*Protein Science* (1999), *8*:283–290. Cambridge University Press. Printed in the USA.


# Crystal structure of wild-type human procathepsin K

J. SIVARAMAN,[1,2,3] MANON LALUMIÈRE,[1] ROBERT MÉNARD,[1,2] AND MIROSLAW CYGLER[1,2,3]

[1]Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Avenue, Montréal, Québec H4P 2R2, Canada
[2]Protein Engineering Network of Centers of Excellence, Montréal, Québec H4P 2R2, Canada
[3]Montreal Joint Centre for Structural Biology, Montréal, Québec H4P 2R2, Canada

(RECEIVED July 20, 1998; ACCEPTED October 14, 1998)

**Abstract**

Cathepsin K is a lysosomal cysteine protease belonging to the papain superfamily. It has been implicated as a major mediator of osteoclastic bone resorption. Wild-type human procathepsin K has been crystallized in a glycosylated and a deglycosylated form. The latter crystals diffract better, to 3.2 Å resolution, and contain four molecules in the asymmetric unit. The structure was solved by molecular replacement and refined to an *R*-factor of 0.194. The N-terminal fragment of the proregion forms a globular domain while the C-terminal segment is extended and shows substantial flexibility. The proregion interacts with the enzyme along the substrate binding groove and along the proregion binding loop (residues Ser138–Asn156). It binds to the active site in the opposite direction to that of natural substrates. The overall binding mode of the proregion to cathepsin K is similar to that observed in cathepsin L, caricain, and cathepsin B, but there are local differences that likely contribute to the specificity of these proregions for their cognate enzymes. The main observed difference is in the position of the short helix $\alpha$3p (67p–75p), which occupies the S′ subsites. As in the other proenzymes, the proregion utilizes the S2 subsite for anchoring by placing a leucine side chain there, according to the specificity of cathepsin K toward its substrate.

**Keywords:** crystal structure; cysteine protease; procathepsin K; proregion; prosegment

In the last few years there has been a dramatic growth in the number of known mammalian cysteine proteases belonging to the papain superfamily (Chapman et al., 1997; Linnevers et al., 1997; Santamaria et al., 1998). They can be divided into two groups based on their tissue distribution. Cathepsins B, C, H, and L tend to be found in most cell types and appear to play a general role in protein turnover. In contrast, cathepsins K, L2, O, S, and W are found in relatively few cell types (Turk et al., 1997), suggesting a more specialized role. Cathepsin K represents an extreme example of the second group since it is unique to the osteoclast, the cell specifically devoted to bone resorption (Littlewood-Evans et al., 1997). In this cell, removal of the organic (mostly type I collagen) and inorganic (hydroxyapaptite) components is carried out in a sealed, low pH compartment termed the ruffled border, where high levels of cathepsin K are located (Littlewood-Evans et al., 1997). The importance of cathepsin K in the removal of the organic phase is strikingly demonstrated by the finding that the human disease pycnodysostosis is a consequence of inactivating mutations in this enzyme (Gelb et al., 1996). Cathepsin K is the major mediator of organic matrix turnover in bone and represents an important target for inhibitor development to treat diseases such as osteoporosis, a condition affecting a large number of post-menopausal women (McGrath et al., 1997; Thompson et al., 1997).

As with the other members of the papain superfamily, cathepsin K is synthesized as an inactive proenzyme, containing a 99 residue proregion, and becomes activated in a low pH environment. In vitro processing of the proenzyme occurs by an autocatalytic cleavage (McQueney et al., 1997). The three-dimensional structure of cathepsin K complexes have recently been reported (McGrath et al., 1997; Zhao et al., 1997), and novel inhibitors of this enzyme have been described (Thompson et al., 1997).

Cysteine proteases from the papain superfamily have been divided into two subfamilies based on sequence analysis (Karrer et al., 1993). Cathepsin K belongs to the larger, cathepsin L subfamily, while cathepsin B is a member of the second subfamily. In addition to some differences within the enzyme sequences, these subfamilies differ significantly in the sequences and lengths of their proregions. Cathepsin B-like enzymes have proregions of approximately 60 residues, while cathepsin L-like proteases have proregions of ~100 residues (Cygler & Mort, 1997). There is little sequence similarity between the proregions across the two subfamilies.

For several of the cysteine proteases, it has been shown that the peptide corresponding to the major portion of the proregion is a potent inhibitor of the proteolytic activity of its parental protease. More detailed studies showed that these propeptides also inhibit

Reprint requests to: Dr. Mirek Cygler, Biotechnology Research Institute, 6100 Royalmount Avenue, Montréal, Québec H4P 2R2, Canada; e-mail: mirek.cygler@bri.nrc.ca.

related proteases but that the degree of inhibition varies significantly (Fox et al., 1992; Taylor et al., 1995; Carmona et al., 1996; Maubach et al., 1997). This ability of propeptides to differentiate between closely related proteases is of considerable interest for development of specific inhibitors of these proteases. The molecular architecture of the proregion-protease interactions for the papain superfamily has been brought to light by the determination of the three-dimensional structures of procathepsin B (Cygler et al., 1996; Turk et al., 1996), procathepsin L (Coulombe et al., 1996), and procaricain (Groves et al., 1996). Despite the differences in the lengths and sequences of the proregions from the two subfamilies, their fold, mode of binding, and inhibition to the enzymes are very similar (Cygler & Mort, 1997; Groves et al., 1998). The structure of procathepsin K adds to the database and provides new structural details for this important therapeutic target.

