Crystal structure of human cathepsin S

ARTICLE *in* PROTEIN SCIENCE · JUNE 1998

Impact Factor: 2.85 · DOI: 10.1002/pro.5560070604 · Source: PubMed

4 AUTHORS, INCLUDING:

James T Palmer
Biota Pharmaceuticals Incorporated
39 PUBLICATIONS **2,256** CITATIONS
SEE PROFILE

John Somoza
Gilead Sciences
40 PUBLICATIONS **1,385** CITATIONS
SEE PROFILE

Protein Science (1998), 7:1294–1302. Cambridge University Press. Printed in the USA.

Crystal structure of human cathepsin S

MARY E. McGRATH, JAMES T. PALMER, DIETER BRÖMME,[1] AND JOHN R. SOMOZA

Axys Pharmaceuticals, Inc., 180 Kimball Way, South San Francisco, California 94080

(RECEIVED December 22, 1997; ACCEPTED February 20, 1998)

Abstract

We have determined the 2.5 Å structure (R_{cryst} = 20.5%, R_{free} = 28.5%) of a complex between human cathepsin S and the potent, irreversible inhibitor 4-morpholinecarbonyl-Phe-hPhe-vinyl sulfone-phenyl. Noncrystallographic symmetry averaging and other density modification techniques were used to improve electron density maps which were nonoptimal due to systematically incomplete data. Methods that reduce the number of parameters were implemented for refinement. The refined structure shows cathepsin S to be similar to related cysteine proteases such as papain and cathepsins K and L. As expected, the covalently-bound inhibitor is attached to the enzyme at Cys 25, and enzyme binding subsites S3–S1′ are occupied by the respective inhibitor substituents. A somewhat larger S2 pocket than what is found in similar enzymes is consistent with the broader specificity of cathepsin S at this site, while Lys 61 in the S3 site may offer opportunities for selective inhibition of this enzyme. The presence of Arg 137 in the S1′ pocket, and proximal to Cys 25 may have implications not only for substrate specificity C-terminal to the scissile bond, but also for catalysis.

Keywords: antigen presentation; cysteine protease; structure-based drug design; vinyl sulfone

The lysosomal cathepsins are papain-like cysteine proteases that are responsible for normal cellular protein degradation. They are, perhaps, more well known for their purported roles in pathological tissue destruction. Excessive or inappropriate activity of these enzymes has been linked to diseases such as emphysema, arthritis, osteoporosis, asthma, glomerulonephritis, and metastases of various cancers. Consequently, specific inhibition of individual cathepsins may assuage these afflictions. Design of potent inhibitors is greatly aided by knowledge of the atomic resolution structure of the target molecule. Though the atomic coordinates of cathepsin B (Musil et al., 1991) have been available for some time, it is only recently that the X-ray structures of related human enzymes have been elucidated: structures of the pro- (Coulombe et al., 1996) and mature (Fujishima et al., 1997) forms of cathepsin L were reported last year, as was the structure of the osteoclastic enzyme, cathepsin K (McGrath et al., 1997; Zhao et al., 1997). The structure of human procathepsin B has also just been described (Podobnik et al., 1997). Most recently, the structure of porcine cathepsin H was published (Gunčar et al., 1997). Though a number of less abundant cathepsins continue to be identified, cathepsins S is the last of the high-profile lysosomal cysteine proteases to be structurally characterized.

Cathepsin S is quite similar to the above-mentioned enzymes, with 57% identity to cathepsins L and K, and 31% identity to cathepsin B. However, it is distinguished from the related proteases by its high stability at neutral and slightly basic pH (Kirschke et al., 1989; Brömme et al., 1993), and by its limited tissue distribution. While cathepsins L and B are found in a large variety of tissues, cathepsin S is elaborated mainly in lymphatic tissues, with high levels measured in spleen, and in lung macrophages (Kirschke et al., 1989; Qian et al., 1991; Shi et al., 1994). This localization coupled with the broad substrate specificity observed for cathepsin S suggested involvement in the immune response. Recent work has characterized cathepsin S as playing a major role in MHC-II mediated antigen presentation (Riese et al., 1996). Though other proteases appear to be responsible for generation of the 15–25 amino acid peptides that are displayed by the antigen presenting cells, studies demonstrate that cathepsin S readies the peptide-binding site of MHC-II molecules. Apparently, cathepsin S is responsible for the degradation of invariant chain, Ii, which is bound in the antigen-binding groove of class II molecules as they traverse the Golgi apparatus (Riese et al., 1996; Villadangos et al., 1997). Cathepsin L has just been shown to play an equivalent role in the cortical thymus for the display of self-peptides (Nakagawa et al., 1998). Only after this proteolytic step are MHC-II molecules competent for antigen binding. Selective inhibition of cathepsin S has been shown to alter processing of Ii, resulting in reduced immune reaction to ovalbumin (Riese et al., 1998). The resultant attenuation of immune response may be therapeutically useful in disease states corresponding to hyperimmune responses, such as asthma. Thus, the atomic resolution structure of cathepsin S is useful for the design of specific inhibitors which could be developed for unmet medical needs.

In addition, the structure of cathepsin S may be an important component of drug design programs where it is not the therapeutic

Reprint requests to: Mary E. McGrath, Axys Pharmaceuticals, Inc., 180 Kimball Way, South San Francisco, California 94080; e-mail: mcgrath@arris.com.

[1]Present address: Department of Human Genetics, Mount Sinai Hospital School of Medicine, Box 1498, Fifth Avenue at 100th Street, New York, New York 10029.

target. Attempts to curtail the activity of related disease-causing cathepsins are more likely to succeed when they strategically pinpoint the causative protease, but do not inactivate cathepsin S or other antitargets.

Results and discussion

Structure determination and refinement

Structure determination of cathepsin S was hampered by prolonged efforts to obtain diffraction-quality crystals. Thus, while the high-resolution structure of cathepsin K (McGrath et al., 1997) was reported within three years of first discovery of the enzyme (Tezuka et al., 1994), this description of cathepsin S comes more than 20 years after its initial characterization (Turnsek et al., 1975).

