Crystal Structure of Human Cathepsin V[†,‡]

John R. Somoza,* Hangjun Zhan, Krista K. Bowman, Lily Yu, Kyle D. Mortara, James T. Palmer, James M. Clark, and Mary E. McGrath

Axys Pharmaceuticals, Inc., 385 Oyster Point Boulevard, Suite 1, South San Francisco, California 94080

Received April 26, 2000

ABSTRACT: Cathepsin V is a lysosomal cysteine protease that is expressed in the thymus, testis and corneal epithelium. We have determined the 1.6 Å resolution crystal structure of human cathepsin V associated with an irreversible vinyl sulfone inhibitor. The fold of this enzyme is similar to the fold adopted by other members of the papain superfamily of cysteine proteases. This study provides a framework for understanding the structural basis for cathepsin V's activity and will aid in the design of inhibitors of this enzyme. A comparison of cathepsin V's active site with the active sites of related proteases revealed a number of differences, especially in the S2 and S3 subsites, that could be exploited in identifying specific cathepsin V inhibitors or in identifying inhibitors of other cysteine proteases that would be selective against cathepsin V.

The cathepsins are proteases that are active at low pH and are primarily responsible for protein degradation in the lysosome. In mammals, the cathepsins play important roles in a variety of cellular processes, and defects in the proper regulation of these enzymes have been linked to diseases such as osteoporosis, cancer, emphysema, asthma and glomerulonephritis. The relation between these diseases and the cathepsins has led to the targeting of these proteases in a number of drug development efforts. For example, cathepsin S is pursued as an inflammation target and cathepsin K as an osteoporosis target, and cathepsins B and L are potential oncology targets.

Cathepsin V (also known as cathepsin L2) was recently cloned by three independent research groups (*1*−*3*). Its sequence shows that it is a member of the papain family of cysteine proteases (*4*), and several pieces of evidence demonstrate an especially close link between this enzyme and cathepsin L. The sequences of these two enzymes are quite similar, with a sequence identity of 77% for the proenzyme and 80% for the mature enzyme. In addition, the position and chromosomal organization of the genes coding for these two cathepsins are similar (*3*, *5*). The cathepsin V gene has been mapped to a region near the cathepsin L locus, and both genes consist of eight exons that code for analogous regions of the cathepsins V and L structures. Overall, the similarities in sequence and in genomic organization suggest that cathepsins V and L diverged late in evolution.

Despite the similarities between cathepsins V and L, these two proteases differ both in substrate specificity (*3*) and in tissue distribution (*1*, *2*). An examination of the tissue distribution of cathepsin V has shown that it is specifically expressed in the thymus, testis and corneal epithelium (*1*−*3*). This limited distribution contrasts with that of cathepsin L, which is expressed throughout the body. Santamaria et al. (*1*) also reported that cathepsin V is expressed in colorectal and breast carcinomas but not in normal colon or mammary tissue, and Schilling and Ahlquist (*6*) showed that cathepsin V is expressed at a higher level in adenomas and adenocarcinomas of the colorectum than in normal mucosa. This increase in expression suggests that monitoring cathepsin V levels may provide a way of tracking the progression of colorectal carcinomas. It is also possible that cathepsin V plays a part in cancer progression and that it might be an oncology target.

Although the nonpathological role of cathepsin V is not clear, there is evidence pointing toward a possible role in T-cell selection. In the mouse, the protein that is most closely related to human cathepsin V is mouse cathepsin L. The sequence of mouse cathepsin L places it between human cathepsin L and human cathepsin V, and some of the surface properties of mouse cathepsin L more closely resemble human cathepsin V than human cathepsin L (*3*). Brömme et al. (*3*) have suggested that mouse cathepsin L plays the roles of both human cathepsins. Nakagawa et al. (*7*) have shown that, in the mouse, cathepsin L is an important part of $CD4^+$ T-cell selection in the thymus. This finding, coupled with the similarity between mouse cathepsin L and human cathepsin V and the observation that cathepsin V is specifically expressed in the thymus, suggests that cathepsin V may be involved in T-cell selection in humans.