## Materials and methods

### *Expression and purification*

The vector (pPIC9) and *Pichia pastoris* strain GS115 were purchased from Invitrogen Corporation (San Diego, California). Human procathepsin K was amplified by PCR from a human spleen cDNA library (CLONTECH Laboratories, Inc., Palo Alto, California) using gene-specific primers (5′-GCATCCGCTCGAGAAA AGAGAGGCTGAAGCTCTGTACCCTGAGGAGATACTGGAC-3′ and 5′-CCGGCGGCCGCTCACATCTTGGGGAAGCTGGCC AG-3′). The cDNA construct included a 99 amino acid long proregion (1p–99p) and 215 amino acid long mature enzyme (1–215). Two potential N-glycosylation sites are present in the sequence, one in the proregion (Asn89p) and one in the mature enzyme (Asn99). The gene was cloned into the vector pPIC9 and expressed in the yeast *P. pastoris* as a prepro-$\alpha$-factor fusion construct using the culture conditions recommended by Invitrogen. The supernatant from the culture medium was concentrated and equilibrated with 1.2 M ammonium sulfate and 50 mM Tris·HCl, pH 7.4 in a stirred cell (pro YM10 membrane, Amicon, Beverly, Massachusetts), loaded on a butyl sepharose column and eluted with a 1.2–0 M $(NH_4)_2SO_4$ gradient. Fractions with procathepsin K were pooled, equilibrated with 1.2 M $(NH_4)_2SO_4$, concentrated and re-applied on the same column to remove completely the yellow pigment. Pooled procathepsin K fractions were dialyzed overnight at 4 °C against 20 mM Tris·HCl, pH 7.4 and 150 mM $(NH_4)_2SO_4$ and concentrated using Centriprep (10K cut-off, Amicon). Approximately half of this protein was deglycosylated by incubation with Endoglycosidase F/N Glycosidase mixture (0.8 $\mu$g of glycosidases for 1 mg of procathepsin K) (Boeringher Mannheim, Germany) for 15 h at 23 °C. The deglycosylated proenzyme was applied to Hi-prep 16/60 Sephacryl S100 column, the enzyme fractions pooled and concentrated.

### *Crystallization and data collection*

Crystals of the glycosylated and nonglycosylated procathepsin K were obtained using the hanging drop vapor diffusion method. Initial crystallization conditions were found through screening (screen I, Hampton Research, Laguna Niguel, California) and were optimized further. Glycosylated procathepsin K crystallized from drops containing 2 $\mu$L of protein solution at 5.4 mg/mL concentration, mixed with 2 $\mu$L of mother liquor containing 50 mM monobasic ammonium phosphate, 50 mM Tris·HCl, pH 8 and 15% poly(ethylene glycol) 3350 (PEG3350). These crystals belong to the hexagonal system and have cell dimensions $a = 127.3$, $c = 111.3$ Å. There are likely three molecules in the asymmetric unit, leading to a reasonable $V_m = 2.4$ Å$^3$/Da. Unfortunately, these crystals diffract only to 3.4 Å resolution and decay rapidly. Their investigations were not continued at this time.

Suitable crystallization conditions were also found for deglycosylated procathepsin K through factorial screening. After optimization, the best crystals appeared in the drops containing 2 $\mu$L of protein solution 10.7 mg/mL and 2 $\mu$L of well solution (100 mM Tris·HCl pH 7.0, 5% 2-methyl-2,4-pentanediol, 12 mM $(NH_4)_2SO_4$ and 9% PEG3350). Crystals appeared after 1–2 days and had to be used while fresh. They are monoclinic, space group $P2_1$, cell dimensions $a = 58.7$, $b = 84.4$, $c = 155.6$ Å, $\beta = 90.3°$. Assuming four molecules in the asymmetric unit the $V_m$ coefficient is 2.75 Å$^3$/Da, within the expected range. Numerous attempts made to flash freeze these crystals failed. Under all conditions tested there was a significant loss of the resolution concomitant with very high mosaicity. Consequently, diffraction data were collected at room temperature to a maximum resolution of 3.2 Å. Two crystals were necessary for complete dataset. MAR Research area detector was used for data collection. DENZO and SCALPACK were used for data processing and scaling. A total of 121,990 observations between 15–3.2 Å resolution were reduced to 21,801 unique reflections ($I > 1\sigma(I)$) with $R_{merge} = 0.12$ and 87.2% completeness.

### *Structure solution and refinement*

Molecular replacement using AmoRe (Navaza, 1994) and procathepsin L as a search model (PDB entry code 1CJL) readily gave four expected solutions. The rotation and translation functions were calculated with reflections in 8–4 Å resolution range. Rigid body refinement in AmoRe with four molecules gave an *R*-factor of 0.39 and correlation coefficient of 0.47. This model was subsequently refined using X-PLOR (Brünger, 1993) including data in 8–3.2 Å resolution range. A set of 6% randomly selected reflections was used for the calculation of $R_{free}$. The initial map calculated at this stage had discontinuous density in several parts of the proregion indicating local differences between the proregions of cathepsin L and K. The density was improved by fourfold noncrystallographic symmetry averaging and solvent flattening using programs MAMA and RAVE (Kleywegt & Jones, 1994). Two rounds of rebuilding (O (Jones et al., 1991)), refinement and density averaging allowed to extend the model to include residues 7p–80p and 2–215. The averaged map ended abruptly at the residue occupying the S3 site, suggesting that the C-terminal segment of the proregion is either disordered or that its conformation differs between the four molecules. The inspection of the unaveraged $3F_o - 2F_c$ map provided support for the second hypothesis. The map showed electron density for the remainder of the proregion in two out of four molecules and part of this segment for the third molecule. These were added to the refinement and noncrystallographic restrains were applied only to parts clearly visible in the averaged map (Leu7p–Lys79p and Pro2–Met215).

