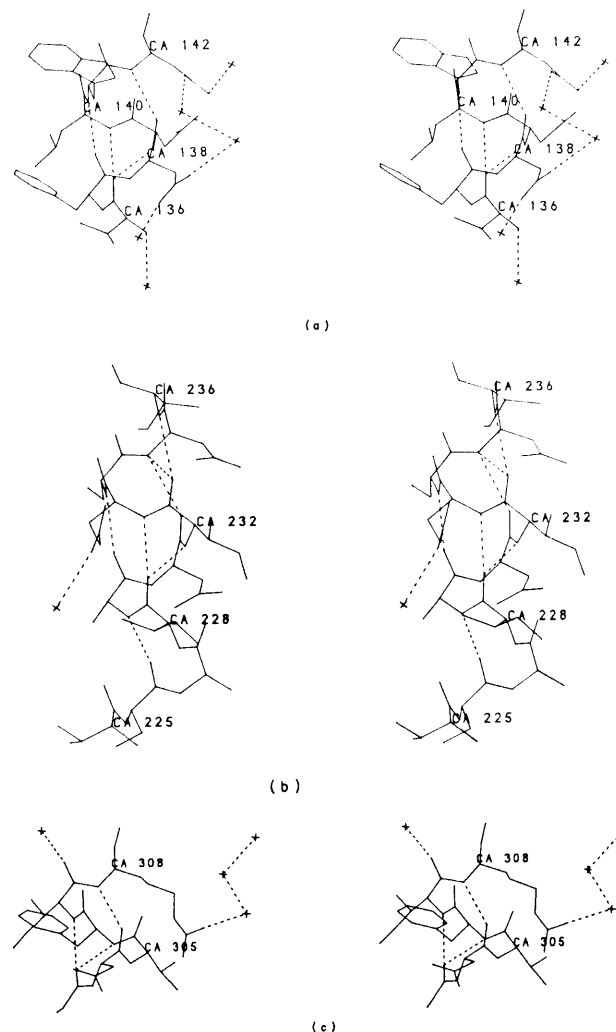


Figure 16. Stereo views of the topologically related h'_N , h_C and h'_C α -helical segments of pepsin (downwards order).

conserved glycine residues. More precisely, the outer carboxyl oxygen atom of Asp215 is within hydrogen-bonding distance of the Thr218 hydroxyl group and the inner oxygen atoms of both aspartate residues appear to accept hydrogen bonds from the amide groups of glycine residues 34 and 217. The hydrogen bonds to outer carboxyl oxygen atoms are somewhat variable in the different aspartic proteinase structures and are presumably not essential for

catalysis, since residue 218 is Ala in human renin. The bound water molecule (W1), as shown in Figure 17, is within hydrogen-bonding distance of both essential carboxyl groups forming contact distances of 2.7 Å, 2.9 Å, 3.4 Å and 2.9 Å to aspartate oxygen atoms 32 O^{δ1} (outer), 32 O^{δ2} (inner), 215 O^{δ1} (inner) and 215 O^{δ2} (outer), respectively. The water and carboxyl group atoms including O^δ atoms have a r.m.s. deviation from planarity of

Table 8
The β -hairpin loops of porcine pepsin classified according to Sibanda et al. (1989)

Loop	H-bond†	Type	Residue and ϕ/ψ angles (deg.)
$a'_N b_N$	NH13–O9	3:3 I'	Leu10(53, 41) Asp11(44, 25) Thr12(–125, –29)
$b'_N c_N$	NH25–O20	4:4	Gly21(100, –173) Thr22(–142, 112) Pro23(–59, 169) Ala24(–79, 127)
$a'_N b'_N$	NH75–O78	2:4	Gly76(–42, 32) Thr77(–123, –54)
$b'_N c'_N$	NH94–O91	2:2 I'	Gly92(68, 23) Gly93(92, –1)
$q'_N r'_N$	NH164–O154	9:9 I'	Asn158(30, 52) Asp159(74, 1)
$b'_C c'_C$	NH202–O199	2:2 I'	Asp200(62, 25) Gly201(78, 2)
$a'_C b'_C$	NH241–O245	3:5 I	(240–245) loop disordered in hexagonal but is 3:5 type I in $P2_1$ pepsin
$b'_C c'_C$	NH266–O263	2:2 I'	Asn264(57, 24) Gly265(97, –19)
$c'_C d'_C$	NH282–O279	2:2	Asp280(–67, 17) Asp281(–103, –39)
$q'_C r'_C$	NH320–O315	4:4	Arg316(–83, –5) Ala317(–78, –50) Asn318(–111, 8) Asn319(53, 55)

Other turns: defined by Kabesh & Sander (1983)

Loop	H-bond $_{i,i+3}$	Type	Torsion angles ($i+1$ and $i+2$ residues)
$c'_N d'_N$	NH35–O32	I	Thr33(–82, 6) Gly34(–105, 5)
$d'_N a'_N$	NH45–O42	I	Val43(–80, –8) Tyr44(–90, 18)
$d'_N a'_N$	NH64–O61	I	Ser62(–73, –29) Thr63(–110, 11)
$r'_N a'_C$	NH174–O171	I	Ser172(–53, –9) Ser173(–69, –22)
$a'_C b'_C$			Gly188(–167, –80) Tyr189(–82, 180)
$c'_C d'_C$	NH209–O206	II	Ser207(–70, 131) Gly208(70, 16)†
$c'_C d'_C$	NH218–O215	I	Thr216(–77, –12) Gly217(–107, 13)

† The last hydrogen bond of the β -ladder.

‡ Turn enclosed by disulphide bridge.