Only one data-quality crystal was obtained over an 18 month period, and the resulting data set was significantly and systematically incomplete. More specifically, a substantial cone of reflections along the L axis was absent from the data, and these missing reflections led to two predictable sets of problems. First, the incompleteness of the data decreased the number of data points that could be used to constrain the cathepsin S model during refinement. Second, the missing data led to streakiness in the electron density maps along the z direction, and to substantial Fourier series truncation artifacts throughout the maps.

Use of related enzymes for structure determination

The incompleteness of the data did not appear to complicate or hinder the structure determination by molecular replacement. Two homology models were used as search models. The first was based on the structures of papain and cruzain, which are, respectively, 45% and 38% identical with cathepsin S. A self-rotation function with the $\kappa = 180°$ plane strongly confirmed (peak height = 17.5σ, next highest peak = 2.3σ) the local symmetry of two molecules in the asymmetric unit. The subsequent locked rotation function produced a 6.4σ solution (next highest peak was 4.3σ), while the translation searches gave correct solutions which were convincing, but not outstanding (8.1σ peak vs. incorrect solution of 6.5σ). Though initially useful, refinement of this model proved difficult: it commenced with R_{cryst} = 48%, and converged at a fairly high R_{free} (~40%). Fortunately, the structure of procathepsin (Coulombe et al., 1996), whose mature region shares 57% sequence identity with cathepsin S, became available, and the mature region was used as a search model. Rotation and translation searches were carried out with AmoRe (Navaza, 1994) and yielded a residual of 37.5% with a correlation coefficient of 56.5. The initial R_{cryst} upon starting refinement was 39.5%. This new model refined well and led directly to the current structure. The only region that did not have sufficient electron density for building and does not have a homolog in the related enzymes is the 58–61 loop. Consequently, we have not placed residues 58a–c in our final model.

Strategies for refinement with incomplete data

Several tactics were pursued to mitigate the effects brought about by the missing data. During refinement we sought to minimize the number of adjustable parameters used to describe the cathepsin S model. Throughout most of the refinement, bond distances and angles were kept fixed, and only the torsional angles were adjusted. This approach substantially reduced the number of refineable parameters compared to conventional refinement (Rice & Brünger, 1994; Brünger & Rice, 1997). Also, since cathepsin S crystallized with two molecules in the asymmetric unit, the number of parameters was further reduced by forcing the two monomers to behave identically during refinement.

Interpretability of the electron density maps was ameliorated in several ways. Most importantly, the maps were improved through density modification techniques, including solvent flattening, histogram matching, and noncrystallographic symmetry averaging (Cowtan & Main, 1996). A further attempt was made to improve the electron density maps by filling in missing data with calculated structure factors. This strategy has the advantage of removing Fourier truncation artifacts, but the disadvantage of biasing the maps toward the model. In practice, the biasing of the maps was too strong to make them useful. A final tool was to make F_c maps using only the data points that are present in the measured data sets. Although these maps contain no information about how to alter the model, they reveal the artifacts present in the electron density maps, and are, therefore, useful for gauging the theoretical limitations of the map content.

Despite these approaches, the current cathepsin S maps still suffer from many of the problems brought about by the incompleteness in the data. There are sections of the protein that remain poorly defined by the electron density. However, the crystallographic and geometric statistics for the current cathepsin S model (Table 1) attest to its validity. In addition, further tests, as described below, were used to confirm the positions of important residues.

Special care was taken to ensure that the catalytic and substrate/inhibitor binding residues were well determined. This group is comprised of residues 19, 25, 61, 65–68, 133, 137, 157, 159, 160, 175–177, 205. It also includes the atoms of the inhibitor, 4-morpholinecarbonyl-Phe-hPhe-vinyl sulfone-phenyl, which shall be referred to as APC 2848. Simulated annealing omit maps were used to obtain an unbiased view of the electron density in these regions (Fig. 1A). Based on these maps, we were able to unambiguously place all of the main chain atoms and most of the side-chain atoms of the active site residues. However, there are a few residues for which the corresponding electron density was insufficient to position all or part of side chain. These exceptions correspond to side chains or parts of side chains that are oriented perpendicular to the z axis, and are consequently more likely to have broken density. In these cases, the refined position of the side chain was taken to be correct, as long as the side chain consistently refined to the same position. To test the accuracy of the refinement, questionable side chains were manually moved by approximately 90° and then refined. In most cases, the side chain returned to its initial position, which was then deemed correct. The two exceptions to this were the ring moiety of Trp 177 and the methyl groups of Val 157, which do not robustly refine to a single orientation. In the case of Trp 177, the position of this side chain is conserved throughout this family of proteases, and is likely to be conserved in cathepsin S as well. In addition, the temperature factors of the ring atoms refine to substantially lower values when the side chain is placed in the consensus orientation that they do when the ring is rotated by approximately 90° (average B for ring atoms = 6.4 Å^2 vs. 27.4 Å^2), lending credence to the idea that this is the correct orientation. The positions of the Cγ1 and Cγ2 atoms of Val 157 remain ambiguous.

The robustness of the refinement was evident in other ways. For example, though cathepsin K and cathepsin L share 57% sequence identity with cathepsin S, cathepsin K was not incorporated into

Table 1. *Crystallographic data for cathepsin S: APC 2848 complex*

Space group	R3
Cell dimensions (Å)	a = 107.61, c = 105.19
	R-Axis IV
Resolution (Å)	2.5
Sigma cutoff	−3
Data reduction package	DENZO
Number of observations	23,957
Number of independent reflections	10,991
Overall completeness to 2.5 Å (%)	71.0
% completeness 2.6 Å – 2.5 Å shell	65.0
Number of crystals	1
R_{merge}[a]	0.097
Structure determination	Molecular replacement using Xsight(Biosym/MSI), and AMoRe
Model	Mature region of procathepsin L
Refinement	X-PLOR
Resolution range	6–2.5 Å
Number of reflections used	10,306
For monomer, # of	
Protein atoms	2,018
Ordered waters	25
Inhibitor molecules	1
R_{cryst}[b]	0.205
R_{free}[c]	0.285
RMSDs	
Bonds (Å)	0.012
Angles (deg)	1.62
Dihedrals (deg)	26.4
Impropers (deg)	1.42

[a] $R_{merge} = \sum(I - \langle I \rangle)^2 / \sum I^2$ where $\langle I \rangle$ is the mean intensity.

