The EMBO Journal vol.12 no.4 pp.1293–1302, 1993

Two crystal structures for cathepsin D: the lysosomal targeting signal and active site

Peter Metcalf[1] and Martin Fusek[2]

European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D6900, Heidelberg, Germany

[2]Present address: Oklahoma Medical Research Foundation, 825 NE 13th St., Oklahoma City, OK 73104, USA

[1]Corresponding author

Communicated by J.Tooze

Two crystal structures are described for the lysosomal aspartic protease cathepsin D (EC 3.4.23.5). The molecular replacement method was used with X-ray diffraction data to 3 Å resolution to produce structures for human spleen cathepsin D and for bovine liver cathepsin D complexed with the 6-peptide inhibitor pepstatin A. The lysosomal targeting region of cathepsin D defined by previous expression studies [Barnaski *et al.* (1990) *Cell*, 63, 281–219] is located in well defined electron density on the surface of the molecules. This region includes the putative binding site of the *cis*-Golgi phosphotransferase which is responsible for the initial sorting step for soluble proteins destined for lysosomes by phosphorylating the carbohydrates on these molecules. Carbohydrate density is visible at both expected positions on the cathepsin D molecules and, at the best defined position, four sugar residues extend towards the lysosomal targeting region. The active site of the protease and the active site cleft substrate binding subsites are described using the pepstatin inhibited structure. The model geometry for human cathepsin D has rms deviations from ideal of bonds and angles of 0.013 Å and 3.2° respectively. For bovine cathepsin D the corresponding figures are 0.014 Å and 3.3°. The crystallographic residuals (R factors) are 16.1% and 15.8% for the human and inhibited bovine cathepsin D models respectively. The free R factors, calculated with 10% of the data reserved for testing the models and not used for refinement, are 25.1% and 24.1% respectively.

Key words: aspartic protease/cathepsin D/inhibitor complex/lysosomal targeting/three-dimensional structure

Introduction

One of the best understood intracellular protein sorting pathways involves the transport of soluble lysosomal enzymes from their site of synthesis in the endoplasmic reticulum via the Golgi and endosomes to lysosomes (reviewed by Kornfeld and Mellman, 1989). The lysosomal terminal compartment is well defined biochemically and morphologically (von Figura, 1991). A sorting tag which is attached in the early secretory pathway has been identified which distinguishes molecules destined for lysosomes. Phosphate attached to the 6 position of mannose residues on carbohydrate groups of soluble lysosomal enzymes is sufficient to target these molecules for lysosomal transport (Hasilik and Neufeld, 1980). Two mannose-6-phosphate receptor molecules involved in the pathway have been cloned and characterized (reviewed by Kornfeld, 1992). However, the primary targeting signal for lysosomal enzymes remains unclear and the features of soluble lysosomal enzymes required for mannose-6-phosphate tagging and consequent transport to lysosomes are still poorly understood.

Cell lines derived from patients with the inherited lysosomal storage disease mucolipidosis II secrete non-phosphorylated lysosomal enzymes and are able to endocytose phosphorylated but not dephosphorylated exogenous lysosomal enzymes. Endocytosis could be specifically inhibited using phosphorylated sugars (Kaplan *et al.*, 1977; Ulrich *et al.*, 1978). These properties led to the discovery of the mannose-6-phosphate mediated lysosomal sorting pathway and to results indicating that the initial selection step for the pathway is most probably carried out by a single phosphotransferase located in the *cis*-Golgi. This enzyme distinguishes soluble lysosomal enzymes from the numerous other glycoproteins passing through this compartment (reviewed by Kornfeld and Mellman, 1989). The purification to homogeneity of this phosphotransferase has still to be achieved and is essential for a better understanding of this key recognition step of the pathway. Partially purified preparations phosphorylate only lysosomal enzymes and are selectively inhibited by native, but not denatured deglycosylated lysosomal enzymes (Ketcham and Kornfeld, 1991a,b; Lang *et al.*, 1984). This structure-dependent inhibition, and the lack of any evident shared primary sequence homology or pattern in the number or location of carbohydrate groups or in the phosphorylated mannose residues found on them in lysosomal proteins, suggested that the lysosomal sorting signal is a three-dimensional protein structural motif shared by soluble lysosomal enzymes (Lang *et al.*, 1984).

Confirmation of this hypothesis would follow from the identification of a shared motif on lysosomal proteins with known three-dimensional structure and the engineering of a non-lysosomal protein that would express the motif and be retargeted by lysosomes. To date only one lysosomal enzyme, the cysteine protease cathepsin B, has a known X-ray structure although the coordinates are not yet available from the Brookhaven database (Musil *et al.*, 1991). The lysosomal aspartic protease cathepsin D (EC 3.4.23.5) has a unique advantage for the identification and verification of the lysosomal targeting signal: three similar but secreted mammalian aspartic proteases have known X-ray structures (renin-pepsin and chymosin: Figure 1). Because of the high sequence homology between these molecules and cathepsin D (~45% identity), chimeric proteins containing part cathepsin D and part secreted aspartic protease can be expected to fold correctly and can be used in expression studies to identify which parts of cathepsin D are required for mannose-6-phosphate tagging and lysosomal transport.