0.02 Å. The water molecule appears slightly closer to Asp32 than to Asp215, which is contrary to what is found for the fungal aspartic proteinases (Suguna et al., 1987b), although this may be due to the inferior resolution of the pepsin analysis.

The arrangement of residues forming the fireman's grip beneath the catalytic dyad is shown in Figure 17. In this network, the hydroxyls of the residues following each aspartate residue form hydrogen bonds to each side of the catalytic aspartate residue in the opposite lobe. The nearly identical environments of the catalytic carboxyl groups provided by the hydrogen bonds and dyad-related

isoleucine residues 120 and 301 make it difficult to identify the negatively charged carboxyl group, and it has been suggested that both catalytic residues of aspartic proteinases may be protonated, although only one carboxyl group would be negatively charged at any one time (Pearl & Blundell, 1984). Proof of this awaits a neutron diffraction study.

(f) The active site cleft

The overall appearance of the active site is similar to that of endothiapepsin (Blundell et al., 1985, 1990), with the exception of the $c'_C d'_C$ loop (292 to

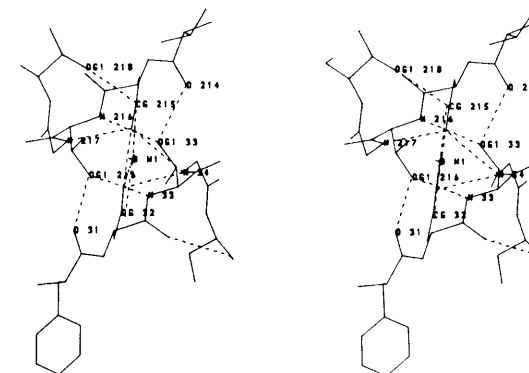


Figure 17. A stereo view of the active site aspartate residues and the fireman's grip.

299), which projects over the prime side of the cleft in a similar manner to the active site "flap" (loop 73 to 79) in the N-terminal lobe. The $c'_c d'_c$ loop is topologically related to a helix (h_{N2}) that occurs in the $c'_N d'_N$ loop and forms part of the S_1 and S_3 pockets. It might be expected from Figure 14 that the loop 240 to 245, which is related by the inter-lobe dyad to the active site flap in the N-terminal lobe, would cover the cleft to some extent. However, the 240 to 245 loop adopts a more exposed orientation with high temperature factors.

In an effort to define the residues that form the binding subsites of this enzyme, the co-ordinates of endotheiapepsin complexed with the pepsin inhibitor H-256 (Cooper *et al.*, 1987) were least-squares fitted to the pepsin structure. This potent inhibitor was designed from the sequences found at pepsin cleavage sites (Powers *et al.*, 1977) and has a reduced bond Phe^R-Phe analogue ($-CH_2-NH-$) in place of the scissile bond (Hallett *et al.*, 1985). The potency of the inhibitor for endotheiapepsin ($K_i = 60$ nM) is only slightly less than that for porcine pepsin ($K_i = 40$ nM). The inhibitor co-ordinates fit the pepsin active site with few unfavourable contact distances.

Aspartic proteinase inhibitors bind by forming a conserved set of hydrogen bonds involving the main chain of the ligand (Foundling *et al.*, 1987). For example, the Thr219 side-chain of endotheiapepsin forms a conserved hydrogen bond to the amide nitrogen of P_3 . In pepsin, residue 219 is serine and the side-chain appears to be appropriately oriented to form an equivalent hydrogen bond with the inhibitor. The NH group of Ser219 appears to donate a hydrogen bond to the P_3 carbonyl of H-256, as is found for the endotheiapepsin complex. Conserved hydrogen bonds involving the carbonyl groups of dyad-related glycine residues 217 and 34, and the NH groups of P_1 and P_2 , respectively, are also found in the model.

A number of potential hydrogen bonds to the inhibitor, in addition to those found with endotheiapepsin, were apparent. The P_3 threonine of the inhibitor is in a good position to form hydrogen bonds with the Glu13 carboxyl group *via* the side-chain hydroxyl group. The Thr77 hydroxyl group, in the active site flap of pepsin, appears able to form a hydrogen bond with the amide nitrogen atom of P_2 .

The bulky Phe- CH_2 -NH-Phe group of H-256 is accommodated remarkably well by the pepsin active site. These binding subsites consist of hydrophobic residues, most of which were predicted from the homology with endotheiapepsin (Cooper *et al.*, 1987). Tyr75, Thr77, Trp39, Ile30, Phe117, Leu112, Phe111 and Ile120 seem to form a hollow that accommodates the P_1 phenyl side-chain of H-256. The S'_1 pocket, although more exposed, is formed by the residues Ile213, Ile301, Tyr189, Met290 and Val292, as well as the main chain of the active site flap. The hydroxyl group of Tyr189 appears to donate a hydrogen bond to the carbonyl group at the P_2 position. Involvement of this residue in

substrate binding was deduced from chemical modification by *p*-nitrophenyldiazonium chloride which inhibits the enzyme (Zavada *et al.*, 1977). Alkylation of the S'_1 residue Met290 by α -bromo-4-amino-3-nitroacetophenone below pH 3 partially inactivates the enzyme (Tarasova *et al.*, 1980). This is consistent with the closeness of the Met side-chains to the catalytic centre.

The length of the 292 to 299 loop in aspartic proteinases is quite variable, gastricsins having the longest and fungal pepsins the shortest. This region may have some influence on the specificity in view of its disposition with respect to the active site cleft. In human renin, it contains three consecutive proline residues, which must rigidify this region of the structure (Sibanda *et al.*, 1984). In pepsin, this loop extends over the cleft to a far greater extent than in endotheiapepsin. However, the prime side residues of the inhibitor in the model appear fairly exposed and are not covered by the loop, although it projects towards this region. In contrast, the central residues of the inhibitor are covered by the 73 to 79 flap.