[b] $R_{cryst} = \sum||F_o| - |F_c|| / \sum|F_o|$ where F_o and F_c are the observed and calculated structure factor amplitudes, respectively.

[c] R_{free} is calculated the same as R_{cryst}, but is for 10% of the data which is withheld from refinement.

the search model. Comparison of the refined structures of these three proteases shows that segments of cathepsin S which more closely resemble cathepsin K in primary structure (e.g., residues 96–110, 114–118, 168a–d), have moved during refinement to occupy positions similar to those found in cathepsin K, and clearly quite different from those in cathepsin L.

Structure description: Overall

The structure of cathepsin S is quite similar to the paradigm for this family, the plant enzyme papain. Consequently, the papain numbering system is used throughout the manuscript. After superpositioning of cathepsin S with related enzymes by a group of 90 "core" residues (Fig. 2), the RMS deviations (RMSDs) for the main-chain atoms of cathepsin S compared with papain, cruzain, cathepsin K, and procathepsin L were found to be 0.55 Å, 0.61 Å, 0.51 Å, and 0.51 Å, respectively. When aligned with papain, cathepsin S has an insertion of three amino acids after residue 58, and four after 168. Single residue insertions are found after 78, 127, and 152, while 104 and 195–197 are missing in cathepsin S.

Cathepsin S is a single chain monomeric protein of 217 amino acids. One domain is partly helical and consists of residues 12 to 111 and 208 to 211 with helices extending from 25–40, 50–56, and 68–78. The second domain is arranged as a six-stranded beta barrel composed of residues 1 to 11 and 112 to 207 with a small helix coiling through residues 119–127. One additional helical turn is found from residues 139–143. The juxtaposition of domains creates the long, narrow active site cleft bordered by the 25–40 helix on one side and the 157–161 and 130–133 strands on the other. The cleft is stabilized by two strands, 204–211 and 108–113, which act as straps crossing from one domain to the other. Side chains of the active site residues Cys 25 and His 159 protrude into the cleft, with the latter residue stabilized by the third catalytic entity, Asn 175. The catalytically active species is believed to be the thiolate-imidazolium ion pair that results in an especially nucleophilic cysteine (Polgar, 1974; Rullmann et al., 1989). In cathepsin S, as well as its structurally characterized relatives, substrate binding sites flank Cys 25 and His 159, with three clear binding pockets observed for substrate residues P3, P2, and P1 (Schechter & Berger, 1967), and only one obvious binding pocket C-terminal to the scissile bond, for P1′.

Binding of the irreversible inhibitor

In this cathepsin S structure, four binding pockets are occupied by the covalent inhibitor, APC 2848 (Fig. 1B). It is one of a series of

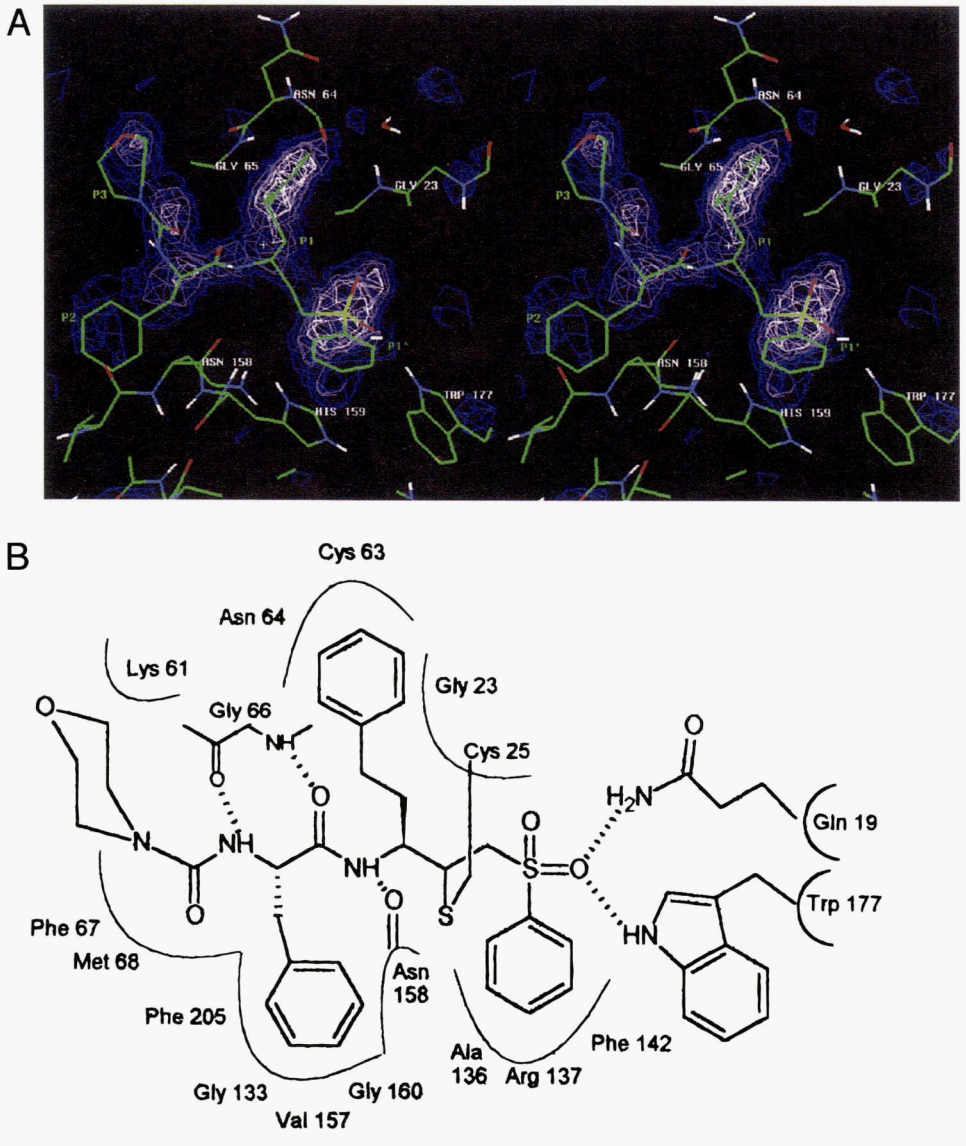

Fig. 1. A: $F_o - F_c$ electron density at the cathepsin S active site showing APC 2848 superimposed on the density, shown in stereo. The P3 through P1′ moieties of the inhibitor are labeled. This is a simulated annealing omit map in which the inhibitor was omitted from the calculation. The contour levels are blue = 2σ, purple = 3σ, and pink = 4σ. **B:** The structure of APC 2848 and schematic showing contacts made with the cathepsin S substrate binding sites.