The structure of cathepsin V will be invaluable in the search for potent and selective inhibitors of this enzyme. These inhibitors would be useful in exploring the in vivo role of cathepsin V, could possibly be used as a way of controlling T-cell selection, and might eventually be useful in a drug design effort should this protease become a

[†] This work is based upon research conducted at the Stanford Synchrotron Radiation Laboratory (SSRL), which is funded by the Department of Energy (BES, BER) and the National Institutes of Health (NCRR, NIGMS).

[‡] The coordinates and observed structure factor amplitudes for human cathepsin V have been deposited in the Protein Data Bank under accession number 1FHO.

* To whom correspondence should be addressed: phone (650) 829-1014; fax (650) 829-1123; e-mail john_somoza@axyspharm.com.

10.1021/bi000951p CCC: $19.00
Published on Web 09/21/2000

therapeutic target. However, the most immediate interest in the cathepsin V structure is due to the protease's role as an anti-target. There are a number of drug design programs that focus on other closely related human lysosomal cysteine proteases. In designing inhibitors that are specific for one of these other enzymes, it will be helpful to be able to pinpoint specific structural differences between the active site of cathepsin V and the active sites of these related cathepsins.

EXPERIMENTAL PROCEDURES

Synthesis of APC-3316 (4-Methylpiperazine-1-carboxylic acid [1-[(3-benzenesulfonyl-1-phenethylallyl]carbamoyl)-2-phenylethyl]amide or MePip-Phe-HphVSPh). Throughout this section, NMR chemical shifts are reported in parts per million (ppm) with the following abbreviations: s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet (m* represents a multiplet signal that cannot be spectroscopically separated from its neighbors).

To a solution of (*S*)-BOC-homophenylalanine (**1**) (Synthetech; 53.63 g, 192 mmol) in CH_2Cl_2 (400 mL) at 0 °C were added *N,O*-dimethylhydroxylamine hydrochloride (Aldrich: 18.73 g, 193 mmol), diisopropylethylamine (Aldrich; 33.4 mL, 192 mmol), and diisopropylcarbodiimide (Aldrich; 30.1 mL, 192 mmol). The mixture was stirred for 2 h while warming to room temperature and was diluted with an equal volume of ethyl acetate. The suspension was filtered, concentrated to a volume of approximately 400 mL, washed with 400 mL each of 1 M HCl, saturated aqueous $NaHCO_3$, and brine, dried over $MgSO_4$, filtered, and concentrated. The residue was suspended in hot hexane (300 mL), filtered to remove any residual diisopropylurea, and concentrated to dryness, yielding 60.63 g (98%) of (*S*)-*tert*-butyl [1-(methoxymethylcarbamoyl)-3-phenylpropyl]carbamate [BOC-Hph-N(Me)OMe, **2]**. This material was used in the next step without further purification.

To a solution of **2** (60.63 g, 188 mmol) in THF (300 mL) was added a 1.0 M solution of $LiAlH_4$ in THF (Aldrich; 188 mL) over 10 min, with vigorous stirring. The mixture was stirred for an additional 20 min, whereupon water was carefully added until effervescence ceased (approximately 50 mL). HCl (1 M, 900 mL) was added to break up the gelatinous emulsion. The product was extracted with ethyl acetate (600 mL), which was washed with 500 mL each of saturated aqueous $NaHCO_3$ and brine, dried over $MgSO_4$, filtered, and evaporated to dryness, giving 49.9 g (100%) of (*S*)-*tert*-butyl (1-formyl-3-phenylpropyl)carbamate [BOC-homophenylalaninal (Boc-HphH, **3**)] (*8*). This material was used immediately in step 4 without further purification.