(g) Solvent structure

In the final refinement, 371 water molecules were included in the asymmetric unit. These and symmetry-related solvent molecules make a total of 137 hydrogen-bond contacts to each protein, with a mean length of 3.02 Å. Protein oxygen atoms dominate these interactions, with main-chain carbonyl groups, side-chain carboxyl oxygen atoms and hydroxyl groups contributing 31%, 21% and 22%, respectively. In contrast, main-chain nitrogen atoms contribute to only 10% of the solvent hydrogen bonds.

Visual inspection revealed that seven solvent molecules are buried well within the protein. These waters are all hydrogen-bonded to enzyme groups with a notable preference for carbonyl oxygen atoms (Table 9), thereby indicating the positions of the hydrogen atoms in most cases. Three of these water molecules (W38, W40 and W305) appear to form a buried cluster with extensive interactions with polar enzyme groups. The buried solvents have U_{iso} values that are significantly lower than the average value of 0.67 Å² for all water molecules built in.

Table 9
Buried solvent sites

Solvent	U_{iso} (Å ²)	Residues within hydrogen-bonding range
W16	0.29	O216, NH14
W38	0.16	OW40, O134, O100
W40	0.34	O38, NH38, N ²³⁷ , OW38
W96	0.20	O124, O ²¹³⁸ , N ⁷³¹⁶
W241	0.34	O112, O117
W305	0.27	NH136, O36, O122
W335	0.42	O307, O153, NH155

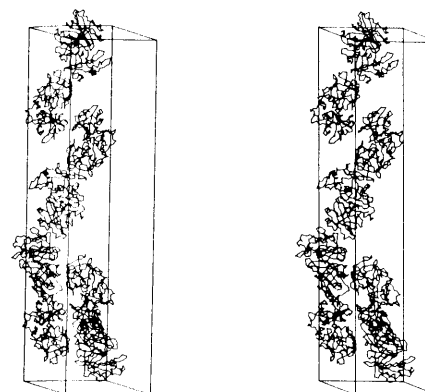


Figure 18. The arrangement of the molecules in the hexagonal lattice.

A small rigid group conformational change was observed for endotheiapepsin when inhibitors are bound. The change consists of a relative movement of two rigid groups, one of which forms a distinct domain (residues 190 to 302) in the C-terminal lobe (Sali *et al.*, 1989). In endotheiapepsin, a solvent-filled groove partially separates this lobe from the remainder of the structure. A similar feature is apparent in pepsin and the groove contains a tightly bound cluster of water molecules (W198, W373 and W370).

Due to the arrangement of the molecules in the crystals, the active site cleft is contiguous with a large solvent-filled cavity that is roughly 30 Å in one dimension. This indicates that it may be possible to form inhibitors complexes by soaking the crystals. A crystal packing diagram is shown in Figure 18.

(h) Conservation and invariance

From an alignment of 24 of the known aspartic proteinase sequences (Blundell *et al.*, 1990), it is clear that apart from the catalytically essential DTG sequences, there is a number of other residues that are absolutely or highly conserved within the family. The pepsin structure was examined in an effort to account for this conservation.

There are ten absolutely conserved glycine residues, most of which occur at turns. The presence of side-chains at the glycine positions 34 and 217 would disrupt the planarity of the catalytic carboxyl groups. The absolutely conserved glycine residues 119 and 122 lie at each end of the β -strand d'_N , which is beneath the active site cleft. Examination reveals that the presence of a side-chain at these positions would disrupt the side-chains of Trp39 and Ser35, both of which are absolutely conserved due to their hydrogen-bonding role in the

active site. Trp39 donates a side-chain hydrogen bond to the hydroxyl group of Tyr75, which forms one side of the S'_1 pocket and the presumptive oxyanion hole. The side-chain of Ser35 is hydrogen-bonded to the outer carboxyl oxygen of Asp32 in the catalytic dyad in some of the aspartic proteinase structures e.g. penicillopepsin. Similar arguments apply to the absolutely conserved glycine residue at position 303, which is at the end of strand d'_C . A side-chain at this position would disrupt Thr218, which stabilizes the conformation of Asp215, although this threonine residue is not essential (e.g. Ala in human renin). These glycine residues, 122 and 303, lie within highly conserved and dyad-related -hydrophobic-hydrophobic-Gly- sequences that are present in the mammalian, fungal and dimeric retroviral aspartic proteinases. These residues lie within the central strands of the ψ structures (d'_N and d'_C). Their conservation appears necessary to allow a twist in the chain and to avoid major steric disruption of the catalytic site residues.

Many of the absolutely or highly conserved aromatic residues have a side-chain hydrogen-bonding interaction with another conserved side-chain or the main-chain of the enzyme, thereby stabilizing the fold of the protein. For example, the side-chains of Trp190 and Tyr275, both highly conserved, hydrogen-bond with the main-chain carbonyl groups at positions 123 and 257, respectively. The hydroxyl group of invariant Tyr14 forms hydrogen bonds with the main-chain carbonyl group of Ser156 and the guanidinium group of Arg308, which is a highly conserved residue. Tyr14 lies at the end of the activation peptide and may act as an anchor during the refolding that occurs on activation (see later). A similar role for several aromatic residues of trypsinogen has been proposed (Huber & Bode, 1978).