The final model comprises the sequence from Leu1p to Met215 for molecule A, 1p–82p and 86p–215 for molecule B, 1p–82p and 88p–215 for molecule C, and 1p–80p, 91p–93p and 1–215 for molecule D. No water molecules are included in the current model due to the limited resolution. The final *R*-factor is 19.4%, $R_{free}$ is 25.3% for data with $I > \sigma(I)$ and in the 8–3.2 Å reso-

lution shell. The root-mean-square deviations (RMSDs) for bond lengths and angles are 0.006 Å and 1.41°, respectively. The Ramachandran plot shows one residue per molecule with unfavorable torsion angles, located in the poorly ordered segment near the end of the proregion. The coordinates of all four molecules of procathepsin K are deposited in the Protein Data Bank (accession code 7PCK).

## Results and discussion

The molecule of procathepsin K consists of 99 residues of the proregion (marked with suffix p in Fig. 1) and 215 residues of the mature enzyme (Fig. 1). Numbers in brackets refer to papain sequence. Although the diffraction data extend to only 3.2 Å resolution, fourfold averaging resulted in a map that is continuous along most of the proregion (Fig. 2) and along the entire mature enzyme. Comparison with the recently determined structure of mature cathepsin K (PDB code 1MEM (McGrath et al., 1997)) indicates that there is no significant rearrangement within the protease upon activation. The superposition of cathepsin K with the corresponding portion of procathepsin K gives an RMSD of 0.5 Å for all backbone atoms of residues 1–215. Like other cysteine proteases from the papain superfamily, cathepsin K consists of two domains that assemble to form the active-site cleft at their interface. The catalytic nucleophile, Cys25, and the oxyanion hole residues are provided by the mostly $\alpha$-helical N-terminal domain of the enzyme, while the His162 (159) and Asn182 (175) are from the mostly $\beta$-sheet, C-terminal domain. The proenzyme is a compact molecule, with approximate dimensions of $58 \times 47 \times 34$ Å. The N-terminal fragment of the proregion is globular and binds to cathepsin K through the interactions with the prosegment binding loop (PBL, residue 138–156). The proregion's C-terminal fragment assumes an extended conformation, following the substrate binding cleft toward the N-terminal residue of the mature cathepsin K, located on the opposite side of the enzyme (Fig. 1). The region beyond residue 84p is rather flexible and its conformation differs among the four crystallographically independent molecules.

The access to the active site in the proenzyme is obstructed by the proregion, explaining its lack of enzymatic activity. The direction of the extended fragment of the proregion that fills the substrate binding site is opposite to that taken by the polypeptide substrate. This reverse orientation of the proregion renders it resistant to a cleavage by the Cys-His active site. The inhibition mechanism observed here is the same as described previously in other proenzymes from the papain superfamily (Cygler & Mort, 1997).

### *The proregion*

The N-terminal segment of the proregion consists of three $\alpha$-helices ($\alpha$1p = 7p–17p, $\alpha$2p = 25p–51p, $\alpha$3p = 68p–75p) and a $\beta$-strand ($\beta$1p = 55p–60p). The helices $\alpha$1p and $\alpha$2p are connected by a six residue long loop. They intersect at an angle of 96° and pack tightly together forming a hydrophobic mini-core, which is comprised of the side chains of Leu7p, Trp11p, Leu39p, Trp14p, Trp35p, and the aliphatic part of Glu36p. The three aromatic rings in the mini-core are stacked against each other (Fig. 3). However, the imidazole ring of the nearby His10p is exposed to the solvent. Salt bridges and hydrogen bonds provide additional stabilization of this fold (Table 1) and also tie these two helices to the helix $\alpha$3p. These features are well conserved among the cathepsin L subfamily (Groves et al., 1998).

The long helix $\alpha$2p is connected to the strand $\beta$1p by a two residue long loop. The two secondary structural elements run nearly antiparallel to each other forming a hairpin-like super-secondary structure (Fig. 3). This arrangement is stabilized by interactions originated from the Asp27p–Arg31p–Trp35p–Asn38p–Ile42p–Asn46p residues that are highly conserved in the cathepsin L subfamily proregions—the so-called ERFNIN motif (Karrer et al., 1993; Coulombe et al., 1996; Groves et al., 1996), and in particular Ile42p and Asn46p. The former makes numerous van der Waals contacts while the side chain of Asn46p in helix $\alpha$2p interacts with the backbone of Leu58p ($\beta$1p) via two hydrogen bonds. Equivalent interactions have been found not only in the proregions of


**Fig. 1.** Stereo view of the C$\alpha$ trace of procathepsin K. Catalytic residues and Tyr56p of the proregion are shown in full. The proregion is drawn in thick lines.