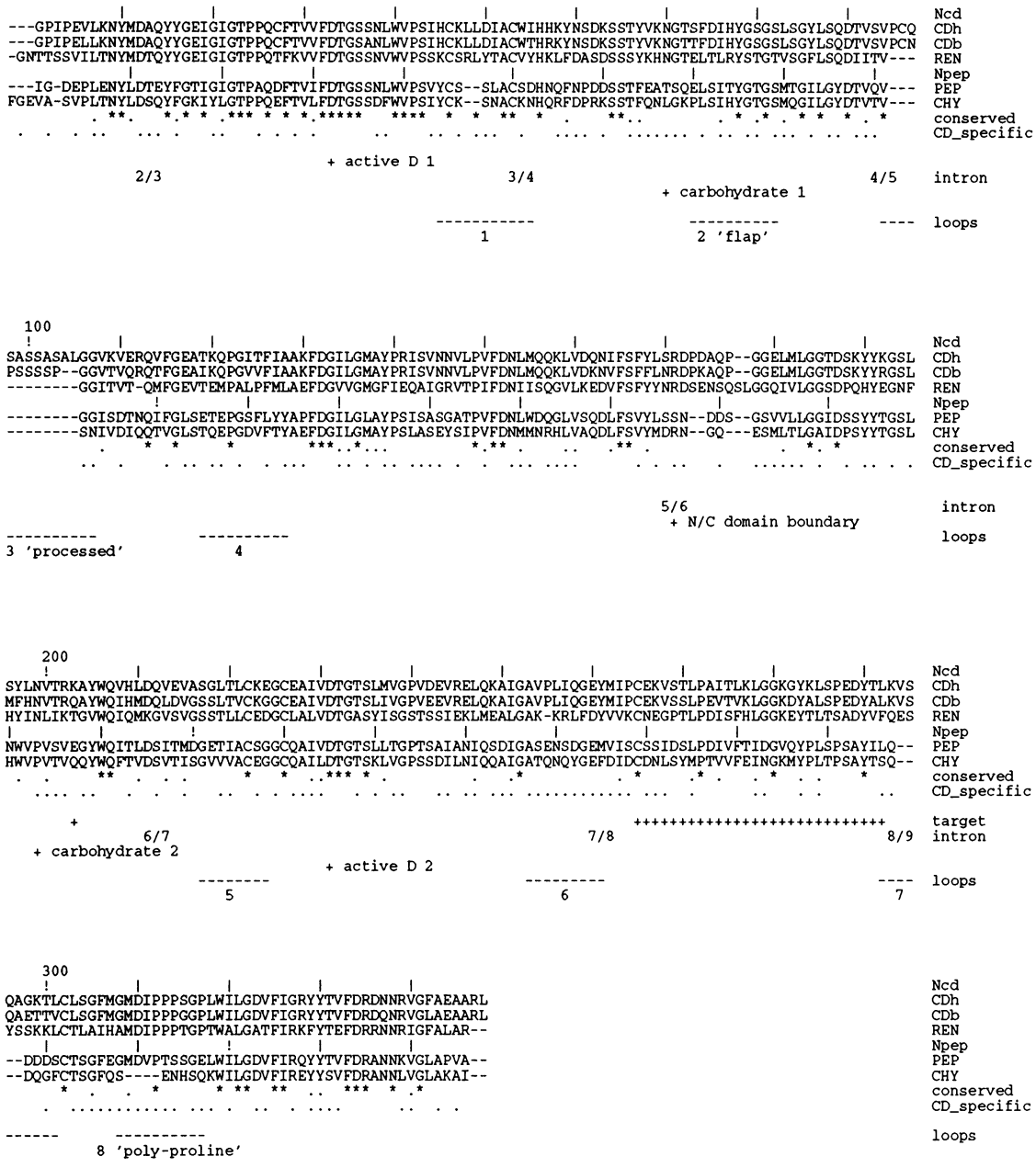

Fig. 1. Sequence alignment for the mammalian aspartic proteases with known three-dimensional structures: cathepsin D, renin, pepsin and chymosin. Sequences of cathepsin D (five species), renin (four), pepsin (four) and chymosin (two) were aligned using the program CLUSTALV (Higgins *et al.*, 1992). The structural alignments were checked with the program O (Jones *et al.*, 1991). Only five sequences from the 15 sequence alignment are shown here: human and bovine cathepsin D, human renin, porcine pepsin and bovine chymosin. **Key**: Ncd, human cathepsin D numbering; CDh, human cathepsin D; CDb, bovine cathepsin D; REN, human renin; Npep, pepsin numbering, as used for the structure of porcine pepsin; PEP, porcine pepsin; CHY, bovine chymosin; conserved, conserved residues are shown by a (.) and identical ones by (*); CD_specific, residues identical in cathepsin D sequences and different in either one of the renin, chymosin or pepsin sequences; intron, human cathepsin D exon boundaries (Redecker *et al.*, 1991); target, cathepsin D residues essential for phosphorylation assay, the minimal lysosomal targeting signal (Baranski *et al.*, 1991); loops, variable loops 1–8. **Sequence and crystal structure sources**: Cathepsin D: Swissprot entries CATD_HUMAN, PIG, MOUSE and RAT; bovine cathepsin D: Chirgwin,J., unpublished cDNA sequence; residues K158–G224 were reconfirmed in this work by PCR cloning and sequencing from a bovine cDNA library (Labeit and Metcalf, unpublished results). Renin: Swissprot entries RENI_HUMAN, MOUSE, RAT, RENS_MOUSE, structure from Dhanaraj *et al.* (1992). Pepsin: Swissprot entries PEPA_HUMAN, CHICK, MACFU; structure from Abad-Zapatero *et al.* (1990) (Brookhaven database entry 3PEP). Chymosin: Swissprot entries CHYM_BOVIN, SHEEP; structure from Gilliland *et al.* (1990) (Brookhaven database entry 1CMS).

A systematic search using chimeric mRNA constructs of human procathepsin D† and glycopepsinogen (pepsinogen with carbohydrates at the two corresponding cathepsin D carbohydrate positions) expressed in *Xenopus* oocytes has led to the identification of regions of the cathepsin D primary sequence required for phosphorylation (Baranski *et al.*, 1990, 1991, 1992; Cantor *et al.*, 1992). Unfortunately the

†The cathepsin D molecule tagged for lysosomal transport by the *cis*-Golgi phosphotransferase is procathepsin D (53 kDa as judged by SDS–PAGE) [reviewed by Hasilik (1992)]. The 46 residue N-terminal prosequence is removed at low pH in the endosome in a manner presumably similar to the low pH activation of pepsinogen to pepsin. Further partial processing occurs in the lysosome and cathepsin D purified from tissue contains 48 kDa intact molecules and clipped molecules with 33 and 15 kDa chains.

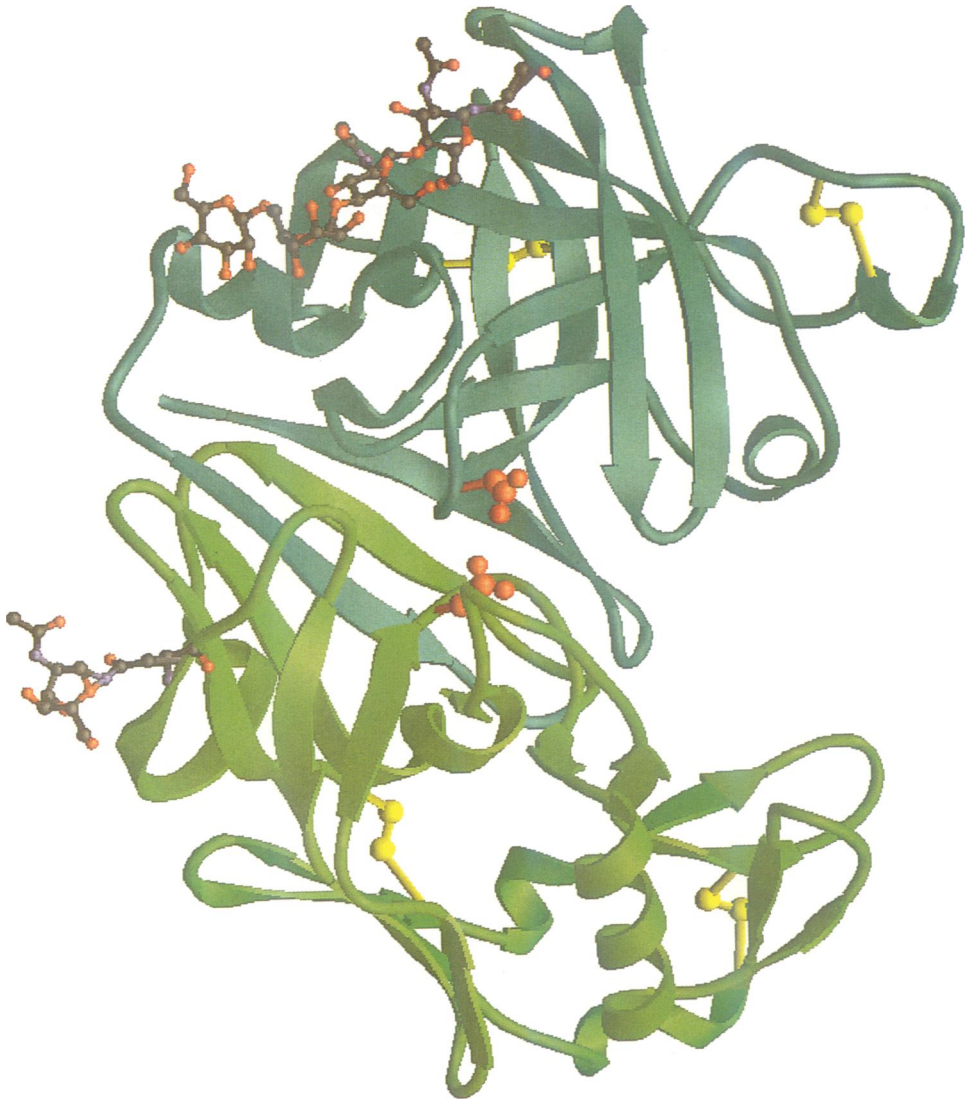

Fig. 2. Ribbon representation of cathepsin D. The N-terminal domain is shown coloured blue at the top of the figure and the C-terminal domain is in green at the bottom. The active site aspartate side chains are the red spheres at the centre of the molecule and the disulfide bonds are shown in yellow. The carbohydrate residues are depicted with smaller atom spheres. The figure was made using the program RIBBONS 2.0 (Carson, 1987) using the bovine cathepsin D coordinates.

homology between aspartic proteases does not extend to the surface and loop regions expected to comprise the phosphotransferase binding site, so little information is available for modelling the structure of the lysosomal targeting signal.