To a solution of diethyl phenylthiomethylphosphonate (**4**) (Lancaster Synthesis; 50 g, 192 mmol) in CH_2Cl_2 (300 mL) at 0 °C was added peracetic acid (Fluka; 72 mL of a 39% dilute acetic acid solution, 2.2 equiv) over 1 h. The mixture was stirred overnight while warming to room temperature. The solution was then cooled to 0 °C and was treated with 50 mL of a freshly prepared 10% sodium bisulfite solution, added dropwise. NaOH (2 M, 500 mL) was then added over 20 min to increase the pH to >12. The organic phase was separated, dried over $MgSO_4$, filtered, and concentrated to dryness, giving 56.1 g (100%) of diethyl benzenesulfonylmethyl-phosphonate (**5**). ^{1}H NMR ($CDCl_3$) 1.25 (6H, t, J = 7 Hz); 3.74 [2H, d, J(P−C−H) = 17 Hz]; 4.09 (4H, q, J = 7 Hz); 7.56−7.66 (3H, m); 7.96−7.99 (2H, m).

Sodium hydride (Aldrich; 8.45 g of a 60% mineral oil dispersion, 211 mmol) was added to a solution of **5** (56.12 g, 192 mmol) in THF (600 mL) at room temperature with rapid stirring. The solution was stirred until effervescence had ceased, resulting in a clear pale-yellow solution. A solution of **3** (49.9 g, 188 mmol) in THF (150 mL, 2 × 25 mL wash) was added quickly. The mixture was stirred for 15 min. HCl (1 M, 500 mL) was added. The product was extracted with ethyl acetate (600 mL), washed with 500 mL each of saturated aqueous $NaHCO_3$ and brine, dried over $MgSO_4$, filtered, and evaporated to dryness. The product, (*S*)-*tert*-butyl (3-benzenesulfonyl-1-phenethylallyl)carbamate (BOC-HphVSPh, **6**) was used in step 5 without further purification, assuming a quantitative yield; however, crystallization of a small amount of the product from CH_2Cl_2/ether/hexane could be effected for analysis. ^{1}H NMR ($CDCl_3$) 1.32 (9H, s), 1.73−1.95 (2H, m), 2.67 (2H, m), 4.35 (1H, m), 4.53 (br d), 6.43 (1H, dd, J = 15 and 1.5 Hz), 6.87 (1H, dd, J = 15 and 5 Hz), 7.1−7.3 (5H, m), 7.48−7.65 (3H, m), 7.83−7.87 (2H, m).

p-Toluenesulfonic acid monohydrate (Aldrich; 130 g, 0.683 mol) was azeotropically dried by repeated coevaporation of a 3.75:1 toluene/2-propanol solution (2 × 950 mL). A portion of this material (81.5 g, 0.473 mol) was dissolved in ether (200 mL). To this solution was added a solution of crude **6** (81.5 g, 203 mmol maximum from two separate batches processed through step 4) in ether (100 mL). The mixture was stirred at room temperature overnight. Ether (1 L) was added. The solids were broken up, filtered, washed with 9:1 ether/ethyl acetate (2 × 1L) and then hexane (2 × 500 mL), and dried in vacuo to give 73.4 g (76% from **3**) of 3-benzenesulfonyl-1-phenethylallylamine tosylate (TsOH•HphVSPh, **7**) (*9*).

(*S*)-Benzyl-2-amino-3-phenylpropionate hydrochloride (HCl•PheOBzl, 8) (Advanced ChemTech; 56.03 g, 192 mmol) and pyridine (Aldrich; 62.12 mL, 0.768 mol) were suspended in CH_2Cl_2 (800 mL) at −10 °C. The mixture was stirred vigorously, whereupon a solution of phosgene in toluene (Fluka; 20%, 132 mL, 249.6 mmol) was added over 30 min. The solution was stirred for an additional 2 h at −10 °C and washed with ice-cold 0.5 M HCl (1.8 L), and the aqueous phase was back-extracted with CH_2Cl_2 (2 × 100 mL). The combined organics were washed with ice-cold brine (500 mL), and the aqueous phase was back-extracted with CH_2Cl_2 (100 mL). The combined organics were quickly filtered through a tall Buchner funnel loaded with $MgSO_4$ (bottom layer) and decolorizing charcoal (top layer). The filtrate was concentrated and crystallized from CH_2Cl_2/hexane to give 45.86 g (85%) of (*S*)-benzyl-2-isocyanato-3-phenylpropionate (OCN-PheOBzl, **9**) (*10*).