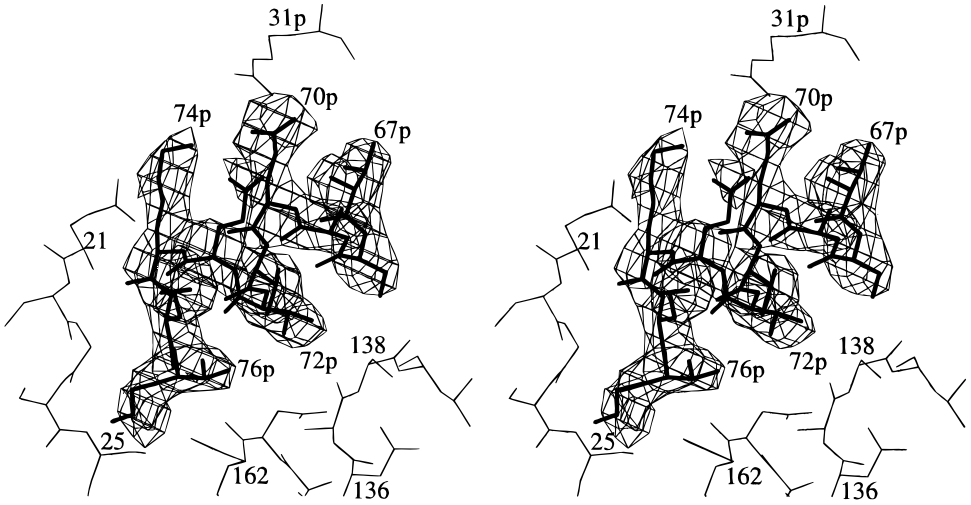


**Fig. 2.** Four-fold averaged ($2F_o - F_c$) electron density map around $\alpha$3p helix of the proregion, displayed at 1$\sigma$ level. The residues of $\alpha$3p are shown in thick lines and the nearby residues from cathepsin K are shown in thin lines.

cathepsin L subfamily (Coulombe et al., 1996) but also in the proregion of cathepsin B (Cygler et al., 1996; Podobnik et al., 1997). Formation of the hairpin super-secondary structure allows the side chain of Tyr56p to extend away from the proregion and to become a pivotal element for the interactions with the PBL of the mature enzyme.

In the region C-terminal to helix $\alpha$3p, the polypeptide chain is in an extended conformation. The residues 75p–84p, which cover the substrate binding site, are well ordered. Beyond residue 84p the proregion displays more flexibility, having somewhat different conformation in each crystallographically independent molecule. Some of the side chains and/or backbone atoms in this segment are disordered and were not located in the electron density. The flexibility observed in this part of the proregion has also been noted in other proenzymes from the papain superfamily (Cygler et al., 1996; Podobnik et al., 1997).

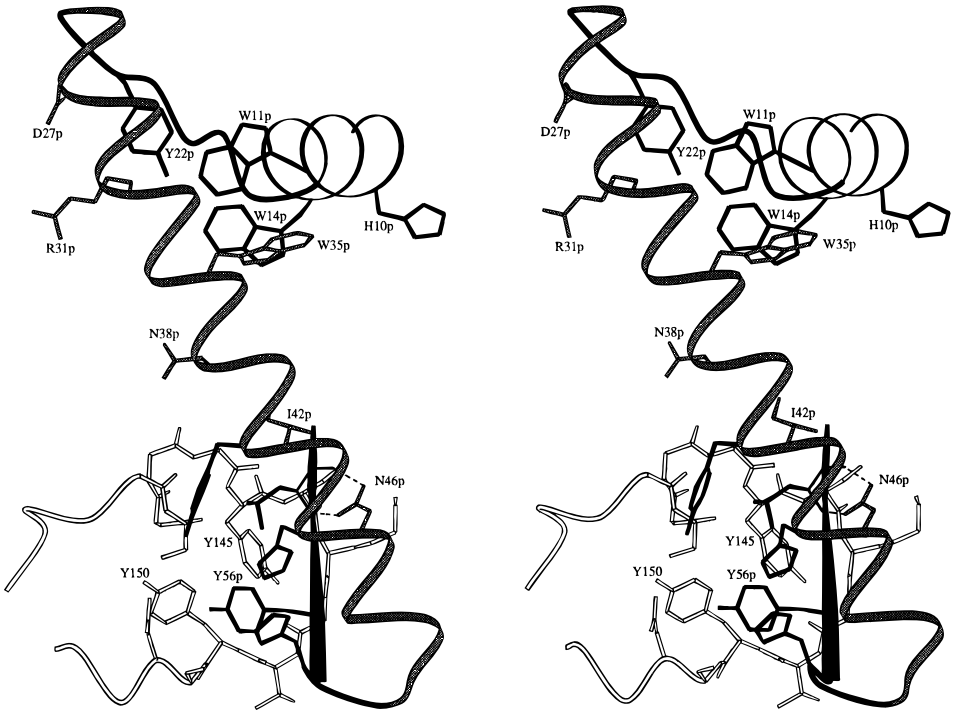


**Fig. 3.** The stereo view of the N-terminal part of the proregion. Packing of helices $\alpha$1p and $\alpha$2p creates a mini-core by stacking of several aromatic residues. The nearby His10p points outside and does not participate in these interactions. The hairpin formed by $\alpha$2p and $\beta$1p is also shown. The residues forming the ERFNIN motif, Tyr56p and several residues from PBL are shown in full.