Cathepsin D is an abundant intracellular endoproteinase located in endosomes and lysosomes and presumably involved in the degradation of endocytosed and endogenous proteins in these organelles. Besides its relevance to the lysosomal sorting question, cathepsin D is also of interest because it is probably involved in antigen processing in endosomes (Factorovich and Puri, 1988), and as the major and ubiquitously expressed aspartic protease, it represents a possible alternative target for drugs designed to specifically inhibit renin and the HIV protease. Aspartic protease inhibitors specific for cathepsin D may also prove to be medically useful since the enzyme has been implicated in a number of diseases (Gopalan *et al.*, 1987; Kenessey *et al.*, 1989) and because secretion of the proenzyme has been associated with metastasis in breast cancer (Rochefort, 1992; Tandon *et al.*, 1990). Here we describe the structures of human spleen cathepsin D and bovine liver cathepsin D with the 6-peptide inhibitor pepstatin A bound in the active site cleft.

Results

Overall fold, variable loops

The structure of cathepsin D is similar to that of other aspartic proteases, as expected from their sequence homology (reviewed by Davies, 1990). Two related 170 residue, mostly β sheet domains lie on either side of the deep 30 Å long active site cleft. Each domain contributes an active site aspartate at the centre of the cleft and each contains a single carbohydrate group and two disulfide bonds (Figure 2). Structural comparisons of aspartic proteases showed that a region comprising two-thirds of the C-terminal domain moves as a separate rigid body rotating up to 18° from the rest of the molecule (Abad-Zapatero *et al.*, 1990; Sielecki *et al.*, 1990; Sali *et al.*, 1992). When the human and bovine cathepsin D structures are aligned, taking account this rigid body movement, four loops in each domain have significantly

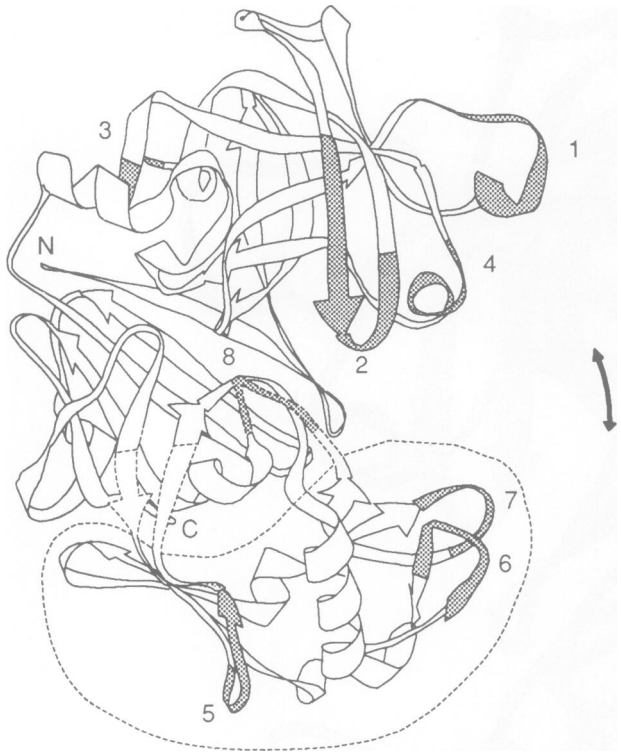

Fig. 3. Diagram showing the location of the loops that differ most between the human and bovine cathepsin D structures and also between the structures of other aspartic proteases. The dotted boundary shows the location of the C-terminal structural domain. The numbered loops are also shown in the sequence alignment (Figure 1).

different conformations between the two molecules (loops 1–8; Figures 1, 3 and 4). All except loop 1 vary to a similar extent in other aspartic protease structures and all the loops have higher than average crystallographic temperature factors and are probably mobile or partially disordered in the crystals (Figure 4). Cathepsin D contains a unique extended loop (loop 3: the 'processed loop') located near the N-terminus which, in addition to the two C-terminal residues, is partially cleaved in purified protein preparations (Yonezawa *et al.*, 1988b) and in the crystals used in this work (Bieber *et al.*, 1992; Fusek *et al.*, 1992). Two loops close over the active site cleft are probably involved in substrate binding. One of these, the 'polyproline loop' (loop 8) contains three consecutive prolines and occurs only in renin and cathepsin D. The other, the 'flap' (loop 2), is mobile in the human structure but is stabilized in the bovine structure where the cleft is occupied by pepstatin. The whole C-terminal structural subdomain of bovine cathepsin D has higher temperature factors than the corresponding region in human cathepsin D and may move as a rigid body in the crystal (Figure 4C). Electron density at the end of loop 6 in bovine cathepsin D is very weak although the same loop in human cathepsin D is interpretable. Loop 6 is replaced by a 104 residue insertion in an aspartic protease from barley grain related to cathepsin D (~52% identity) (Runeberg-Roos *et al.*, 1991). The loops with different configurations in the two cathepsin D structures represent individual conformations of flexible parts of the molecule which probably also occur *in vivo*. The stable existence of a molecule with flexible loops in the proteolytic lysosomal environment demonstrates that protein flexibility and susceptibility to proteolysis are not necessarily correlated.

Both human and bovine cathepsin D crystallized as pseudo-dimers with two molecules related by a 2-fold non-crystallographic rotation axis in the crystal asymmetric unit. The interface between the molecules contains mostly residues from loops 1 and 4 contacting loops 6 and 7 of the symmetry related molecule. The transformations relating the human and bovine molecules differ and consequently the details of the two dimer interfaces are different (Figure 5). A comparison of the dimer structures shows that the loops shift to accommodate the neighbouring molecules. In particular, loop 1 of bovine cathepsin D, which contains eight residues closed by a disulfide bond, is rotated 30° about an axis near the disulfide bond from its position in human cathepsin D, which is similar to the corresponding loop in renin. Because the dimer configurations are not conserved and because analytical sedimentation and gel filtration experiments provided no evidence for dimerization of bovine cathepsin D at concentrations up to 5 mg/ml (M.Fusek, unpublished data), we conclude that the two dimer structures seen in the crystals do not reflect the oligomeric state or function of cathepsin D within lysosomes. A pseudo-dimer and pseudo-tetramer of renin have also been described recently (Dhanaraj *et al.*, 1992). Two other aspartic proteases related to cathepsin D have been reported to be dimeric: cathepsin E has a 10 residue N-terminal extension containing an extra cysteine thought to be involved in dimer formation (Yonezawa *et al.*, 1988a). The mosquito lysosomal protease mLAP has the highest sequence homology of any aspartic protease to cathepsin D (58% identity). However, it has no such N-terminal extension and lacks the whole processed loop, the two cysteines corresponding to C27 and C96 of cathepsin D, the C-terminal processed residues and the C-terminal carbohydrate (Cho and Raikhel, 1992).