To a solution of **9** (45.86 g, 163 mmol) in THF (250 mL) at 0 °C was added *N*-methylpiperazine (Aldrich; 18.08 mL, 163 mmol). The mixture was stirred for 30 min and was concentrated in vacuo to a pale orange oil that crystallized on standing. This material [(*S*)-benzyl-2-[(4-methylpiperazine-1-carbonyl)amino]-3-phenylpropionate, MePip-PheOBzl, **10;** theoretical yield 62.19 g]) was used in the next step without further purification.

A solution of **10** (62.19 g, 163 mmol) in ethanol (400 mL) was charged with 10% palladium on active charcoal (Ald-

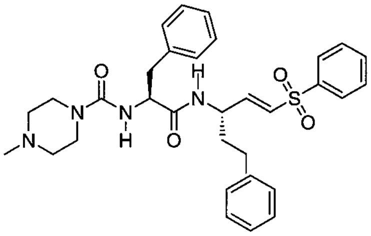

FIGURE 1: 4-Methylpiperazine-1-carboxylic acid [1-[(3-benzenesulfonyl-1-phenethylallyl)carbamoyl]-2-phenylethyl]amide (APC-3316).

rich: 4.5 g). The solution was exposed to hydrogen on a Parr shaker (30 psi) for 6 h, filtered through Celite, and concentrated in vacuo. The residue was triturated with ether (700 mL) to remove residual ethanol and was reprecipitated from CH_2Cl_2/ether to give 44.31 g (93%) of (*S*)-2-[(4-methylpiperazine-1-carbonyl)amino]-3-phenylpropionic acid (MePip-PheOH, **11**). ^{1}H NMR (DMSO-d_6) 2.15 (3H, s), 2.18 (4H, m), 2.84−2.94 (1H, dd, J = 15 and 11 Hz), 2.95−3.04 (1H, dd, J = 15 and 5 Hz), 3.14−3.30 (4H, m), 4.17 (1H, m), 6.65 (1H, d, J = 8 Hz), 7.17−7.27 (5H, m).

A 2-L, three-necked round-bottomed flask equipped with overhead stirrer, thermometer, and an addition funnel was charged with a solution of 11 (44.3 g, 152 mmol) in 3:1 THF/DMF (400 mL). The mixture was cooled to −10 °C, whereupon 4-methylmorpholine (Aldrich; 16.72 mL, 152 mmol) was added, followed by a solution of isobutyl chloroformate in THF (50 mL) over 20 min, with vigorous stirring. The temperature was maintained at or below −5 °C throughout the addition. The mixture was stirred for 1 h. Compound **7** (72.0 g, 152 mmol) was added as a solid. A solution of 4-methylmorpholine (16.72 mL, 152 mmol) was added dropwise, with careful temperature control maintained between −5 and −10 °C to minimize epimerization of the phenylalanine residue. The mixture was stirred for 1 h. Saturated aqueous $NaHCO_3$ (500 mL) was added. The mixture was extracted with ethyl acetate (600 mL), washed with brine, dried over $MgSO_4$, filtered, and concentrated. The thick oily residue was dissolved in acetonitrile (200 mL) and was added to a solution of *p*-toluenesulfonic acid monohydrate (28.91 g, 152 mmol) in acetonitrile (300 mL). The solution was allowed to stand overnight following addition of a small seed crystal. The product, (APC-3316 tosylate, **12**) was filtered, washed with ether (2 × 300 mL), air-dried, and recrystallized from hot acetonitrile to give 70.29 g (62%) of APC-3316 tosylate.