extremely potent mechanism-based inhibitors of cysteine proteases that utilize vinyl sulfones as Michael acceptors. These inhibitors require the enhanced nucleophilicity of Cys 25 in the papain-like cysteine proteases for their activity, and are not reactive with normal cysteines, or with serine proteases (Palmer et al., 1995). Cathepsin S is particularly vulnerable to this class, with k_{inact}/K_i values in excess of 10^7 M^{-1} s^{-1} for several molecules which vary in the P2 substituent, and a value of 5.0×10^6 M^{-1} s^{-1} for the inhibitor used in this study (Brömme et al., 1996b).

APC 2848 is irreversibly linked to the enzyme as a consequence of nucleophilic attack of Cys 25 upon a vinyl carbon. The *si* face of the inhibitor is the site of attack, as seen for a similar cathepsin K structure (McGrath et al., 1997), and as predicted for cysteine-protease inhibitor complexes (James, 1994). The vinyl sulfone inhibitors are more substrate-like in their interactions with the specificity pockets than are other inhibitor types because they are not shifted by one carbon like the halomethyl ketones (Drenth et al., 1976) or oriented backward with respect to substrate as are the peptidyl epoxides (Turk et al., 1995) and pro-regions (Cygler et al., 1996). This more realistic binding undoubtedly contributes to enhanced protease inhibition as the inhibitor side chains fall naturally into the arrayed binding pockets. Furthermore, examination of the atomic resolution interactions should be more useful for understanding binding and catalysis of natural substrates.

Binding determinants C-terminal to the scissile bond

Just C-terminal to the covalent link, the phenyl sulfone occupies the S1′ pocket of cathepsin S. As seen in a similar cathepsin K complex, one sulfone oxygen interacts with Trp 177 Ne1 (3.2 Å)

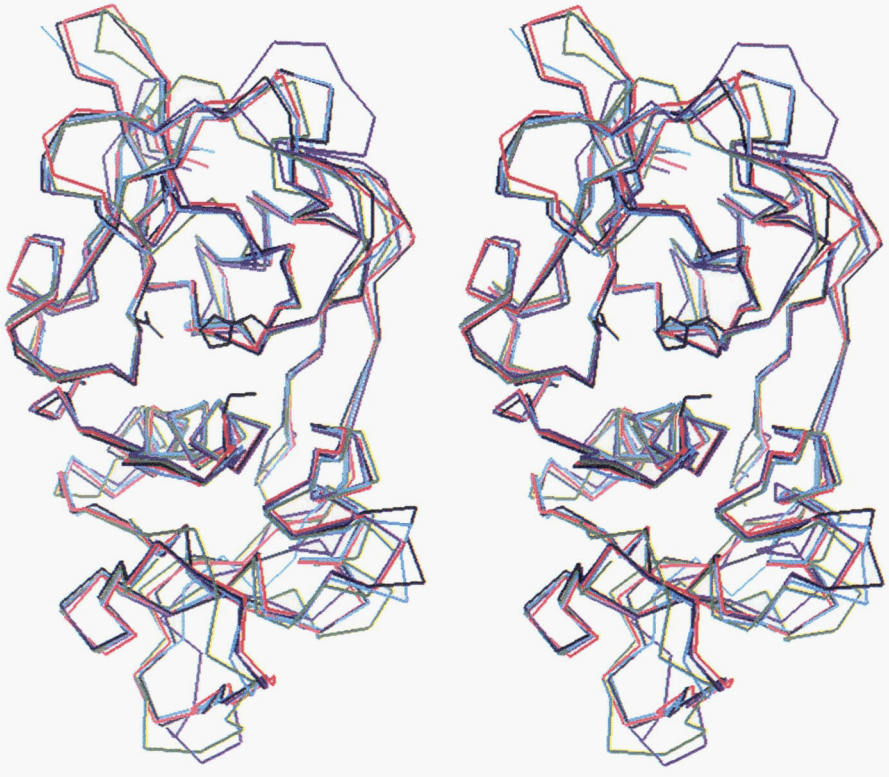

Fig. 2. Traces through the alpha carbon positions of cathepsin S and related enzymes are shown in stereo. The proteases were superimposed using a group of 90 core residues and are colored as follows: cathepsin S, black; cathepsin K (1mem (McGrath et al., 1997)), red; mature region of procathepsin L (Coulombe et al., 1996), blue; papain (Kamphuis et al., 1984), purple; cruzain (McGrath et al., 1995), green. Side chains for the catalytic triad residues Cys 25, His 159, and Asn 175 of cathepsin S are shown.

and oxyanion hole moiety Gln 19 Ne2 (3.2 Å), while the other lies 4.0 Å from Gln 19 Oε1, and 3.9 Å from Gly 23 Cα. The attached phenyl ring occupies S1′, which is comprised of Trp 177, His 159, Asn 158, and Ala 136. Also contributing to S1′ is Arg 137 (Fig. 3). In our structure, the electron density is not sufficient to fit the Arg 137 side-chain atoms beyond the Cβ. However, examination of commonly observed arginine rotomers (Ponder & Richards, 1987) shows that the guanidinium group could easily come within hydrogen bonding distance of a P1′ residue. As such, it could greatly affect the specificity at S1′, for both electrostatic and steric reasons. Comparison with related proteases shows that although the 136–137 region is somewhat variable in this family and can be