The carbohydrates and lysosomal targeting signal

Interpretable carbohydrate density extends from the molecular surface at both carbohydrate positions on each of the two molecules of human and bovine cathepsin D. The longest clearly interpretable chain extends two residues from the N-terminal position on bovine cathepsin D and weak density for two further residues can be seen extending further from the molecule (Figure 6). Phosphate groups attached to lysosomal enzyme carbohydrates occur at mannose positions at least four sugar residues out from the protein surface (Neufeld and Ashwell, 1980), so none of the attachment sites are visible in these structures. The actual phosphates are unlikely to be present on the crystallized protein or on proteins within lysosomes because they are rapidly removed by lysosomal phosphatases (Ludwig *et al.*, 1991).

The simplest model for the initial selection step of lysosomal transport involves a single *cis*-Golgi phosphotransferase binding to a well defined protein region on the surface of glycosylated molecules destined for lysosomes. Several phosphorylation events on different mannose residues might occur as the flexible carbohydrate chains of these molecules extend towards the active site of the bound phosphotransferase. In expression studies (Baranski *et al.*, 1990) the high level of phosphorylation of a chimeric protein containing an N-terminal glycopepsinogen domain and a C-terminal cathepsin D domain clearly suggested a C-

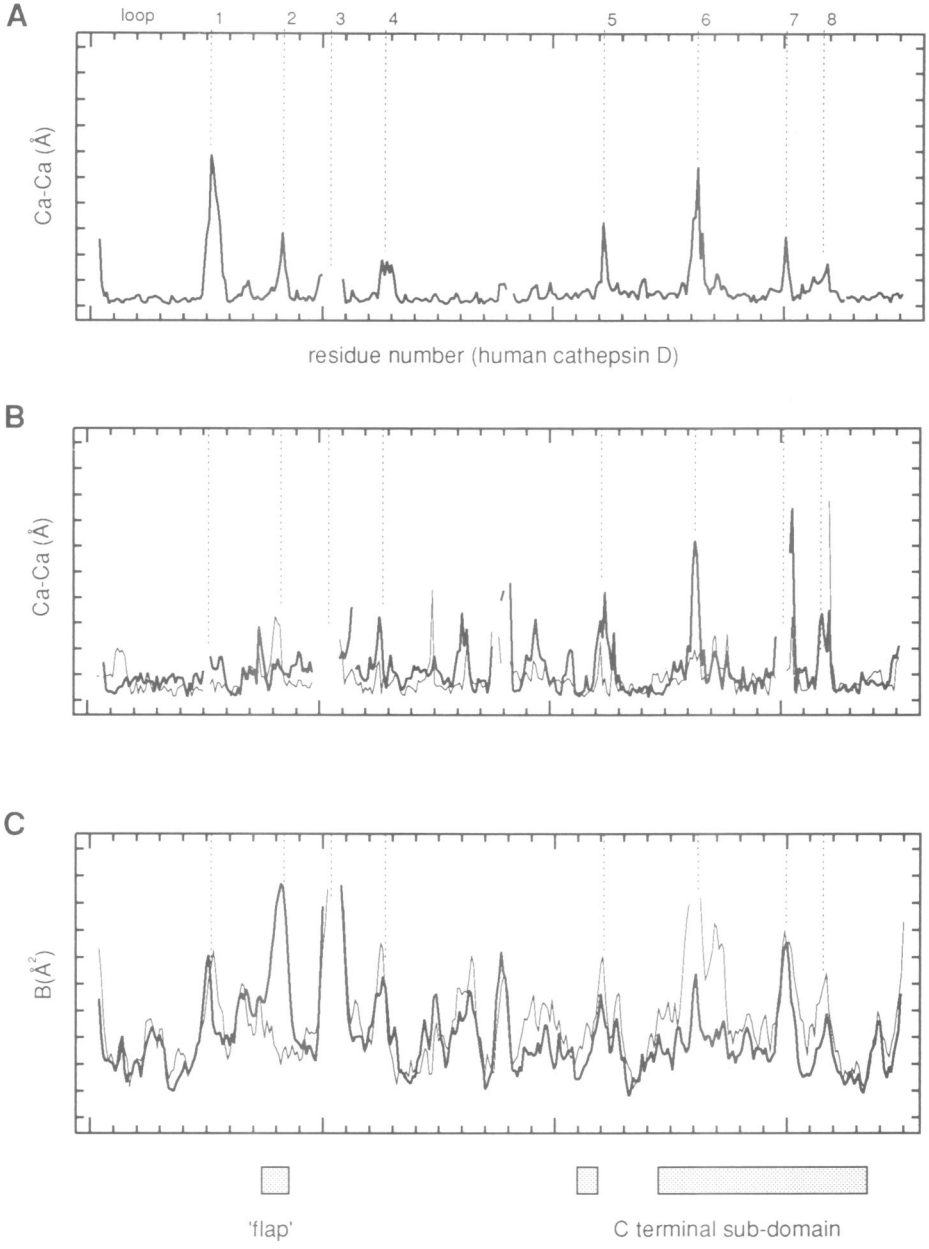

Fig. 4. (**A**) Comparison of human and bovine cathepsin D structures. The N- and C-terminal structural domains of human and bovine cathepsin D were separately aligned by matching the positions of corresponding Cα atoms in the domains using the program O (Jones *et al.*, 1991). The plot shows distances between Cα atoms using the appropriate alignment for residues in the two structural domains. Vertical axis: Cα−Cα distances between corresponding residues of the aligned molecules. Horizontal axis: human cathepsin D residue numbers. The dotted vertical lines show the positions of the variable loops shown in Figures 1 and 3. (**B**) Comparison of the structures of mammalian aspartic proteases: renin, chymosin and pepsin. The structures of mouse submandibular renin (Dhanaraj *et al.*, 1992) and chymosin (Gilliland *et al.*, 1990) were aligned with that of pepsin (Abad-Zapatero *et al.*, 1990) using the same methods as in panel A. Vertical axis: Cα−Cα distances between corresponding residues of pepsin and renin (thick lines) and pepsin and chymosin (thin lines). Horizontal axis: human cathepsin D numbers (the sequence alignment used is that in Figure 1). (**C**) Cathepsin D temperature factors. Average main chain temperature factors (B factors) versus residue number for human cathepsin D (thick lines) and pepstatin inhibited bovine cathepsin D (thin lines). The bars below the axis show the position of the flap loop and the C-terminal structural subdomain.

terminal location for the binding site of the phosphotransferase on cathepsin D. The minimal C-terminal targeting region containing the region C265−L292 and K203 of human cathepsin D when expressed in a glycopepsinogen background produced lower, but still significant phosphorylation. Results with other constructs, often also with comparatively low phosphorylation levels, suggested a less well defined binding site of multiple binding sites, and are sometimes difficult to interpret. Lysine 203, in particular, was required for phosphorylation of the minimal hybrid construct (Baranski *et al.*, 1991), but was not essential for cathepsin D phosphorylation (Baranski *et al.*, 1992) and is furthermore not present in bovine cathepsin D. However, it is clear that the major determinant for cathepsin D phosphorylation is in the C-terminal and that the region C265−L292 is probably an essential part of the phosphotransferase binding site.