Conversion of APC-3316 tosylate to APC-3316 hydrochloride (MePip-Phe-HphVSPh•HCl) (Figure 1)was as follows: Acetonitrile (Burdick and Jackson; 800 mL) was added to 12 (173.4 g, 0.232 mol from multiple combined batches). The mixture was heated on a hot plate/stirrer until all the solids had dissolved. The solution was allowed to cool for 48 h. The solids were filtered, washed with acetonitrile (1 L) and ether (J. T. Baker), and air-dried. The pure tosylate was partitioned between saturated aqueous $NaHCO_3$ (1 L) and ethyl acetate (800 mL), washed with brine (500 mL), dried over $MgSO_4$, filtered, and concentrated to a total weight of 154 g (APC-3316 free base plus solvent). This material was dissolved in CH_2Cl_2 (400 mL) and was added dropwise, over 3 h, to a vigorously stirred solution of 1.0 M hydrogen chloride in ether (Aldrich, 460 mL) diluted with 4 L of dry ether. The solids were stirred for an additional 12 h, filtered, washed with ether (4 × 1 L) and hexane (Burdick and Jackson; 3 × 1 L), and dried to constant weight at 45 °C under vacuum. The yield of APC-3316 hydrochloride was 134 g (94% from the tosylate, 34% total yield from commercially available materials). ^{1}H NMR (DMSO-d_6) 1.8 (1H, m), 1.9 (1H, m), 2.49−2.9 (4H, m*), 2.7 (3H, d, J = 4 Hz), 2.95 (2H, m*), 3.04−3.26 (2H, m*), 3.35 (2H, d, J = 11 Hz), 4.03−4.17 (2H, m), 4.32 (1H, m), 4.48 (1H, m), 6.52 (1H, d, J = 15 Hz), 6.87 (1H, dd, J = 15 and 4 Hz), 7.13−7.31 (10H, m), 7.61−7.74 (3H, m), 7.85 (2H, d, J = 8 Hz), 8.44 (1H, d, J = 8 Hz), 11.30 (1H, br s). ^{13}C NMR (DMSO-d_6) 32.0, 35.3, 37.9. 41.3, 42.5, 49.3, 52.5, 52.6, 57.4, 126.4, 127.7, 128.6, 128.8, 129.0, 129.8, 130.2, 134.2, 138.8, 140.9, 141.8, 147.9, 157.3, 172.8. MS (M^{1+}) 575 (signal of free base).

Gene Cloning, Mutagenesis, and Expression of Human Procathepsin V. Human cathepsin V cDNA was cloned and expressed in *Pichia pastoris* by use of the pPIC9 expression vector as described by Brömme et al. (*3*). PCR-based site-directed mutagenesis was used to remove potential N-linked glycosylation sites at residues 108 and 179. The Asn-108 codon (AAT) was changed to glutamine (CAG) by use of the oligonucleotide primer 5′ CCT GAG AAT TCT GTT GCT AAT GAC ACT GGC TTC ACA G 3′ and its reverse complement. The Asn-179 codon (AAT) was changed to an aspartic acid (GAT) by use of the primer 5′ GC TTT GAA GGA GCA AAT TCG GAT AAC AGC AAG TAT TGG 3′ and its reverse complement. Oligonucleotides were obtained from Operon Technologies, Inc. ProcathepsinV/pPIC9 wild-type expression plasmid (5.0 ng) was mixed with 125 ng of one primer, its reverse complement, and 2.5 units of *Pfu* DNA polymerase (Stratagene). The mixture was subjected to 16 rounds of temperature cycling with the following sequence: 95 °C for 30 s, 55 °C for 60 s, and 68 °C for 20 min. *Dpn*I restriction enzyme was added to eliminate the template DNA. An aliquot of the resulting mixture was used to transform *Escherichia coli* XL1-Blue cells, and the presence of the mutation was confirmed by DNA sequencing (Perkin-Elmer ABI Prism dye terminator cycle sequencing ready reaction kit; ABI Prism 377 DNA sequencer). The second mutation was then created with the second primer set under the same conditions. The expression plasmid for nonglycosylated cathepsin V (N108Q, N179D) was then transformed into the KM71 strain of *P. pastoris*, and a clone expressing high levels of cathepsin V was selected for fermentation. During fermentation, methanol induction was initiated at an OD_{600} of 360, and harvesting was carried out 80 h later.