**Table 1.** *Salt bridges and hydrogen bonds involving side chains that contribute to the stabilization of the proregion*

| H-bond | Dist (Å) | H-bond | Dist (Å) |
|---|---|---|---|
| $^{Lys15p}NZ\ldots OE2^{Glu28p}$ | 2.8 | $^{Glu28p}OE2\ldots NE1^{Trp11p}$ | 3.3 |
| $^{Asp65p}OD2\ldots NZ^{Lys20p}$ | 3.2 | $^{Lys20p}NZ\ldots O^{His62p}$ | 3.3 |
| $^{Glu70p}OE1\ldots NH2^{Arg31p}$ | 2.8 | | |
| $^{Glu70p}OE2\ldots NH1^{Arg31p}$ | 2.9 | $^{Arg31p}NH\ldots O^{Asp65p}$ | 2.7 |
| $^{Asp8p}OD1\ldots NH2^{Arg32p}$ | 2.8 | $^{Arg32p}NE\ldots OE1^{Glu36p}$ | 2.8 |
| $^{Asn46p}OD1\ldots N^{Leu58p}$ | 2.9 | $^{Leu58p}O\ldots ND2^{Asp46p}$ | 3.0 |
| $^{Asn38p}ND2\ldots O^{Gly64p}$ | 2.7 | $^{Asn38p}ND2\ldots O^{Met66p}$ | 3.2 |

*Proregion-cathepsin K interactions*

The proregion is in contact with cathepsin K over an extended area. The surface buried by these interactions is 1,067 Å$^2$ on the enzyme and 1,109 Å$^2$ on the proregion (residues 86p–99p were excluded from these calculations due to their high flexibility). As in procathepsin L and procaricain, the buried surface of cathepsin K can be divided into three adjacent areas: the 138–156 surface loop (prosegment binding loop (PBL) (Cygler et al., 1996)), the hydrophobic groove leading to the active-site (including S′ subsites) and the substrate binding subsites S1–S3.

*Prosegment binding loop (PBL)*

The prosegment binding loop, Ser138–Asn156 (137–155 in papain), is a large omega loop placed on cathepsin K surface. A short section in the middle of this loop, Gly148–Tyr150, forms a fifth strand of the main $\beta$-sheet of the C-terminal domain with two hydrogen bonds between strands four and five, $O^{Leu195}\ldots N^{Tyr150}$ (2.7 Å) and $N^{Ala197}\ldots O^{Tyr\ 150}$ (2.9 Å), and a longer contact $N^{Leu195}\ldots O^{Gly148}$ (4.2 Å). The proregion's $\beta$1p strand (Thr55p–Met60p) extends this main sheet by pairing with Phe144–Val149 segment of PBL in an antiparallel mode (Fig. 3) and making four hydrogen bonds: $N^{Ala59p}\ldots O^{Phe144}$ (2.8 Å), $O^{Glu57p}\ldots N^{Ser146}$ (3.1 Å), $N^{Glu57p}\ldots O^{Lys147}$ (3.3 Å), $O^{Thr55p}\ldots N^{Val149}$ (3.1 Å). Several side chains of the proregion participate in van der Waals contacts with the enzyme. Particularly extensive are the interactions of the aromatic side chain of Tyr56p, which occupies the center of the PBL (Fig. 3) and extends the cluster of aromatic side chains, which include: Phe142 (141), Tyr145 (144), Tyr150 (149), Trp184 (177), and Trp188 (181). Significant contact area is also provided by Leu58p and Thr55p. Other van der Waals interactions involve the N-terminal part of PBL, in particular residues Thr140 (139), Ser141 (140), Phe144 (143), and Gln143 (142), which come into contact with the face of helix $\alpha$2p containing residues Asn38p, Tyr41p, and Ile42p. The total surface area of PBL in contact with the proregion is 512 Å$^2$.

*Interactions within the S′-side groove*

The groove on the S′ side of the active site is lined with prosegment residues that come from the loop following strand $\beta$1p and from the helix $\alpha$3p. The beginning of the loop forms a tight bend clustering together the side chains of three residues: Asn61p, His62p, and Leu63p (Fig. 4). They provide hydrogen bonds through the side chains of Asn61p (to $O{=}C^{Gln143}$) and His62p (to $O{=}C^{Trp184}$ and to the side chain of Asn187)—and van der Waals contacts to Gln143, Phe144, Trp184, Trp188, and Asn187.

Helix $\alpha$3p packs against cathepsin K with its hydrophobic face and contacts the enzyme through residues Val71p, Val72p, and Met75p (Fig. 4). It makes van der Waals contacts with residues Ser138–Gln143 (start of PBL), Trp184, Gln19, and Gln21. In particular, Met75p, which occupies the S2′ subsite, extends along the crevice formed by the indole ring of Trp184 and the Gln19–Gly20–Gln21 stretch. Hydrogen bonds are made by $OG^{Ser68p}$ (to $O{=}C^{Ser138}$) and $O{=}C^{Lys74p}$ (with the NH groups of Gln21 and Cys22). The first residue of the extended segment, Thr76p, occupies the S1′ subsite and fits between Asn161 and His162. Its side chain forms a hydrogen bond to $O{=}C^{Asp161}$ (2.9 Å).

The proper orientation of helix $\alpha$3p relative to the N-terminal fragment of the propeptide is maintained through interactions between $\alpha$3p and $\alpha$2p, in particular a salt bridge between Glu70p and Arg31p (from the ERFNIN motif), and hydrogen bonds formed from $ND2^{Asn38p}$ to $O{=}C^{Met66p}$ and $O{=}C^{Gly64p}$ (Table 1).