The C265−L292 region, coded by most of exon 8,

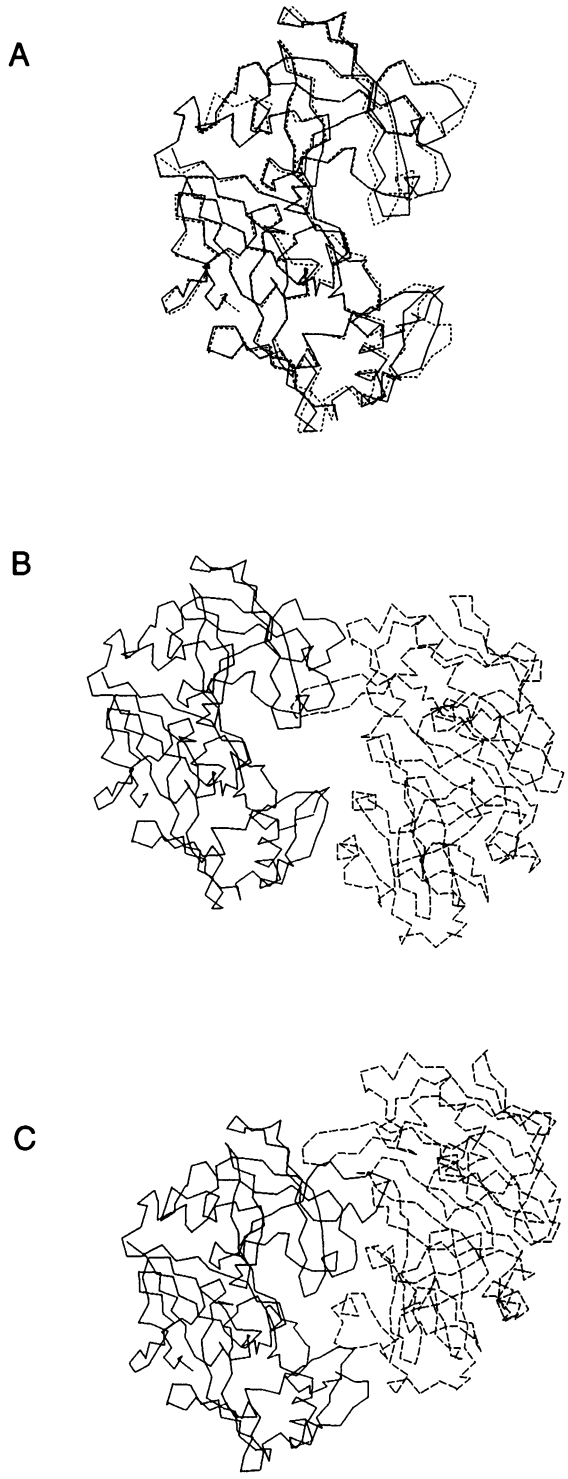

Fig. 5. (**A**) Superimposed Cα plots for human (solid) and bovine cathepsin D (dotted lines). (**B**) Human cathepsin D pseudo dimer. The molecule on the left has the same orientations as that shown in panel A. (**C**) Bovine cathepsin D pseudo dimer with the molecule on the left oriented as in panel A.

comprises about a quarter of the C-terminal hinged structural subdomain and is located near to the C-terminus (Figure 7). The mobile subdomain of aspartic proteases has previously been proposed to play a role in substrate binding or in proenzyme activation (Abad-Zapatero *et al.*, 1990) and the location of the lysosomal targeting signal of cathepsin D on the subdomain suggests that its movement may aid the bound phosphotransferase to reach its targets. The closeness of the lysosomal targeting signal to the C-terminus also raises the possibility that C-terminal processing (i.e. the removal of the last two residues) may also be involved in sorting although expressed cathepsin D constructs with a C-terminal extension appeared to be normally phosphorylated (Schorey and Chirgwin, 1991).

The active site cleft and active site

Aspartic proteases are thought to function by binding the target extended polypeptide chain in a distorted configuration and then attacking the strained peptide bond adjacent to the enzyme active site (Pearl, 1985). The selectivity of the enzymes is poorly understood and may involve substrate interactions far from the active site which induce a distorted conformation in the substrate as it binds to the protease. The active site cleft width depends on the inter-domain movement and the binding of substrate into the cleft is presumably directly affected by this domain movement and by the conformations of loops at the entrance to the cleft. The average width of the cleft is reduced ~0.5 Å between human cathepsin D and bovine cathepsin D where the cleft is occupied by pepstatin. The conformation of the polyproline loop in the two molecules is similar to that in renin. The mobility of loop 2 (the 'flap') is substantially reduced on binding inhibitor as indicated by residue temperature factors (Figure 4C).

Residues in the vicinity of the active site are well defined in the cathepsin D structures and the active site is similar to those of other aspartic proteases, as expected from the strong sequence homology in this part of the molecule. Hydrogen bonds to the pepstatin inhibitor are shown in Figure 8 and the residues which create binding pockets for the substrate are listed in Table I. Higher resolution native aspartic protease structures have a catalytic water between the active aspartates which was not evident in the human cathepsin D map. The pepstatin P1 statine hydroxyl occupies this position in the bovine complex. The inhibitor is almost completely buried apart from the C-terminal P3′ residue, which is exposed to solvent.

Kinetic studies with a range of substrates (Offermann *et al.*, 1983; Dunn *et al.*, 1986; Imoto *et al.*, 1987; van Noort and ver der Drift, 1989; Jupp *et al.*, 1990) have shown that cathepsin D, like other aspartic proteases, has a general preference for hydrophobic residues in P1 and P1′ subsites and cleaves substrates longer than five residues. The pH optimum is pH 3.5 (Bond and Butler, 1987), between that of pepsin (pH 1.9; Tang, 1970) and renin (approximately pH 6.5; Inagami *et al.*, 1985). Unlike pepsin, cathepsin D is stable at pH 7, although inactive (Baudyš,M., personal communication). The active site cleft shares specific features with renin, pepsin and chymosin: Cathepsin D, like pepsin and chymosin, has a tyrosine (Y205) in the S2′ pocket replacing the smaller residue valine of renin. This is the only aromatic residue in the base of the cleft. However, only renin and cathepsin D have the nearby polyproline loop (P312–P317) located near the P3′ end of the inhibitor. A unique feature of the cathepsin D active site cleft is a second methionine (M307) forming part of the S2 subsite. In other aspartic proteases the corresponding residue is charged or

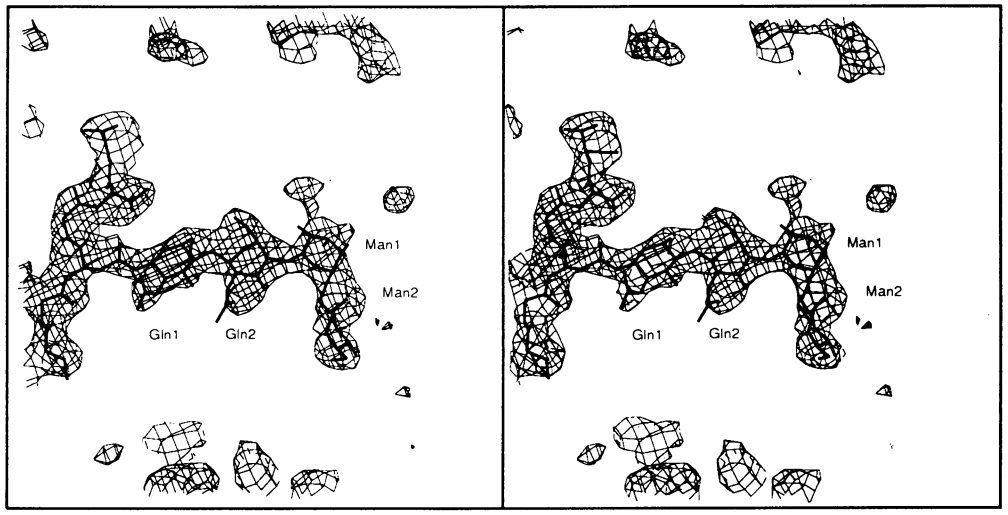

Fig. 6. The N-terminal carbohydrate of bovine cathepsin D. The stereo diagram shows unaveraged $2F_o-F_c$ map density contoured at 0.9σ. The inner two residues labelled Gln1 and Gln2 have high average crystallographic temperature factors (60 and 77) and the outer two mannose residues are poorly defined, probably because of flexibility or chemical heterogeneity, and are not included in the atomic model.

polar. Residue 125 in the S3 subsite is the only cathepsin D active cleft residue that varies in different species (threonine in human and pig; valine in cow, mouse and rat). There are 30 lysine and arginine residues on the surface of cathepsin D, 20 on renin and only three on pepsin. The role of these charged residues in the stability, pH optimum and substrate selectivity of the enzymes is unclear although several are in the vicinity of the active site cleft and are likely to influence proteolytic activity. A better understanding of these features, and in particular the role of the polyproline loop, will require higher resolution inhibitor complex structures and further kinetic studies with mutant aspartic proteases.