In general, charged residues occur at the surface of proteins and are therefore not expected to be highly conserved. However, there is a number of highly or absolutely conserved polar residues in addition to the catalytic aspartate residues. Inspection reveals that most of these residues are buried

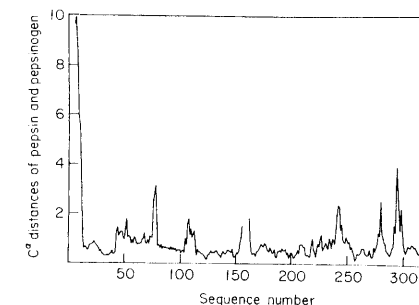


Figure 19. The $C^\alpha-C^\alpha$ distances for pepsin and pepsinogen after least-squares fitting using MNYFIT (Sutcliffe *et al.*, 1987).

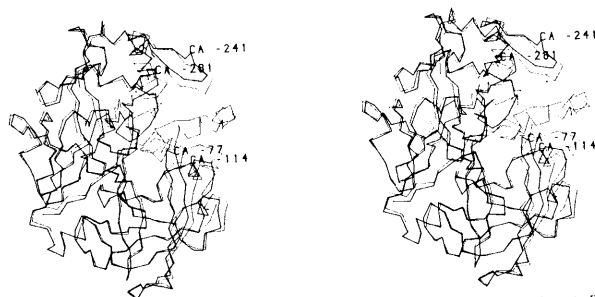


Figure 20. (^{100} tracings showing the close fit of pepsin (thick lines) to pepsinogen (thin lines). The activation peptide may be seen filling the active site cleft.

and are involved in hydrogen bonds to the main chain or other conserved residues. For example, the side-chain NH_2 group of Gln99 (which is conserved in 23/24 sequences) is hydrogen-bonded to the main-chain carbonyl groups of residues 100 and 134 and

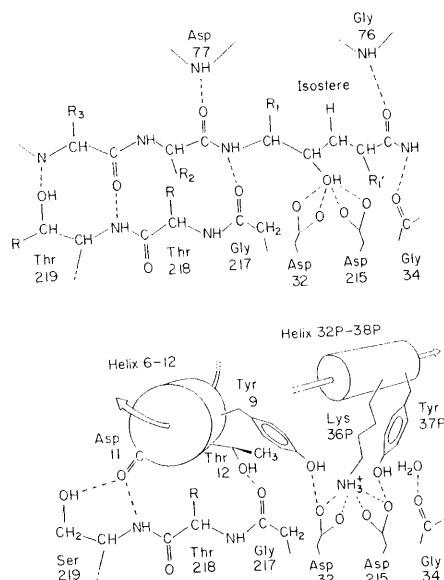


Figure 21. A schematic view of the interactions between aspartic proteinase inhibitors and the active site cleft (top) compared with the propeptide interactions of pepsinogen (below). Broken lines show hydrogen-bond contact distances.

Figure 22. Plots showing the variation in U_{100} versus sequence number of pepsin (top), pepsinogen (middle). The difference in U_{100} (pepsin minus pepsinogen) is shown last.