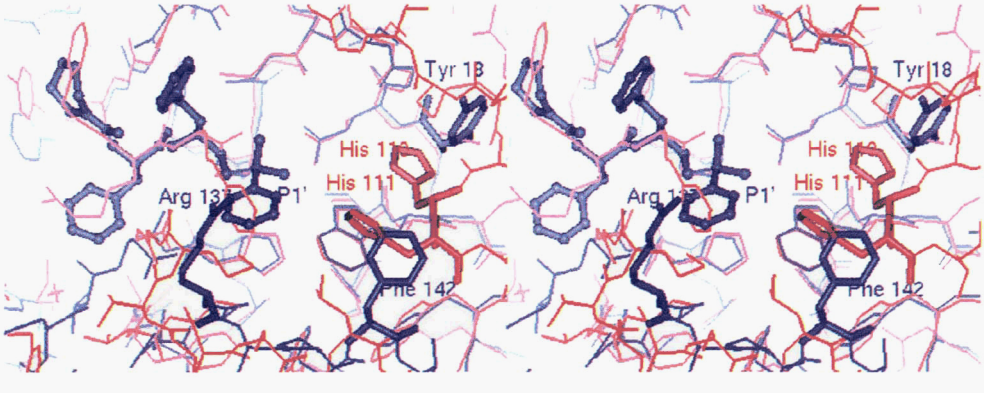

Fig. 3. Comparison (in stereo) of the prime side binding determinants of cathepsin S (blue) and cathepsin B (red) (Jia et al., 1995). APC 2848, as it appears at the cathepsin S active site, is rendered as balls-and-sticks, and the P1′ phenyl group is labeled. Immediately to the left of and below P1′, Arg 137, a component of the cathepsin SS1′ pocket is shown. The intended P1 residue of a cathepsin B complex with benzyloxycarbonyl-Arg-Ser (O-Bzl) is seen bound in the region of S1′. Also shown are the important prime side binding determinants for cathepsin B, His 110, and His 111. They are proximal to two residues, Tyr 18 and Phe 142, respectively, which are found in cathepsin S, but not the related enzymes.

the site of insertions or deletions of one or two residues, Arg at position 137 is quite unusual. Papain, cruzain, and cathepsins L, K, and B have Ala, Ser, Gly, Ser, and Tyr, respectively. In cathepsin K, the Ser side chain faces away from S1′, while in cruzain deletion of two residues in this area places Ser 137 far from S1′. The four χ angles accessible with Arg 137 results in an array of possibilities for interaction with P1′.

Another ramification of Arg at position 137 may be an effect on the electrostatic potential in the region of the catalytic triad. Examination of the superimposed structures of cathepsins S, B, L, and K along with those of papain and cruzain show a distinct absence of positive charge in the vicinity of the catalytic triad. The overall charge on these molecules varies: while the majority of cysteine proteases are negatively charged, cathepsin S has a +1 charge at neutral pH and cathepsin K has a net positive charge of 6. However, it is clear that the charge differences of these molecules overwhelmingly cluster in the back of the molecule and that there is conservation of charge at the active site face.

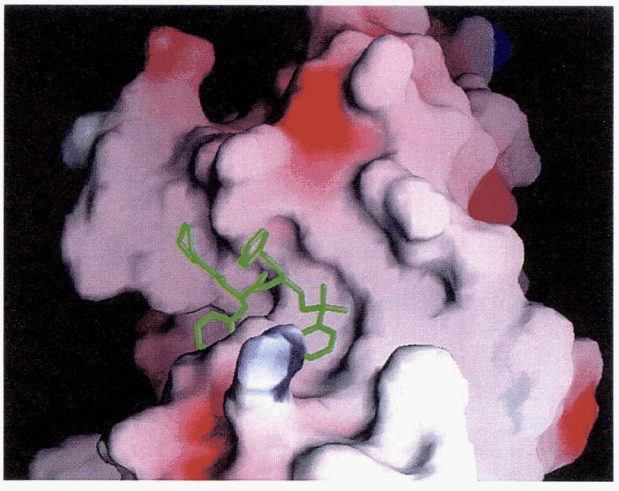

Fig. 4. GRASP (Nicholls et al., 1991) representation of the cathepsin S: APC 2848 complex. The inhibitor is shown in green and the electrostatic potential for cathepsin S is contoured from −16.0 kT/e (red) to 8.0 kT/e (blue). The positive potential corresponding to Arg 137 at the S1′ site can be seen near the bottom center of the figure, where it approaches the P1′ phenyl of APC 2848.

Nucleophilicity of Cys 25

Easily accessible Arg rotomers place the guanidinium group within 10 Å of Cys 25 Sg. This Cys at the papain-like protease active site contributes the acidic half of an ion pair with His 159, as mentioned above. The pK_a of Cys 25 has been measured at 4.5 (Rullmann et al., 1989), which greatly enhances its nucleophilicity. Though the unique electrostatic environment leading to this altered behavior is not well understood, the presence of the Cys at the N-terminus of a long, buried helix, is certainly contributory. This motif is also seen at the active site of thioredoxin family proteins where a nucleophilic cysteine is critical for activity (Gan et al., 1990). A recent study (Kortemme & Creighton, 1995) used model alpha helical peptides to quantify ionization of terminal cysteines, and showed that the alpha helix dipole interacts with the thiolate anion resulting in a significant pK_a drop even in the context of nonshielded model peptides. Small variations in the peptides were shown to have significant effects on the ionization. Thus, Arg 137, by introducing positive charge (Fig. 4), may further modulate the local electrostatics at the cathepsin S active site, and play a role in the enhanced reactivity of this enzyme toward certain inhibitor classes. It may also bear some responsibility for the unusual potency of cathepsin S at neutral pH.