*Interactions within the S subsites*

The proregion follows the S subsites in an extended conformation. The protein we have crystallized has the wild-type active-site and it was of considerable interest to determine if the cysteine had been oxidized. The electron density map suggested that the nucleophilic Cys25 was not oxidized in the crystals. A similar observation was made in the case of wild-type human procathepsin B (Podobnik et al., 1997) but in the case of procathepsin L the active-site cysteine was found oxidized (R. Coulombe & M. Cygler, unpubl. data). The proregion of procathepsin K approaches the catalytic Cys25 closely (Fig. 4). The shortest distance is to the Thr76p–Gly77p peptide bond (3.4 Å), made possible due to the lack of a side chain on residue 77p. The carbonyl of this Gly77p is hydrogen bonded to $NH^{Gly66}$ (2.8 Å). The proregion utilizes also the oxyanion hole. The carbonyl oxygen of Thr76p extends into this region and forms hydrogen bonds with $NE2^{Gln19}$ (3.1 Å) and $N^{Cys25}$ (3.1 Å).

The S2 subsite plays the most important role in determining substrate specificity of enzymes from the papain superfamily (Storer, 1991) The preferred P2 residue for cathepsin K is a leucine (Bromme et al., 1996). The proregion utilizes this subsite for binding and, like the best substrate, inserts there the side chain of Leu78p.

*Comparisons with other proenzymes*

The level of sequence similarity between the proregions of cathepsin L subfamily of cysteine proteases is lower than that for the mature enzymes and is in most cases below 30%. Nevertheless, the three known three-dimensional structures of these proregions are very similar (Fig. 5) and it is expected that this will be the case for other members of this subfamily. All of them utilize the S2 subsite for binding of a branched residue or a phenylalanine. The insertion into the S2 subsite is fastened by two hydrogen bonds to Gly66 (cathepsin K) or its equivalent, located opposite. The binding of the proregion and the recognition of the enzyme are then confined to the area between the two constant attachment points: the aromatic residue in the center of PBL and a hydrophobic residue in the S2 subsite. The main determinants of the recognition have to be located within this area of the proregion.

*Helix $\alpha$3p*

The present structure confirms that the location of helix $\alpha$3 differs the most between the proregions (Fig. 5). The center of


**Fig. 4.** The interactions between the proregion and the enzyme in the S and S′ sites. Helix $\alpha$3p and the preceding loop occupies the S′ side. The carbonyl group of Thr76p occupies the oxyanion hole and the S2 subsite is occupied by Leu78p. The relevant hydrogen bonds between the proregion and the enzyme are shown in dashed lines. The proregion is shown in dark lines and hydrogen bonds in dotted lines.

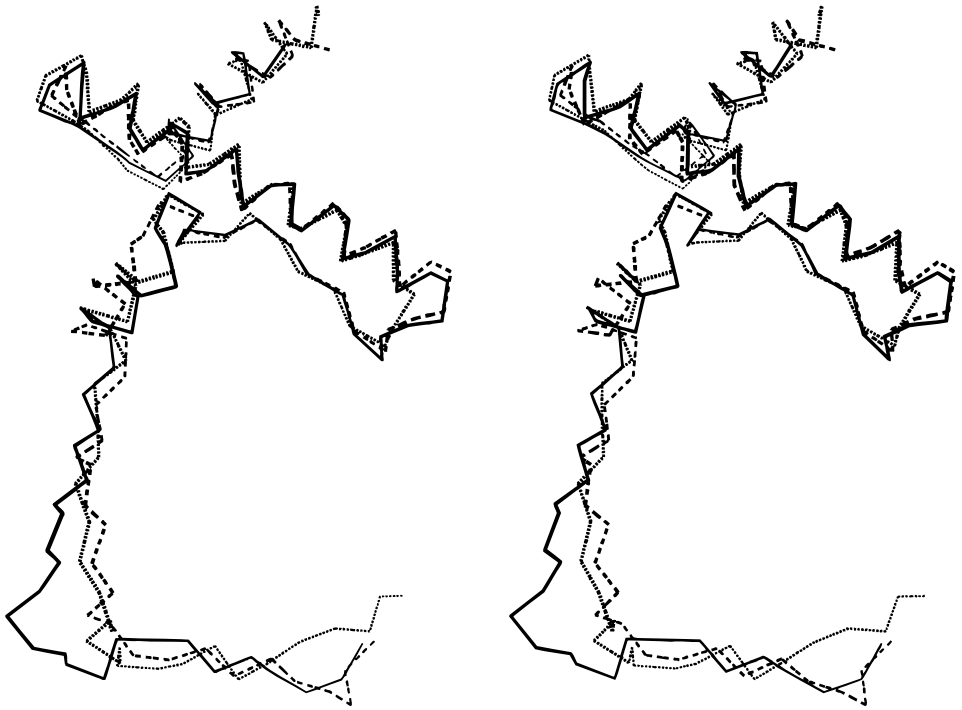

**Fig. 5.** The superposition of C$\alpha$ traces of the proregions of cathepsin K (solid line), cathepsin L (dashed line), and caricain (dotted line). The transformation matrix was calculated based on the corresponding C$\alpha$ atoms of residues from the N-terminus to residues in the enzyme active site.