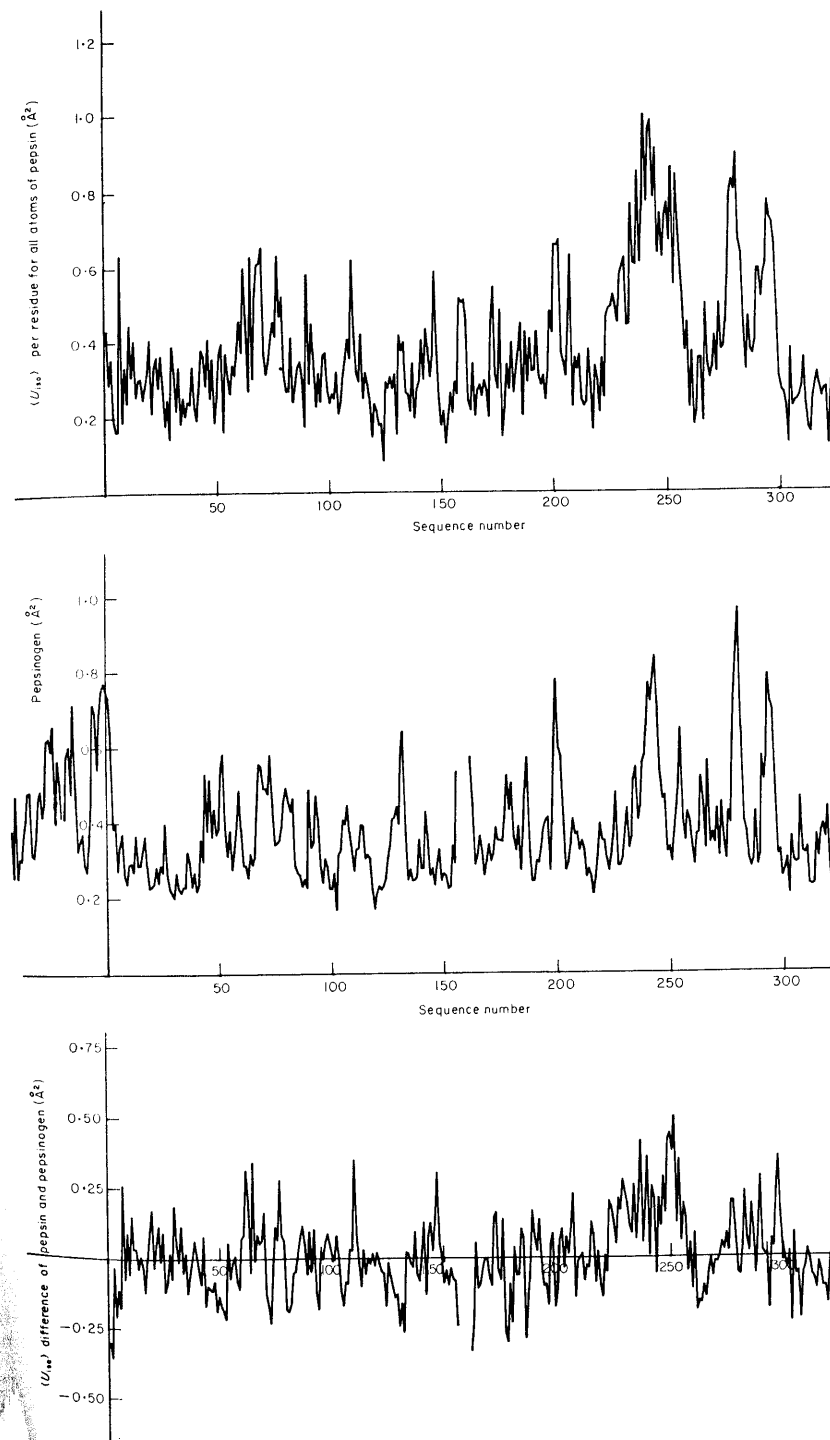


Fig. 22.

the carboxyl group of Asp87 (23/24) forms hydrogen bonds with the side-chains at 61 and 63, which are invariably Ser or Thr. Two other conserved buried side-chains forming hydrogen bonds to main chain carbonyl oxygen atoms are those of Asn37 (invariant as Asn or Asp) and Ser42 (20/24), which interact with the main chain at positions 130 and 104, respectively. Similarly, the Asp118 side-chain is hydrogen-bonded to the main-chain NH group of residue 54, and the carboxyl group of Asp304 interacts with the carbonyl groups of residue 216. Asp304 has been implicated by site-directed mutagenesis as one factor in lowering the pH optimum of catalysis (Yamauchi *et al.*, 1989) via its network interaction with Asp215. Changing Ala304 of renin, which is maximally active at neutral pH, to Asp lowered the optimum pH by approximately 0.5 pH unit. The highly conserved residues Arg308 and Asp11 appear close enough to interact via a salt link and residue 315, which is absolutely conserved as Asp or Asn, appears to form a side-chain hydrogen bond to the main-chain carbonyl group of Asp149. However, similar arguments could not be applied to Asp171, which is very highly conserved (22/24) and appears to lie on the surface on the opposite side of the enzyme to the active site cleft. The aspartate residue at position 11 shows a similar degree of exposure and high degree of conservation. These side-chains were found to form salt-bridges with the propeptide of pepsinogen involving 3P Lys and 13P Arg, respectively (James & Sielecki, 1986). These residues may therefore be involved in the pH-dependent denaturation of the propeptide that precedes the autolytic processing of proenzyme to enzyme. With pepsin, this is achieved by the large drop in pH as pepsinogen is secreted into the stomach. Activation first involves autolytic cleavage at 16P-17P. The remainder of the activation peptide is sufficiently exposed for further cleavage that occurs at 44P-1 unleashing fully processed pepsin. Tyr9 is also highly

conserved, and yet has no side-chain hydrogen bonds except to solvent. It is, however, involved in an interaction with the aspartate dyad in pepsinogen (see below).

(i) Comparison of pepsinogen with pepsin

Least-squares fitting of the pepsinogen co-ordinates (kindly provided by Dr J. Remington) to the hexagonal pepsin structure resulted in 289 structurally equivalent residues (C^α separation less than 3.5 Å for segments of more than 3 adjacent residues), which overlap with a r.m.s. deviation of 0.92 Å. The enzyme and proenzyme structures are very similar, as can be seen in Figures 19 and 20. Most of the differences occur in the vicinity of the cleft which, in pepsinogen, is covered and filled by the propeptide (1P-44P) and the first 13 residues of pepsin.

The propeptide forms several salt-bridges to the enzyme, including one involving the aspartate dyad and Lys36P, which is wedged between two tyrosine rings (9 and 37P), the hydroxyl groups of which hydrogen-bond to the outer carboxyl oxygen atoms. Of these residues, Lys36P is almost invariant (18/19) and Tyr9 is highly conserved (17/24). The activation peptide blocks the active site in a different manner to peptide inhibitors. However, some of the residues of the cleft that hydrogen-bond to inhibitors, as defined by several inhibitor complexes of endothiapepsin (e.g. see Sali *et al.*, 1989), are involved in interactions with amino acids of the activation peptide that form the first loop of fully processed pepsin, e.g. both NH and O⁷ of Ser219 are within hydrogen-bonding distance of O11, O217 forms a hydrogen bond to the O⁸ of Thr12. Some of the differences between binding of an inhibitor and the activation peptide, which forms a distorted helix (3_{10}) in the active site, are illustrated by Figure 21.

Interestingly, the propeptide residues 2P to 9P structurally align with residues 2 to 9 of pepsin, since the first ten amino acid residues of the propeptide form β -strand a_N of pepsinogen. The pro-peptide is cleaved autolytically at position 16P-17P, unleashing active pepsin before cleavage of the 44P-1 peptide bond, which may be a bimolecular reaction. Whether residues 3 to 7 replace the first strand of pepsinogen before or after these cleavage events is not known.

Both pepsin and pepsinogen have a similar variation in the average U_{iso} value along the polypeptide chain, e.g. the regions 240 to 245, 278 to 282 and 294 to 298 have high thermal factors in both structures. Since activation of serine proteinases involves a disorder-order transition (Huber & Bode, 1978), a difference temperature factor plot for pepsin *versus* pepsinogen was calculated (Fig. 22). The most significant differences may be in the region 157 to 161, which was omitted from the pepsinogen co-ordinates, presumably due to disorder, yet is clearly defined in the pepsin structure. The conformation of this loop in pepsin is stabilized by hydrogen bonds

to Asn8 and Asp11 ($a_N b_N$ loop region), which are in totally different positions in the proenzyme. In pepsinogen, the residues nearest to the 157 to 161 loop are from the propeptide (7P-8P), which may not provide as much stability as the $a_N b_N$ loop residues do in pepsin. The a_N strand of pepsin represents another region that has lower U_{iso} values than in the proenzyme. These residues form a 3_{10} helix within the active site cleft of pepsinogen but, on activation, they become part of sheet 3 and are held tightly by hydrogen bonds.