Additional prime side determinants

Phe 142 lies approximately 5–6 Å from the bound P1′ phenyl ring and the conserved Trp 177. Though found in the bovine, rat, and human cathepsin S's, Phe is unusual at this site in the papain family. The side chain of Phe 142 closely approaches the position seen for the important His 111 (Musil et al., 1991) in cathepsin B (Fig. 3), and may provide shielding or π bonding opportunities for bound P1′ and P2′ residues. Phe 142 may also be important with regard to other conserved aromatic residues in the area. For example, cathepsin S is unusual in having Phe and not Trp at position 181. An aromatic residue at position 181 can interact with Trp 177 through the energetically favorable (Burley & Petsko, 1985) perpendicular juxtaposition of the aromatic rings. Earlier protein engineering studies sought to dissect out the contribution of conserved prime side aromatic residues in the context of cathepsin S, and showed that replacing Phe 181 with Trp had minimal impact (Brömme et al., 1996a). The presence of Phe 142 only in cathepsin S, where it lies 5.1 Å from Trp 177, and 6.5 Å from Phe 181, may be part of the reason the latter residue is not conserved as Trp in this enzyme.

Proximal to the position of the other prime-side histidine in cathepsin B (His 110) is Tyr 18 of cathepsin S (Fig. 3). This residue is usually an Asp or Asn in the cathepsins. Again, it is found in the three sequenced cathepsin S's, and may affect specificity C-terminal to the scissile bond.

S1 pocket

The S1 pocket of cathepsin S is occupied by homophenylalanine in this structure. This binding site in the papain-like cysteine proteases is, at best, a groove in the wall of the left domain. It is comprised of main-chain atoms of Asn 64 and Gly 65, which are 3.5–4.3 Å from the ring and Cβ2 of hPhe, and Cys 63 and Gly 23, whose carbonyl oxygens make van der Waals contacts with the ring moiety, and Cβ1, respectively. In general, this family of enzymes prefers Ala, Arg, and hPhe at P1. Examination of this structure, and that of a similar cathepsin K complex (McGrath et al., 1997), indicates that beta- or gamma-branched side chains would not work well at P1 because of steric clashes with Gly 23. The distinct lack of binding determinants in S1 makes the P1 site an unlikely candidate for design of specific inhibitors that target one cathepsin, and not another. Inhibitors in the vinyl sulfone series do offer a realistic view of the P1–S1 interaction, though, since binding is substrate-like and not out of register by one carbon. A shifted binding mode can lead to anomalous results such as seen in a cathepsin B-inhibitor structure where the intended P1 residue has collapsed into S1′ (Jia et al., 1995) (Fig. 3). Interestingly, the electron density for hPhe in this cathepsin S structure is especially strong, and the temperature factors quite low (average B for P1 side-chain atoms = 8.0 Å^2 vs. 24.2 Å^2 for all cathepsin S atoms) compared to a related cathepsin K structure (McGrath et al., 1997). This suggests that subtle differences in the cathepsin S S1 binding

pocket, which are not visible in this 2.5 Å structure, result in better binding of P1 hPhe.

S2 pocket

The S2 pocket of cathepsin S is comprised of Phe 67, Met 68, Gly 133, Val 157, Gly 160, and Phe 205. In this cathepsin S complex with APC 2848, the P2 Phe is bound at S2. The S2–P2 interaction is considered the primary determinant of specificity in the papain-like enzymes, and, as such, has received the most study. Previous experiments have shown that the S2 site of cathepsin S can be engineered to mimic those of cathepsin B or L, with the appropriately modified substrate preferences (Brömme et al., 1994). Comparison of the activity of a series of vinyl sulfone inhibitors against cathepsins S, K, and L indicated that cathepsin S has a broader specificity, which may be derived from a larger S2 pocket (Brömme et al., 1996b). Examination of the S2 pocket in this crystal structure shows that Phe 67, Met 68, and Val 157 occupy positions equivalent to those found for the same or similar residues in the related enzymes. Phe 205 forms the base of the S2 pocket and is what steers the specificity of cathepsin S toward branched hydrophobic side chains. The presence of Gly 160, and especially the unusual Gly 133, though, probably account for the more capacious binding site. The pose of residue 133 is such that, if there is a side chain, it points directly at bound P2. Thus, the orientation of the P2 Phe of APC 2848 places Cϵ2 4.0 Å from Gly 133 Cα, and would be only 3.0 Å from Ala 133 Cβ in any of a number of related enzymes. It appears that the extra room in the cathepsin S S2 pocket allows substrate side chains to adopt a number of conformations and still achieve productive binding.

S3 and additional interactions

Phe 67 and Lys 61 form the S3 binding site in cathepsin S. Substrate or inhibitor P3 is held in a pincer-like grasp between these residues. In this structure, the side-chain density of Lys 61 is insufficient for assignment of the positions of atoms beyond Cβ, though it is clear that this residue is oriented to interact with P3. Lys 61 is unusual in this family of enzymes, and may provide an opportunity for the design of specific cathepsin S inhibitors.

Binding of APC 2848 is also stabilized by a number of main-chain hydrogen bonds. As seen for several other complexes, bonds are found between Gly 66 N and the P2 carbonyl oxygen, Gly 66 O and the P2 amide nitrogen, and Asn 158 O and the P1 amide nitrogen.

Ordered solvent

Examination of the water structure in the highly related enzymes cathepsin K and procathepsin L allowed identification of conserved waters that were likely to be present in cathepsin S. Twenty-seven such molecules were modeled into the cathepsin S structure. They all were capable of one or more hydrogen bonds to solute atoms. After refinement, two waters were deleted due to a high temperature factor (>65.0 Å^2). While the effect of each individual water molecule was not determined, inclusion of the group of 25 reduced R_{free} by greater than 1%.

Conclusions

The atomic resolution structure of cathepsin S bound to a potent, irreversible inhibitor provides a wealth of information on the general topology, and on the specificity determinants of this cysteine protease. APC 2848 maps out four substrate binding sites and has facilitated the identification of structural features peculiar to cathepsin S. These idiosyncracies are consistent with the distinct catalytic behavior of cathepsin S, and with substrate specificity profiles. Moreover, the structure provides a sound basis for the design of potent and selective inhibitors against this therapeutically relevant enzyme.