Proteolytic cleavage: the processed loop and the C-terminus

Unlike other aspartic proteinases, cathepsin D is processed to a two chain protein by the excision of several residues near residue 100 (Shewale and Tang, 1984). The disulfide bond between residues C27 and C96 predicted by three-dimensional modelling (Yonezawa *et al.*, 1988b) was confirmed in these crystal structures and joins one end of the processed loop to the main body of the molecule. The number and sequence of residues in the processed loop varies in different species; this fact has been used to argue against a role for the loop in lysosomal sorting (Yonezawa *et al.*, 1988b). However, the proximity of the loop to the expected position of the prosequence N-terminus was used to suggest that the loop might stabilize the proenzyme before it reaches a low pH environment (Yonezawa *et al.*, 1988b). The sequences of processed loops from five species are listed in Table II together with the loop residues absent in the bovine and human cathepsin D electron density maps. The human cathepsin D crystals contained no unprocessed protein as judged by SDS electrophoresis (Fusek *et al.*, 1992) and N-terminal sequencing of the heavy chain from SDS gels produced evidence only for a heterogenous N-terminus, in agreement with previous results (Horst and Hasilik, 1991). The bovine crystals contained ~50% intact unprocessed protein (Bieber *et al.*, 1992), but there is little evidence for the loop in the electron density map, presumably because it is mobile or disordered.

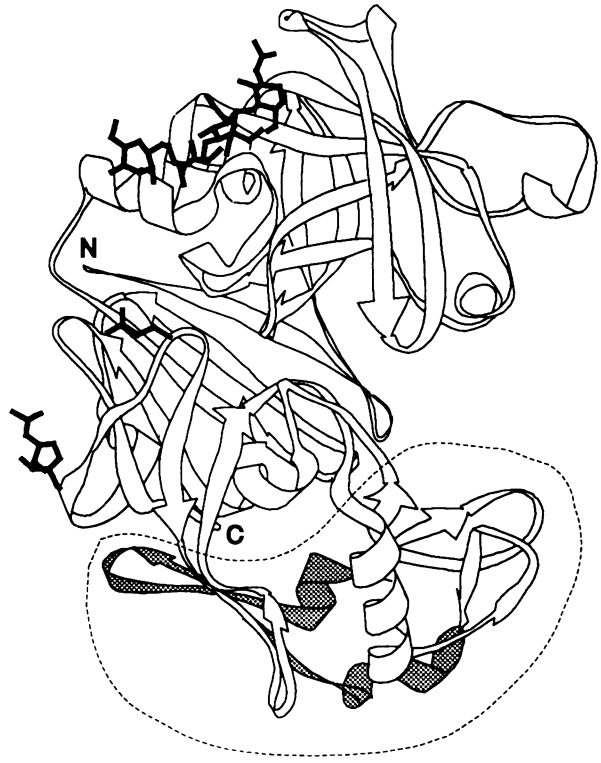

Fig. 7. The minimal lysosomal targeting region of cathepsin D. When residues corresponding to C265–L292 and K203 of non-phosphorylated glycopepsinogen were replaced with human cathepsin D residues, the carbohydrates of the resultant chimeric protein were significantly phosphorylated using an oocyte expression system (Baranski *et al.*, 1991). The figure shows a ribbon representation of bovine cathepsin D. Loop C265–L292 (shaded) is part of the C-terminal structural subdomain (dotted outline). In bovine cathepsin D residue 203 is glutamine. Side chains for Q203 and the carbohydrates are shown.

The sequence of porcine cathepsin D obtained by protein sequencing (Shewale and Tang, 1984) does not contain the

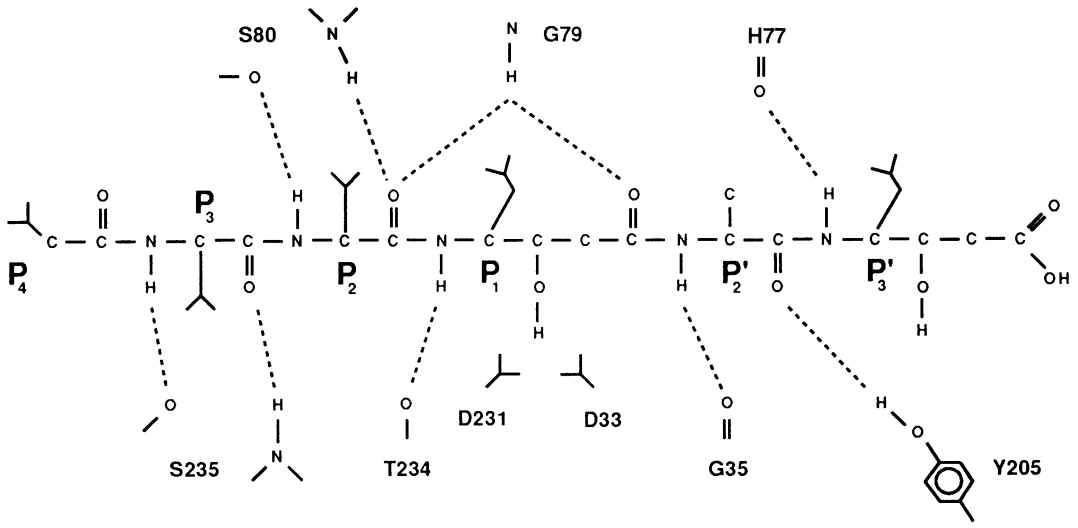

Fig. 8. Schematic diagram showing hydrogen bonds between bovine cathepsin D and the 6-peptide aspartic protease inhibitor pepstatin A. A 3.5 Å bond length cut-off was used.

two C-terminal residues present in cathepsin D cDNA sequences but absent in other aspartic proteases (Figure 1). In the electron density maps for bovine and human cathepsin D there is no interpretable density for the C-terminal conserved leucine L349 and the penultimate arginine R347 can be seen only in the bovine cathepsin D map. The chemical structure of the C-termini of these molecules remains to be determined and may well prove interesting since evidence for a non-mannose-6-phosphate dependent pathway (Kornfeld, 1986) and for membrane bound procathepsin D (Diment *et al.*, 1988; McIntrye and Erickson, 1991; Rijnbout *et al.*, 1991; Casciola-Rosen *et al.*, 1992) has focused attention on parts of the molecule that may be attached to a hydrophobic moiety and perhaps be part of a novel sorting mechanism.

Discussion

An important first step in understanding the mechanism of an intracellular protein transport pathway is to identify which features of the sorted molecules are necessary for the initial selection step of the pathway. The lack of primary sequence homology of the sorted molecules in most intracellular transport pathways means that detailed chemical and structural studies of both the purified targeted molecules and the molecules with which they interact are necessary to identify the targeting signal shared by the sorted molecules. In this work we have exploited the fact that cathepsin D, by virtue of its strong pH dependent binding to a pepstatin affinity column, is comparatively easy to purify in crystallizable quantities. We have used two crystal forms, together with previous information from chimera expression studies, to learn the three dimensional structure of the region on cathepsin D required for its lysosomal sorting. Other lysosomal enzymes have been the subject of crystallographic studies (e.g. Musil *et al.*, 1991; Church *et al.*, 1992) and in future a comparison of their structures, together with the continued development of better expression systems to assay sorting mutants quantitatively, will allow the identification of lysosomal targeting signals on more lysosomal molecules. This information, together with the purification and characterization of the phosphotransferase which selectively phosphorylates soluble lysosomal molecules, should eventually lead to a detailed understanding of the initial selection step of the lysosomal transport pathway.