Several stretches of polypeptide near the active site cleft appear to adopt very different conformations in the proenzyme due to interactions with the propeptide. These include the active site flap (75 to 80) and loop residues 291 to 298, which are forced away from the cleft in pepsinogen and residues 240 to 244 and 277 to 282, which are closer to the cleft region when the propeptide is present. All these differences are reflected in the C^α - C^α distance plot of pepsin least-squares fitted to pepsinogen, which is shown in Figure 19. It can be seen that the b_{N2} helix (110 to 114), which contributes to the S_1 and S_3 pockets, is also somewhat perturbed by the propeptide. In native pepsin, the Phe111 and Leu112 side-chains are positioned more deeply in the active site cleft.

From comparison of the difference distance and temperature factor plots (Figs 19 and 22), it is clear that the regions where pepsin and pepsinogen differ significantly correspond to the most flexible regions of both proteins. The inherent flexibility of these loops and the active site helix (b_{N2}) may facilitate unfolding of the propeptide on activation.

In addition to these local differences due to the propeptide, there appears to be a rigid body shift of part of the C-terminal lobe away from the cleft region (Fig. 20), similar to that observed for endothiapepsin upon co-crystallization with inhibitors (Sali *et al.*, 1989). This mobile rigid body of aspartic proteinases was defined as residues 190 to 302 using difference distance maps. A similar rigid group analysis was made for pepsin and pepsinogen. Least-squares (C^α superposition of hexagonal pepsin onto pepsinogen using only the first rigid body (residues 15 to 189 and 303 to 327) in the fitting program but applying the transformation to the whole pepsin molecule gives a r.m.s. deviation of 2.0 Å for the second rigid group. Fitting only the second rigid bodies of pepsin (hexagonal form) and pepsinogen lowers the r.m.s. deviation to 1.1 Å. This relative movement is best described by a 5.8° screw rotation and a 0.6 Å slide parallel to the rotation axis. Figures of 4.1° and 0.3 Å were obtained for endothiapepsin on complex formation with one inhibitor (Sali *et al.*, 1989). These effects could be due to the different crystal environments of the molecules in each comparison. However, this is unlikely, since the same analysis for the two crystal forms of pepsin (hexagonal and the more tightly packed monoclinic form) gave a smaller relative rigid body movement of 1.9° and 0.02 Å with insignificant improvement in the r.m.s. deviation for the second rigid group from 0.8 Å to 0.7 Å. Comparison

of pepsin with the fungal aspartic proteinases indicates that the orientation of the second rigid body is most different for endothiapepsin. The r.m.s. deviations from pepsin for this region of rhizopuspepsin and penicillopepsin after fitting with the first rigid group are almost equal (≈ 1.7 Å) but significantly less than the value of 2.7 Å obtained for endothiapepsin. This may have contributed to the failure of the rotation function with this enzyme.

Therefore, the aspartic proteinase fold has a tendency to flex in a concerted way when the active site cleft is occupied by a peptide, be it inhibitor, propeptide or presumably substrate. This was not observed earlier in inhibitor complexes (e.g. see Cooper *et al.*, 1987) where it occurs to a lesser extent with isomorphous inhibitor complexes. Recent kinetic experiments on the temperature dependence of k_{cat}/K_m (Hofmann *et al.*, 1988) and older data by Fruton (1976) on the change of k_{cat}/K_m as the S_3 pocket is filled also suggest a conformational change affecting the activation barrier. Whether this involves the enzyme or substrate or both cannot be determined from the available kinetic data.

We are indebted to the AFRC (U.K.) for project support and to the MRC for a studentship (to J.B.C.). We thank Drs N. Andreeva and A. Fedorov (Moscow Academy of Sciences) for provision of the monoclinic pepsin co-ordinate and reflection data. We are very grateful to Dr J. Remington (University of Oregon) for provision of the pepsinogen co-ordinates in advance of publication. We thank Drs C. Abad-Zapatero, J. Erickson and T. Rydel (Abbott Laboratories) for the kind donation of their monoclinic pepsin co-ordinates, which served as a useful comparison with our own structure. We are indebted to Dr A. Sali for his detection of the rigid group movement and its analysis in various aspartic proteinase structures, and to Dr J. A. Jenkins for growing the original crystals.