Materials and methods

Crystallization and data collection

Human cathepsin S was expressed in *Sf*9 cells using a baculovirus system, purified, and inhibited with Mu-Phe-hPhe-vinylsulfone-phenyl (APC 2848) as previously described (Brömme & McGrath, 1996). The previously characterized ammonium phosphate crystal form was not used to solve the structure due to problems with twinning and secondary nucleation events. A second crystal form was obtained by vapor diffusion at 17 °C, where the reservoir contained 1.25 M ammonium sulfate, 70 mM MES, pH 6.5, and 7 mM $CoCl_2$. Two weeks after setup, the drops were streak-seeded from a drop containing tiny crystals obtained under similar conditions. Over the next six months, the seeds were maintained but failed to grow. The ammonium sulfate concentration in the reservoir was then increased to 2.2 M, and the drops were left undisturbed for almost ten months. At this point accretion of one seed crystal was sufficient to allow data collection.

The new crystal form proved to be rhombohedral, space group R3 (#146), with a = 107.61 Å and c = 105.19 Å. There are two molecules in the asymmetric unit, with V_M = 2.49 Å/Da (Matthews, 1968) and a solvent content of 51%. Data were collected (at 25 °C) using a R-AXIS IV image plate system (MSC, The Woodlands, Texas) equipped with focusing mirrors and using a 0.001 in. nickel filter. The generator was powered at 50 kV × 100 mA. The Biotex (MSC, The Woodlands, Texas) and DENZO (Otwinowski & Minor, 1993) packages were used for data reduction and space group determination.

Structure solution

Molecular replacement methods were used to obtain initial phases. The search model was a homology model of cathepsin S (MEM, unpubl. results) constructed from the related enzymes papain and cruzain. Similarity of cathepsin S to these and additional related proteases can be assessed by consulting a very useful alignment of 48 papain-like sequences (Berti & Storer, 1995). The molecular replacement experiments were done using the Xsight program package (MSI, San Diego, California). Self-rotation functions, in which the κ angle was held to 180° or 120°, were used to determine whether there were two or three molecules in the asymmetric unit. Once this was determined, a cross rotation search with the appropriate locked local symmetry was successful and translation searches were then used to correctly place the molecules in the cell. When the structure of procathepsin L became available (Coulombe et al., 1996), the mature region of this enzyme was then used as a more accurate basis for model-building. The model was mostly comprised of segments of cathepsin L, with residues 154–157 donated from actinidin. The molecular replacement results were confirmed by use of the new model with AMoRe (Navaza, 1994).

Density modification and refinement

The refinement of cathepsin S was carried out by alternating cycles of automated refinement with cycles of manual rebuilding of the model. All of the automated refinement was carried out using the X-PLOR v. 3.851 software package (Brünger et al., 1987) (MSI, San Diego, California). During the early stages of refinement, bond distances and angles were kept fixed, and only the torsional angles were adjusted. Toward the end of the refinement, the "slow-cool" protocol was employed. Through the entire process, the two monomers in the asymmetric unit were forced to behave identically, using the "ncs-strict" option in X-PLOR.

The electron density maps used for the manual refitting were made using various programs in the CCP4 suite (CCP4, 1994). Sigma-weighted $2F_o - F_c$ and $F_o - F_c$ maps were used, as well as maps that had been improved using solvent flattening, histogram matching and noncrystallographic symmetry averaging (Cowtan & Main, 1996).

Cathepsin S refinement was monitored by setting aside approximately 10% of the data (~1,000 reflections) for the calculation of a free *R*-factor (Brünger, 1993). Simulated annealing omit maps were used to assess the validity of specific regions of the model. In order to facilitate comparisons, cathepsin S was superimposed with related enzymes (Fig. 2) using a group of 90 core residues as follows: 14–39, 45–54, 66–68, 82–89, 108–113, 129–135, 157–167, 170–177, 186–190, 205–210. The quality of the electron density maps was insufficient for the placement of ordered water molecules. Therefore, the structures of cathepsin K and procathepsin L were examined in order to identify conserved, tightly-bound solvent molecules which were likely to be present in cathepsin S.

A Ramachandran plot (data not shown) of the refined model indicates four nonglycine residues that lie outside of the favored regions of the diagram. The phi/psi occupancy is 77%, similar to that found for related cysteine proteases, such as cathepsin K (76.5%).

Acknowledgments

We thank Eleanor Dodson, Kevin Cowtan, Roger Williams, and Victor Lamzin for advice and Michael C. Venuti for support.

References

Berti PJ, Storer AC. 1995. Alignment/phylogeny of the papain superfamily of cysteine proteases. *J Mol Biol 246*:273–283.

Brömme D, Bonneau PR, Lachance P, Storer AC. 1994. Engineering the S2 subsite specificity of human cathepsin S to a cathepsin L- and cathepsin B-like specificity. *J Biol Chem 269*:30238–30242.

Brömme D, Bonneau PR, Lachance P, Wiederanders B, Kirschke H, Peters C, Thomas DY, Storer AC, Vernet T. 1993. Functional expression of human cathepsin S in *Saccharomyces cerevisiae*. *J Biol Chem 268*:4832–4838.

Brömme D, Bonneau PR, Purisma E, Lachance P, Hajnik S, Thomas DY, Storer, A C. 1996a. Contribution to activity of histidine-aromatic, amide-aromatic, and aromatic-aromatic interactions in the extended catalytic site of cysteine proteinases. *Biochem 35*:3970–3979.

Brömme D, Klaus JL, Okamoto K, Rasnick D, Palmer JT. 1996b. Peptidyl vinyl sulphones: A new class of potent and selective cysteine protease inhibitors: S2P2 specificity of human cathepsin O2 in comparison with cathepsins S and L. *Biochem J 315*:85–89.

Brömme D, McGrath ME. 1996. High level expression and crystallization of recombinant human cathepsin S. *Protein Sci 5*:789–791.

Brünger AT. 1993. Assessment of phase accuracy by cross validation: The free *R* value—Methods and applications. *Acta Crystallogr D 49*:24–36.

Brünger AT, Kuriyan J, Karplus M. 1987. Crystallographic *R* factor refinement by molecular dynamics. *Science 235*:458–460.

Brünger AT, Rice LM. 1997. Crystallographic refinement by simulated annealing: Methods and applications. *Methods in enzymology*. Vol 277. New York: Academic Press. pp 243–269.

Burley SK, Petsko GA. 1985. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. *Science 229*:23–28.