Table I. The amino acid residues forming binding subsites of bovine cathepsin D with pepstatin

Subsite	Amino acid residues								
S4	A13	S235	L236						
S3	Q14	S80	V125	F126	F131	G233	T234	S235	
S2	Y78	G79	S80	G233	T234	V239	M307	M309	I320
S1	V31	D33	G35	H77	Y78	F126	F131	I134	D231
S1′	G35	S36	H77	Y78	T234				
S2′	H77	Y78	G79	Y205	I311				

Cathepsin D residues within 4.5 Å of pepstatin are listed. Residues at the position of M307 (underlined) are charged or polar in other aspartic proteases.

Table II. Processed loop sequences for cathepsin D molecules

human	v	P	C	Q	S	A	S	S	A	S	A	L	G	g
bovine	v	P	C	N	P	S	S	S	S			P	G	g
porcine	v	P	C	N	S	A	L	S	G	V			G	g
rat	v	P	C	K	S	D						L	G	g
mouse	v	P	C	K	S	D	Q	S	K	A	R			g

Underlined human and bovine residues are not visible in the electron density maps either because they have been proteolytically cleaved or because they are mobile or disordered. Residues in bold are missing from purified protein preparations (Yonezawa *et al.*, 1988b). Sequence sources are as in Figure 1.

Materials and methods

The purification, crystallization and data collection procedures used in this work have been previously described for both human cathepsin D (CDh; Fusek *et al.*, 1992) and pepstatin A bovine cathepsin D (CDbp; Bieber *et al.*, 1992). Orthorhombic crystals of CDh were obtained from 21% PEG8000 at pH 4.0 and of CDbp from 15% PEG8000 at pH 5.9. Recently, hexagonal crystals of human cathepsin D prepared using ammonium sulfate have been reported (Gulnik *et al.*, 1992). Data from both proteins were collected at 20°C with a rotating anode X-ray source using a Siemens X100A area detector (Blum *et al.*, 1987) and processed using the programs XDS and XSCALE (Kabsch, 1988). Table III shows data parameters for each protein. The structures were solved using the molecular replacement features of the program X-PLOR (Brünger, 1990) with a 2.3 Å resolution pepsin structure

Table III. Diffraction data

	Human spleen	Bovine liver/pepstatin A
Space group	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions (Å)	59.9 × 99.6 × 133.6	74.8 × 76.0 × 157.7 Å
Asymmetric unit	two molecules	two molecules
Number of observed reflections to 3.0 Å	53 779	100 236
Number of unique reflections to 3.0 Å	15 735	18 332
Diffraction limit (statistics for the 3.2–3.0 Å resolution shell only)		
Completeness (%)	92	95
R_{merge} (I) (%)	42	35
Completeness to 3.0 Å (all data) (%)	95	98
R_{merge} (I) (all data) (%)	11.5	9.8

Table IV. Quality of structures

	Human spleen	Bovine liver/pepstatin A
Model parameters (monomer)		
Amino acids	339	(341 + 6)
Sugar residues	3	3
Number of non-hydrogen atoms	3236	3354
2-fold constrained	2538	2446
Diffraction agreement		
Resolution range (Å)	6–3	6–3
Number of reflections ($F < 2\sigma_F$)	13 424	15 854
R(free) (%)	25.1	24.1
R (%)	16.1	15.8
R (3–3.2 Å shell) (%)	23	22
Stereochemical ideality		
Bonds (Å)	0.013	0.014
Angles	3.2°	3.3°

as a search model (Abad-Zapatero *et al.*, 1990) code 3PEP in the Brookhaven protein database).

Molecular replacement was carried out first with human cathepsin D using data in the resolution range 8–3.4 Å. Patterson correlation refined peaks of height 0.07 and 0.055 revealed two molecules in the asymmetric unit related by a 2-fold rotation axis 5.7° from the crystallographic *c* axis. Clear translation function peaks were used to produce a model with reasonable packing and a crystallographic residual error R of 45%. The program O (Jones *et al.*, 1991) was then used to replace the pepsin search model with cathepsin D using the sequence of human cathepsin D (Faust *et al.*, 1985; Code CATD_HUMAN in the SwissProt sequence database) and preliminary refinement with the program TNT (Tronrud *et al.*, 1987) and X-ray restrained molecular dynamics refinement using X-PLOR (Brünger, 1992b; Brünger *et al.*, 1987) produced a model with R = 24% and a map in which carbohydrate density was visible at the positions expected for cathepsin D. The model still contained loops in uninterpretable density and regions with poor geometry.

At this point the CDbp crystals were obtained and the structure was solved using the better CDbp data by same methods, except that the pepsin model used for the molecular replacement was replaced with the CDh structure prior to building in the bovine cathepsin D sequence. Refinement was carried out with X-PLOR version 3.0 using the free R value criteria for model accuracy and applying strict non-crystallographic symmetry (Brünger, 1992a). The first cycle of refinement was carried out before pepstatin was incorporated into the model and produced a map in which pepstatin density was clearly visible in both molecules (Bieber *et al.*, 1992). Pepstatin was then added to the model, the polyproline loop was adjusted using coordinates of a similar loop in mouse submandibular renin (Dhanaraj *et al.*, 1992) and further rounds of model adjustment and refinement were used to improve the model.

The CDh model was then rebuilt using the CDbp structure and refined. X-PLOR, O and the program WHATIF (Vriend and Sander, 1993) were used to help identify parts of the models requiring attention. A series of simulated annealing refinements were carried out with X-PLOR to locate parts of the molecules departing from non-crystallographic symmetry using differently selected and weighted symmetry restraints. Minor changes were made to the models followed by conventional atomic position and B factor refinement. The final model refinement statistics are listed in Table IV.

Coordinates and structure factors for both proteins will be deposited with the Brookhaven protein database.

Acknowledgements

We thank S.Labeit for help with bovine cathepsin D cloning and sequencing, J.Chirgwin for a preliminary bovine cDNA sequence, C.Dealwis and T.Blundell for renin coordinates prior to publication, and J.Neumann for analytical ultracentrifugation. We are grateful to S.Kornfeld for providing results prior to publication, and to M.Mareš, B.Hoflack and T.Ludwig for reading the manuscript. M.F. was supported by a 9 month EMBL post-doctoral fellowship.

References

Abad-Zapatero,C., Rydel,T.H. and Erickson,J. (1990) *Proteins*, **8**, 62–81.
Baranski,T.J., Faust,P.L. and Kornfeld,S. (1990) *Cell*, **63**, 281–291.
Baranski,T.J., Koelsch,G., Hartsuck,J.A. and Kornfeld,S. (1991) *J. Biol. Chem.*, **266**, 23365–23372.
Baranski,T.J., Cantor,A.B. and Kornfeld,S. (1992) *J. Biol. Chem.*, **267**, 23342–23348.
Bieber,F., Brachvogel,V., Arni,R., Fusek,M. and Metcalf,P. (1992) *J. Mol. Biol.*, **227**, 1265–1268.
Blum,M., Metcalf,P., Harrison,S.C. and Wiley,C.D. (1987) *Appl. Crystallogr.*, **20**, 235–242.
Bond,J.S. and Butler,P.E. (1987) *Annu. Rev. Biochem.*, **56**, 333–364.
Brünger,A.T. (1990) *Acta Crystallogr.*, **A46**, 46–57.
Brünger,A.T. (1992a) *Nature*, **355**, 472–475.
Brünger,A.T. (1992b) XPLOR manual, Yale University.
Brünger,A.T., Kuriyan,J. and Karplus,M. (1987) *Science*, **235**, 458–460.
Cantor,A.B., Baranski,T.J. and Kornfeld,S. (1992) *J. Biol. Chem.*, **267**, 23349–23356.
Carson,M. (1987) *J. Mol. Graphics*, **5**, 103–106.
Casciola-Rosen.L.A., Renfrew,C.A. and Hubbard,A.L. (1992) *J. Biol. Chem.*, **267**, 11856–11864.
Cho,W.-L. and Raikhel,A. (1992) *J. Biol. Chem.*, **267**, 21823–21829.