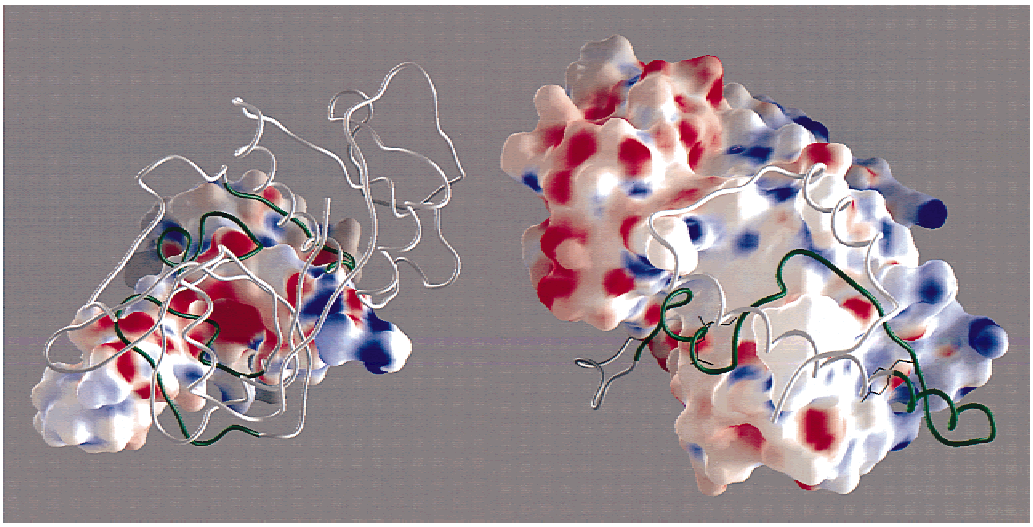

**Fig. 6.** The electrostatic potential of the contacting surfaces of the proregion and cathepsin K. The potentials were calculated independently for each part. **A:** Surface of the proregion facing the enzyme, residues 80–99 were omitted from the picture. Cathepsin K is shown as a backbone worm. The parts of the enzyme in contact with the proregion are shown in green. The long loop (bottom left) is the PBL. **B:** The surface of cathepsin K facing the proregion. The proregion is shown as a backbone worm with the part in close contact with the enzyme painted green. The residues Tyr56p, Val71p, and Val72p are shown in full. The two views are related by a 180° rotation along vertical axes. The complementarity of the electrostatic potential distributions along contacting surfaces is also visible.

helix $\alpha$3p is located approximately 2 Å closer to the surface of cathepsin K than in procathepsin L and procaricain. The difference in the position of $\alpha$3p between procaricain and procathepsin L corresponds to a shift along the helix axis and was postulated to be a result of an insertion between the P1 and P2 subsites in the former. There is no such insertion in procathepsin K when compared to procathepsin L and the shift is in a direction perpendicular rather than parallel, to the helix axis.

The observed difference likely results from a difference in the sequence in the center of the helix $\alpha$3p face that is directed toward the enzyme. The two key side chains for contacting the enzyme are Phe71p–Arg72p in procathepsin L and Phe78p–Asn79p in procaricain. The phenylalanine points directly toward the enzyme and contacts many residues. The corresponding side chains in procathepsin K, Val71p–Val72p, are small and to maintain good van der Waals contacts, the helix has to move closer to the protein surface. Sequence alignment (Coulombe et al., 1996) of the proregions shows that most of them contain phenylalanine in the first of these two positions and a large side chain (Arg-Lys-Asn-Gln) in the second. Apart from procathepsin K the other proregions with small side chains in these positions are those of cathepsins S, with Val-Met/Ile sequences, and of cathepsins H, with Val/Thr-Lys/Ala sequences. Additional sequence similarity between procathepsins K and S in the subsequent segment covering the substrate binding site suggests that the helix $\alpha$3p in these two enzymes is positioned in a similar way.

*PBL*

The mode of binding to PBL is very much the same in the known structures of the proenzymes from the papain superfamily. An aromatic residue packs against the center of the PBL and adds to the aromatic cluster of highly conserved residues. The short, two-stranded $\beta$-sheet is also formed in a similar manner in all of them. What differs somewhat is the location of the N-terminal part of the long helix $\alpha$2 resulting from a slightly different curvature and tilt of the helix in each proenzyme. The largest deviation is observed for procaricain.

As in the other proenzymes, electrostatic interactions are also of importance for procathepsin K. The electrostatic potential distributions, calculated for the proregion alone or for the enzyme, are complementary along the interacting surfaces (Fig. 6), as in other proenzymes (Groves et al., 1998). However, the signature of the cathepsin K proregion differs from the other enzymes, supporting the notion that electrostatic forces play a role in determining the specificity of recognition between the proregion and its cognate enzyme.

## Acknowledgment

We would like to thank Dr. J.S. Mort for careful reading of the manuscript and helpful comments.

## References

Bromme D, Okamoto K, Wang BB, Biroc S. 1996. Human cathepsin O2, a matrix protein-degrading cysteine protease expressed in osteoclasts. Functional expression of human cathepsin O2 in *Spodoptera frugiperda* and characterization of the enzyme. *J Biol Chem 271*:2126–2132.

Brünger AT. 1993. *X-Plor Version 3.1*. New Haven, CT: Yale University.

Carmona E, Dufour E, Plouffe C, Takebe S, Mason P, Mort JS, Menard R. 1996. Potency and selectivity of the cathepsin L propeptide as an inhibitor of cysteine proteases. *Biochemistry 35*:8149–8157.