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

Cowtan KD, Main P. 1996. Phase combination and cross validation in iterated density-modification calculations. *Acta Cryst D 52*:43–48.

Cygler M, Sivaraman J, Grochulski P, Coulomb 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.

Drenth J, Kalk KH, Swen HM. 1976. Binding of chloromethyl ketone substrate analogues to crystalline papain. *Biochem 15*:3731–3738.

Fujishima A, Imai Y, Nomura T, Fujisawa Y, Yamamoto, Y, Sugawara T. 1997. The crystal structure of human cathepsin L complexed with E-64. *FEBS Lett 407*:47–50.

Gan Z-R, Sardana MK, Jacobs JW, Polokoff MA. 1990. Yeast thioltransferase: The active site cysteines display differential reactivity. *Arch Biochem Biophys 282*:110–115.

Gunčar G, Podobnik M, Pungerčar J, Štrukelj B, Turk V, Turk D. 1997. Crystal structure of porcine cathepsin H determined at 2.1 Å resolution: Location of the minichain C-terminal carboxyl group defines cathepsin H aminopeptidase function. *Structure 6*:51–61.

James MNG. 1994. Serine proteinases and the convergence of active site geometries among the four classes of proteolytic enzymes. *Keystone symposium on the structural and molecular biology of protease function and inhibition*. Santa Fe, New Mexico, 1994. *J. Cellular Biochemistry*, Suppl. 18D. New York: Wiley-Liss.

Jia Z, Hasnain S, Hirama T, Lee X, Mort JS, To R, Huber CP. 1995. Crystal structures of recombinant rat cathepsin B and A cathepsin B-inhibitor complex. *J Biol Chem 270*:5527–5533.

Kamphuis IG, Kalk KH, Swarte MBA, Drenth J. 1984. Structure of papain refined at 1.65Å resolution. *J Mol Biol 179*:233–256.

Kirschke H, Wiederanders B, Brömme D, Rinne A. 1989. Cathepsin S from bovine spleen: Purification, distribution, intracellular localization, and action on proteins. *Biochem J 264*:467–473.

Kortemme T, Creighton TE. 1995. Ionisation of cysteine residues at the termini of model alpha-helical peptides. Relevance to unusual thiol pka values in proteins of the thioredoxin family. *J Mol Biol 253*:799–812.

Matthews BW. 1968. Solvent content of protein crystals. *J Mol Biol 33*:491–497.

McGrath ME, Eakin AE, Engel J, McKerrow JH, Craik CS, Fletterick RJ. 1995. The crystal structure of cruzain: A therapeutic target for Chagas' disease. *J Mol Biol 247*:251–259.

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

Musil D, Zucic D, Turk D, Engh RA, Mayr I, Huber R, Popovic T, Turk V, Towatari T, Katunuma N. 1991. The refined 2.15Å X-ray crystal structure of human liver cathepsin B: The structural basis for its specificity. *EMBO J 10*:2321–2330.

Nakagawa T, Roth W, Wong P, Nelson A, Farr A, Deussing J, Villadangos JA, Ploegh H, Peters C, Rudensky AY. 1998. Cathepsin L: Critical role in Ii degradation and CD4 T cell selection in the thymus. *Science 280*:450–453.

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

Nicholls A, Sharp KA, Honig B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins 11*:281–296.

Otwinowski Z, Minor W. 1993. Data collecting and processing. *Data collecting and processing*. Warrington, UK: SERC Daresbury Laboratory.

Palmer JT, Rasnick D, Klaus L. 1995. Vinyl sulfones as mechanism-based cysteine protease inhibitors. *J Med Chem 38*:3193–3196.

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.

Polgar L. 1974. Mercaptide-imidazolium ion-pair: The reactive nucleophile in papain catalysis. *FEBS Lett 47*:15–18.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.

Qian F, Chan SJ, Gong Q, Bajkowski AS, Steiner DF, Frankfater A. 1991. The expression of cathepsin B and other lysosomal cysteine proteases in normal tissues and in tumors. *Biomed Biochim Acta 50*:4–6.

Rice LM, Brünger AT. 1994. Torsion angle dynamics: Reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins Struct Funct Genet 19*:277–290.

Riese RJ, Mitchell RN, Villadangos JA, Karp ER, DeSanctis GT, Ploegh HL, Chapman HA. 1998. Cathepsin S activity regulates antigen presentation and immunity. *J Clin Invest*. In press.

Riese RJ, Wolf PR, Bromme D, Natkin LR, Villadangos JA, Ploegh HL, Chapman HA. 1996. Essential role for cathepsin S in MHC class II-associated invariant chain processing and peptide loading. *Immunity 4*:357–366.

Rullmann JAC, Bellido MN, van Duijnen PT. 1989. The active site of papain: All-atom study of interactions with protein matrix and solvent. *J Mol Biol 206*:101–118.

Schechter I, Berger A. 1967. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun 27*:157–162.

Shi G-P, Webb AC, Foster KE, Knoll JH, Lemere CA, Munger JS, Chapman HA. 1994. Human cathepsin S: Chromosomal localization, gene structure, and tissue distribution. *J Biol Chem 269*:11530–11536.

Tezuka K, Tezuka Y, Maejima A, Sato T, Nemoto K, Kamioka H, Hakeda Y, Kumegawa M. 1994. Molecular cloning of a possible cysteine proteinase predominantly expressed in osteoclasts. *J Biol Chem 269*:1106–1109.

Turk D, Podobnik M, Popovic T, Katunuma N, Bode W, Huber R, Turk V. 1995. Crystal structure of cathepsin B inhibited with CA030 at 2.1 Å resolution: A basis for the design of specific epoxysuccinyl inhibitors. *Biochem 34*:4791–4797.

Turnsek T, Kregar I, Lebez D. 1975. Acid sulfhydryl protease from calf lymph nodes. *Biochim Biophys Acta 403*:514–520.

Villadangos JA, Riese RJ, Peters C, Chapman HA, Ploegh HL. 1997. Degradation of mouse invariant chain: Roles of cathepsins S and D and the influence of major histocompatibility complex polymorphism. *J Exp Med 186*:549–560.

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