References

- Andreeva, N. S., Zdanov, A. S., Gustchina, A. E. & Fedorov, A. A. (1984). *J. Biol. Chem.* **259**, 11353-11365.
- Arndt, U. W. & Wonacott, A. J. (1977). *The Rotation Method in Crystallography*. North-Holland, Oxford.
- Baudys, M., Kostka, V., Gruner, K. & Pohl, J. (1982). *Collect. Czech. Chem. Commun.* **47**, 709-718.
- Bernal, J. D. & Crowfoot, D. (1934). *Nature (London)*, **133**, 794-795.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542.
- Blundell, T. L., Jenkins, J. A., Sewell, T., Pearl, L. H. & Pederson, V. (1985). In *Aspartic Proteinases and their Inhibitors* (Kostka, V., ed.), pp. 151-161. Walter de Gruyter, Berlin.
- Blundell, T. L., Jenkins, J. A., Sewell, T. L., Pearl, L. H., Cooper, J. B., Wood, S. P. & Veerapandian, B. (1990). *J. Mol. Biol.* **211**, 919-941.
- Cooper, J. B., Foundling, S. L., Hemmings, A. M., Blundell, T. L., Jones, D. M., Hallett, A. & Szelke, M. (1987). *Eur. J. Biochem.* **169**, 215-221.
- Crowther, R. A. (1972). In *The Molecular Replacement Method* (Rossmann, M. G., ed.), pp. 173-178. Gordon & Breach, New York.
- Crowther, R. A. & Blow, D. M. (1967). *Acta Crystallogr.* **23**, 544-548.
- Dodson, E. J. (1984). *CCP4 Program Suite Documentation*. Daresbury Laboratory, U.K.
- Dreyer, T., Halkier, B., Svendsen, I. & Ottesen, M. (1986). *Carlsberg Res. Commun.* **51**, 27-41.
- Foundling, S. L., Cooper, J., Watson, F. E., Cleasby, A., Pearl, L. H., Sibanda, B. L., Hemmings, A., Wood, S. P., Blundell, T. L., Valler, M. J., Norey, C. G., Kay, J., Boger, J., Dunn, B. M., Leckie, B. J., Jones, D. M., Atrash, B., Hallett, A. & Szelke, M. (1987). *Nature (London)*, **327**, 349-352.
- Fruton, J. S. (1976). *Advan. Enzymol. Relat. Areas Mol. Biol.* **44**, 1-36.
- Fruton, J. S. (1987). In *Hydrolytic Enzymes* (Neuberger, A. & Brocklehurst, K., ed.), pp. 1-37. Elsevier Science Publishers.
- Gilliland, G. L., Winbourne, E. L., Nachman, J. & Wlodawer, A. (1988). *18th Linderstrom-Lang Conference Proceedings*, p. 29.
- Hallett, A., Jones, D. M., Atrash, B., Szelke, M., Leckie, B. J., Beattie, S., Dunn, B. M., Valler, M. J., Rolph, C. E., Kay, J., Foundling, S. L., Wood, S. P., Pearl, L. H., Watson, F. E. & Blundell, T. L. (1985). In *Aspartic Proteinases and their Inhibitors* (Kostka, V., ed.), pp. 467-478. Walter de Gruyter, Berlin.
- Haneef, I., Moss, D. S., Stanford, M. J. & Borkakoti, N. (1985). *Acta Crystallogr. sect. A*, **41**, 426-433.
- Hartsuck, J. A. & Remington, S. J. (1988). *18th Linderstrom-Lang Conference Proceedings*, p. 28.
- Hodgkin, D. M. C. (1970). *Phil. Trans. Roy. Soc. ser. B*, **257**, 65.
- Hofmann, T., Allen, B., Bendiner, M., Blum, M. & Cunningham, A. (1988). *Biochemistry*, **27**, 1140-1146.
- Huang, W. Y. & Tang, J. (1972). *J. Biol. Chem.* **247**, 2704-2710.
- Huber, R. (1965). *Acta Crystallogr.* **19**, 353-356.
- Huber, R. & Bode, W. (1978). *Acc. Chem. Res.* **11**, 114-121.
- Imai, T., Miyazaki, H., Hirose, S., Hori, H., Hayashi, T., Kageyama, R., Ohkubo, H., Nakanishi, S. & Murakami, K. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 7405-7409.
- James, M. N. G. & Sielecki, A. R. (1983). *J. Mol. Biol.* **163**, 299-361.
- James, M. N. G. & Sielecki, A. R. (1986). *Nature (London)*, **319**, 33-38.
- Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). *J. Mol. Biol.* **125**, 357-386.
- Jenkins, J. A. (1979). Ph.D. thesis, University of Sussex.
- Jones, T. A. (1978). *J. Appl. Crystallogr.* **11**, 268-272.
- Kabsch, W. & Sander, M. (1983). *Biopolymers*, **22**, 2577-2637.
- Kitson, T. M. & Knowles, J. R. (1971). *FEBS Letters*, **16**, 337-338.
- Lin, X.-L., Wong, R. N. S. & Tang, J. (1989). *J. Biol. Chem.* **264**, 4482-4489.
- Northrop, J. H. (1930). *J. Gen. Physiol.* **13**, 739-766.
- Pearl, L. H. (1983). Ph.D. thesis, University of London.
- Pearl, L. H. & Blundell, T. L. (1984). *FEBS Letters*, **174**, 96-101.
- Perutz, M. (1949). *Research*, **2**, 52-61.
- Powers, J. C., Harley, A. D. & Myers, D. V. (1977). *Advan. Expt. Med. Biol.* **95**, 141-157.
- Read, R. J. (1986). *Acta Crystallogr. sect. A*, **42**, 140-149.