Church,W.B., Swenson,L. and James,M.N.G. (1992) *J. Mol. Biol.*, **227**, 577–580.
Davies,D.R. (1990) *Annu. Rev. Biophys. Biophys. Chem.*, **19**, 189–215.
Dhanaraj,V., *et al.* (1992) *Nature*, **357**, 466–472.
Diment,S., Leech,M.S. and Stahl,P.D. (1988) *J. Biol. Chem.*, 6901–6907.
Dunn,B.M., Jimenez,M., Parten,B.F., Valler,M.J., Rolph,C.E. and Kay,J. (1986) *Biochem. J.*, **237**, 899–906.
Factorovich,Y. and Puri,J. (1988) *J. Immunol.*, **141**, 3313–3317.
Faust,P.L., Kornfeld,S. and Chirgwin,J.M. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4910–4914.
Fusek,M., Baudys,M. and Metcalf,P. (1992) *J. Mol. Biol.*, **266**, 555–557.
Gilliland,G.L., Winborne,E.L., Nachman,J. and Wlodawer,A. (1990) *Proteins*, **8**, 82–101.
Gopalan,P., Dufrense,M.J. and Warner,A.H. (1987) *Can. J. Physiol. Pharmacol.*, **65**, 124–129.
Gulnik,S., Baldwin,E., Tarasova,N. and Erickson,J. (1992) *J. Mol. Biol.*, **227**, 265–270.
Hasilik,A. (1992) *Experientia*, **48**, 130–150.
Hasilik,A. and Neufeld,E.F. (1980) *J. Biol. Chem.*, **255**, 4946–4950.
Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *Comput. Appl. Biosci.*, **8**, 189–191.
Horst,M. and Hasilik,A. (1991) *Biochem. J.*, **273**, 355–361.
Imoto,T., Okazaki,K., Koga,H. and Yamada,H. (1987) *J. Biochem.*, **101**, 575–580.
Inagami,T., Misono,K., Chang,J., Takii,Y. and Dykes,C. (1985) In Kostka,V. (ed.), *Aspartic Proteinases and their Inhibitors*. Walter de Gruyer, Berlin, pp. 319–339.
Jones,T.A., Zou,J.-Y. and Cowan,S.W. (1991) *Acta Crystallogr.*, **A47**, 110–119.
Jupp,R.A., *et al.* (1990) *Biochem. J.*, **265**, 871–878.
Kabsch,W. (1988) *J. Appl. Crystallogr.*, **21**, 916–924.
Kaplan,A., Achord,D.T. and Sly,W.S. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 2026–2030.
Kenessey,A., Banay-Schwartz,M., de Guzman,T. and Lajtha,A. (1989) *Neurosci. Res.*, **23**, 454–1456.
Ketcham,C.M. and Kornfeld,S. (1991a) *J. Biol. Chem.*, **267**, 11654–11659.
Ketcham,C.M. and Kornfeld,S. (1991b) *J. Biol. Chem.*, **267**, 11645–11653.
Kornfeld,S. (1986) *J. Clin. Invest.*, **77**, 1–6.
Kornfeld,S. (1992) *Annu. Rev. Biochem.*, **61**, 307–330.
Kornfeld,S. and Mellman,I. (1989) *Annu. Rev. Cell. Biol.*, **5**, 483–525.
Lang,L., Reitman,M.L., Tang,J., Roberts,R.M. and Kornfeld,S. (1984) *J. Biol. Chem.*, **259**, 14663–14671.
Ludwig,T., Griffiths,G. and Hoflack,B. (1991) *J. Cell Biol.*, **115**, 1561–1572.
McIntrye,G.F. and Erickson,A.H. (1991) *J. Biol. Chem.*, **266**, 15438–15445.
Musil,D., Zucic,D., Turk,D., Mayr,I., Engh,R.A., Huber,R., Popovic,T., Turk,V., Towatari,T., Katunuma,N. and Bode,W. (1991) *EMBO J.*, **10**, 2321–2330.
Neufeld,E.F. and Ashwell,G. (1980) In Lennarz,W. (ed.), *The Biochemistry of Glycoproteins and Proteoglycans*. Plenum Press, New York, pp. 241–266.
Offerman,M.K., Chlebowski,J.A. and Bond,J.S. (1983) *Biochem. J.*, **211**, 529–534.
Pearl,L. (1985) In Kostka,V. (ed.), *Aspartic Proteinases and their Inhibitors*. Walter de Gruyter, Berlin, pp. 189–195.
Redecker,B., Heckendorf,B., Grosch,H.-W., Mersmann,G. and Hasilik,A. (1991) *DNA Cell Biol.*, **10**, 423–431.
Rijnbout,S., Aerts,H.M., Geuze,H.J., Tager,J.M. and Strous,G.J. (1991) *J. Biol. Chem.*, **266**, 4862–4868.
Rochefort (1992) *Acta Oncol.*, **31**, 125–130.
Runeberg-Roos,P., Törmäkangas,K. and Östman,A. (1991) *Eur. J. Biochem.*, **202**, 1021–1027.
Šali,A., Veerapandian,B., Cooper,J., Moss,D., Hofmann,T. and Blundell,T. (1992) *Proteins*, **12**, 158–170.
Schorey,J. and Chirgwin,J. (1991) In Dunn,B.M. (ed.), *Structure and Function of the Aspartic Proteases*. Plenum Press, New York, pp. 339–342.
Shewale,F.G. and Tang,J. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 3703–3707.
Sielecki,A.R., Federov,A.A., Boodhoo,A., Andreeva,N.S. and James,M.N.G. (1990) *J. Mol. Biol.*, **214**, 143–170.
Tandon,A.K., Clark,G.M., Chammess,G.C., Chirgwin,J.M. and McGurie,W.L. (1990) *N. Eng. J. Med.*, **322**, 297–302.
Tang,J. (1970) *Methods Enzymol.*, **19**, 406–421.
Tronrud,D.E., Ten Eyck,L.F. and Mathews,B.W. (1987) *Acta Crystallogr.*, **A43**, 489–501.
Ulrich,K., Mersmann,G., Weber,E. and von Figura,K. (1978) *Biochem. J.*, **170**, 643–650.
van Noort,J.M. and ver der Drift,A.C. (1989) *J. Biol. Chem.*, **264**, 14159–14164.
von Figura,K. (1991) *Curr. Opin. Cell. Biol.*. **3**, 642–646.
Vriend,G. and Sander,C. (1993) *J. Appl. Crystallogr.*, **26**, 47–60.
Yonezawa,S., Fujii,K., Maejima,Y., Tamoto,K., Mori,Y. and Muto,N. (1988a) *Arch. Biochem. Biophys.*, **15**, 176–183.
Yonezawa,S., Takahashi,T., Wang,X.-J., Wong,R.N.S., Hartsuck,J.A. and Tang,J. (1988b) *J. Biol. Chem.*, **263**, 16504–16511.

Received on November 12, 1992