Chapman HA, Riese RJ, Shi GP. 1997. Emerging roles for cysteine proteases in human biology. *Annu Rev Physiol 59*:63–88.

Coulombe R, Grochulski P, Sivaraman J, Menard R, Mort JS, Cygler M. 1996. Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *EMBO J 15*:5492–5503.

Cygler M, Mort JS. 1997. Proregion structure of members of the papain superfamily. Mode of inhibition of enzymatic activity. *Biochimie 79*:645–652.

Cygler M, Sivaraman J, Grochulski P, Coulombe R, Storer AC, Mort JS. 1996. Structure of rat procathepsin B: Model for inhibition of cysteine protease activity by the proregion. *Structure 4*:405–416.

Fox T, de Miguel E, Mort JS, Storer AC. 1992. Potent slow-binding inhibition of cathepsin B by its propeptide. *Biochemistry 31*:12571–12576.

Gelb BD, Shi GP, Chapman HA, Desnick RJ. 1996. Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. *Science 273*:1236–1238.

Groves MR, Coulombe R, Jenkins J, Cygler M. 1998. Structural basis for specificity of papain like cysteine protease proregions toward their cognate enzymes *Proteins Struct Funct Genet 32*:504–514.

Groves MR, Taylor MAJ, Scott M, Cummings NJ, Pickersgill RW, Jenkins JA. 1996. The prosequence of procaricain forms an alpha-helical domain that prevents access to the substrate-binding cleft. *Structure 4*:1193–1203.

Jones TA, Zhou JY, Cowan SW, Kjeldgaard M. 1991. Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A47*:110–119.

Karrer KM, Peiffer SL, DiTomas ME. 1993. Two distinct gene subfamilies within the family of cysteine protease genes. *Proc Natl Acad Sci USA 90*:3063–3067.

Kleywegt GJ, Jones TA. 1994. Halloween . . . masks and bones. In: Bailey S, Hubbard R, Waller D, eds. *From first map to final model*. Warrington, UK: SERC Daresbury Laboratory. pp 59–66.

Linnevers CJ, Smeekens SP, Bromme D. 1997. Human cathepsin W, a putative cysteine protease predominantly expressed in CD8+ T-lymphocytes. *FEBS Lett 405*:253–259.

Littlewood-Evans A, Kokubo T, Ishibashi O, Inaoka T, Wlodarski B, Gallagher JA, Bilbe G. 1997. Localization of cathepsin K in human osteoclasts by in situ hybridization and immunohistochemistry. *Bone 20*:81–86.

Maubach G, Schilling K, Rommerskirch W, Wenz I, Schultz JE, Weber E, Wiederanders B. 1997. The inhibition of cathepsin S by its propeptide—Specificity and mechanism of action. *Eur J Biochem 250*:745–750.

McGrath ME, Klaus JL, Barnes MG, Bromme D. 1997. Crystal structure of human cathepsin K complexed with a potent inhibitor. *Nat Struct Biol 4*:105–109.

McQueney MS, Amegadzie BY, D'Alessio K, Hanning CR, McLaughlin MM, McNulty D, Carr SA, Ijames C, Kurdyla J, Jones CS. 1997. Autocatalytic activation of human cathepsin K. *J Biol Chem 272*:13955–13960.

Navaza J. 1994. AMoRe: An automated package for molecular replacement. *Acta Crystallogr A50*:157–163.

Podobnik M, Kuhelj R, Turk V, Turk D. 1997. Crystal structure of the wild-type human procathepsin B at 2.5 Å resolution reveals the native active site of a papain-like cysteine protease zymogen. *J Mol Biol 271*:774–788.

Santamaria I, Velasco G, Pendas AM, Fueyo A, Lopez-Otin C. 1998. Cathepsin Z, a novel human cysteine proteinase with a short propeptide domain and a unique chromosomal location [in process citation]. *J Biol Chem 273*:16816–16823.

Storer AC. 1991. Engineering of proteases and protease inhibition. *Curr Opin Biotechnol 2*:606–613.

Taylor MAJ, Baker KC, Briggs GS, Connerton IF, Cummings NJ, Pratt KA, Revell DF, Freedman RB, Goodenough PW. 1995. Recombinant proregions from papain and papaya proteinase IV-are selective high affinity inhibitors of the mature papaya enzymes. *Protein Eng 8*:59–62.

Thompson SK, Halbert SM, Bossard MJ, Tomaszek TA, Levy MA, Zhao B, Smith WW, Abdel-Meguid SS, Janson CA, D'Alessio KJ, et al. 1997. Design of potent and selective human cathepsin K inhibitors that span the active site. *Proc Natl Acad Sci USA 94*:14249–14254.

Turk D, Podobnik M, Kuhelj R, Dolinar M, Turk V. 1996. Crystal structures of human procathepsin B at 3.2 and 3.3 Angstroms resolution reveal an interaction motif between a papain-like cysteine protease and its propeptide. *FEBS Lett 384:*211–214.

Turk B, Turk V, Turk D. 1997. Structural and functional aspects of papain-like cysteine proteinases and their protein inhibitors. *Biol Chem 378*:141–150.

Zhao B, Janson CA, Amegadzie BY, D'Alessio K, Griffin C, Hanning CR, Jones C, Kurdyla J, McQueney M, Qiu X. 1997. Crystal structure of human osteoclast cathepsin K complex with E-64. *Nat Struct Biol 4*:109–111.