- Richmond, F. M. & Richards, T. (1978). *J. Mol. Biol.* **119**, 537-555.
- Sali, A., Veerapandian, B., Cooper, J. B., Foundling, S. I., Hoover, D. J. & Blundell, T. L. (1989). *EMBO J.* **8**, 2179-2188.
- Sepulveda, P., Marciniyszyn, J., Liu, D. & Tang, J. (1975). *J. Biol. Chem.* **250**, 5082-5088.
- Shewale, J. G., Takahashi, T. & Tang, J. (1985). In *Aspartic Proteinases and their Inhibitors* (Kostka, V., ed.), pp. 101-116. Walter de Gruyter, Berlin.
- Sibanda, B. L., Blundell, T. L., Hobart, P. M., Fogliano, M., Bindra, J. S., Dominy, B. W. & Chirgwin, J. M. (1984). *FEBS Letters*, **174**, 102-110.
- Sibanda, B. L., Blundell, T. L. & Thornton, J. M. (1989). *J. Mol. Biol.* **206**, 759-777.
- Sielecki, A. R., Hayakawa, K., Fujinaka, M., Murphy, M. E. P., Faser, M., Muir, A. K., Carilli, C. T., Lewicki, J. A., Baxter, J. D. & James, M. N. G. (1989). *Science*, **243**, 1346-1351.
- Singh, J., Thornton, J. M., Snarey, M. & Campbell, S. F. (1987). *FEBS Letters*, **224**, 161-171.
- Sly, W. S. & Fischer, H. D. (1982). *J. Cell. Biochem.* **18**, 67-85.
- Steigemann, W. (1974). Ph.D. thesis. Technical University of Munich, F.R.G.
- Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. & Davies, D. R. (1987a). *J. Mol. Biol.* **196**, 877-900.
- Suguna, K., Padlan, E. A., Smith, C. W., Carlson, W. D. & Davies, D. R. (1987b). *Proc. Nat. Acad. Sci., U.S.A.* **84**, 7009-7013.
- Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). *Protein Eng.* **1**, 377-384.
- Tang, J., Sepulveda, P., Marciniyszyn, J., Chen, K. C. S., Huang, W., Tao, N., Lui, D. & Lanier, J. P. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 3437-3439.
- Tarasova, N. I., Lavrenova, G. I. & Stepanov, V. M. (1980). *Biochem. J.* **187**, 345-352.
- Taylor, G. (1983). *J. Mol. Graphics*, **1**, 5-8.
- Ten Eyck, L. F. (1973). *Acta Crystallogr. sect. A*, **29**, 183-191.
- Tickle, I. J. (1985). In *Molecular Replacement, Proceedings of the Daresbury Study Weekend* (Machin, P., ed.), pp. 22-26. Daresbury Laboratory, DL/SCL/R23.
- Williams, R. C. & Rajagopalan, T. G. (1966). *J. Biol. Chem.* **241**, 4951-4954.
- Wilmot, C. M. & Thornton, J. M. (1988). *J. Mol. Biol.* **203**, 221-232.
- Yamauchi, T., Nagahama, M., Hori, H. & Murakami, K. (1989). *FEBS Letters*, **230**, 205-208.
- Zavada, L. L., Kozlov, L. V., Volkova, L. I. & Antonov, V. K. (1977). *Biorg. Khim.* **3**, 1671-1681.

Edited by R. Huber

Time-averaged Nuclear Overhauser Effect Distance Restraints Applied to Tendamistat

Andrew E. Torda, Ruud M. Scheek and Wilfred F. van Gunsteren

Laboratory of Physical Chemistry
University of Groningen, Nijenborgh 16
9747 AG Groningen, The Netherlands

(Received 12 December 1989; accepted 25 February 1990)

A penalty function is introduced into molecular dynamics simulations that improves on current methods for enforcing nuclear magnetic resonance-based distance restraints. Rather than treating nuclear Overhauser effects as static distance bounds, they are considered as quantities that must be satisfied on average over the course of a simulation trajectory.

The efficacy of the method is demonstrated on the previously determined structure of tendamistat. The molecular dynamics simulations show that the time-averaged constraints increase the mobility allowed to molecules, produce better agreement with distance bounds, improve searching properties and give a better estimate of the conformational space occupied by the molecule in solution.

1. Introduction

Historically, one of the strengths of nuclear magnetic resonance (n.m.r.) spectroscopy has been its ability to provide information about dynamic processes within molecules (Wüthrich, 1986). More recently, however, the use of n.m.r. as a tool for structure determination has brought about some change of emphasis. Instead of looking to see what motions are occurring in a molecule, current distance geometry and restrained molecular dynamics simulation techniques are all geared towards generating single structures that are consistent with all the experimental information. Given the known dynamic properties of molecules, we would suggest that this may not be a generally correct treatment.

In the extreme case of a small molecule hopping between two distinct conformations, the approach is demonstrably inappropriate. For example, the small peptide antamanide appears to exist in four conformations related by the flip of two peptide planes (Kessler *et al.*, 1988). Nuclear Overhauser effect (NOE) measurements, however, reflect both conformations simultaneously. In this case, attempting to satisfy all the experimental constraints simultaneously is equivalent to trying to generate some practically non-existent transition state. In larger molecules, it may not be possible to identify such distinct conformational states, but

measured NOEs are still average values reflecting the whole region of configurational space visited by molecules on the n.m.r. time scale.

The problem with current methods is that they use a simple penalty function that forces each single structure to agree with all the experimental data. A more physically realistic penalty function should allow for some averaging of the NOE through time. This inclusion of time leads naturally to the use of such a term in molecular dynamics (MD) simulations rather than a distance geometry method.

Recently, we proposed such a term and demonstrated its application on a simple, artificial model system (Torda *et al.*, 1989). Each particle was allotted a memory of its history with respect to distance constraint violations and was only required to satisfy the distance constraints over the course of a trajectory. Here, we apply the procedure to the realistic case of refining a large polypeptide using published n.m.r. data. These results are then contrasted with those obtained by a more conventional refinement.

For assessing the use of these time-averaged distance constraints, we used the published distance geometry structures and data from tendamistat (Kline *et al.*, 1988). This is a polypeptide of 74 residues, M_r 8000, and has been referred to as the most precisely determined solution structure of a protein (Wüthrich, 1989). This was a particularly interesting test case as the large volume of experimental data resulted in a family of very tightly clustered distance geometry structures with no evidence of conformational averaging.

† Abbreviations used: n.m.r., nuclear magnetic resonance; NOE, nuclear Overhauser effect; MD, molecular dynamics; r.m.s., root-mean-square.