Purification and Acid Activation of Procathepsin V. Six liters of induction medium from expressing *P. pastoris* cells was collected, filtered through a 0.2 μm cellulose acetate filter, concentrated, and diafiltered into 300 mL of buffer A [50 mM sodium acetate (pH 4.5), 1 mM EDTA, and 0.5 mM DTT] through a 10 kDa cutoff membrane in a Filtron UF/DF apparatus. This solution was loaded onto a 16 mL Source 15S methyl sulfonate cation-exchange column (15 μm resin, Amersham Pharmacia Biotech) at 0.8 mL/min or at a linear rate of 38 cm/h. The column was then washed with 4 column volumes of buffer A and eluted with a gradient increase in NaCl in buffer A over 22 column volumes at 19 cm/h. The low pH in buffer A resulted in the autoactivation of cathepsin V, and homogeneous mature cathepsin V was obtained following this single step, with a trace amount of unactivated procathepsin V eluting at lower salt concentration. The peak fractions containing mature cathepsin V were pooled and the irreversible inhibitor APC-3316 was immediately added to prevent degradation resulting from autocatalysis. N-terminal sequencing of the activated enzyme was carried out by the Protein Structure Lab at the University of California, Davis. The purity, molecular weight, and microheterogeneity of the enzyme was evaluated by SDS−polyacrylamide gel electrophoresis and by electrospray mass spectrometry (University of Washington Mass Spectrometry Center).

Crystallization and Data Collection. Cathepsin V was crystallized by the hanging-drop vapor-diffusion method. Crystals were obtained under two crystallization conditions. The first crystals were obtained by equilibrating a reservoir consisting of 30% (w/v) poly(ethylene glycol) 4000, 0.2 M ammonium sulfate, and 0.1 M sodium citrate (pH 4.77) with a hanging drop consisting of 2 μL of the reservoir solution and 2 μL of a 12 mg/mL protein solution in 25 mM sodium acetate (pH 4.5). A second set of crystals was obtained by equilibrating a reservoir consisting of 0.17 M lithium sulfate, 0.085 M Tris (pH 8.5), 25.5% (w/v) poly(ethylene glycol) 4000, and 15% (v/v) glycerol with a drop consisting of 2 μL of the reservoir solution and 2 μL of a 12 mg/mL protein solution in 25 mM sodium acetate (pH 4.5). The two crystallization conditions led to the same crystal form. All of the crystals were grown at 17 °C.

All of the data that were used to determine the structure were collected on crystals that had been cooled to −180 °C. To cryocool the crystals that were grown without the glycerol, they were quickly dragged through a drop of paratone-N (Hampton Research, Inc.) and flash-cooled directly in the N_2 cold stream. The crystals that were grown in the presence of glycerol could be flash-cooled in the N_2 cold stream without any additional cryoprotectant.

Two data sets were used to determine the cathepsin V structure. First, a 2.3 Å resolution data set was collected from one of the crystals that had grown in the absence of glycerol. These data were recorded on an R-AXIS IV image plate detection system, using X-rays from a Rigaku RU-200 rotating anode generator operating at 50 kV/90 mA. These data were processed with the DENZO and SCALEPACK packages (*11*). The autoindexing routine in DENZO showed that the crystals belonged to the primitive hexagonal Bravais lattice, and the data scaling and examination for systematic absences showed that the space group was $P6_422$. The size of the unit cell suggested that there were two molecules in the asymmetric unit. This data set was used for the structure determination and initial refinement of the cathepsin V model. A second data set was collected from one of the crystals that had grown in the presence of glycerol. These data were collected at beamline 9-1 at the Stanford Synchrotron Radiation Laboratory on a MAR345 image plate detector (the statistics for these data are summarized in Table 1). This second data set was used for the final refinement of the cathepsin V model.

Table 1: Data and Refinement Statistics

Data	
space group	$P6_422$
cell parameters (Å)	$a = b = 104.7$, $c = 179.2$
resolution range (Å)	100.0−1.60
no. of observations	386,837
no. of unique reflections	78,054
completeness (%)	
100.0−1.60 Å	98.1
1.63−1.60 Å	89.9
$R_{sym}(I)^a$ (%)	
100.0−1.60 Å	3.8
1.63−1.60 Å	42.4
$\langle I\rangle/\langle\sigma(I)\rangle$	
100.0−1.60 Å	37.3
1.63−1.60 Å	3.45
Model	
molecules in asymmetric unit	2
no. of water molecules	287
no. of sulfates	2
resolution used for refinement (Å)	50.0−1.60
σ cutoff $[F/\sigma(F)]$	0.0
R-factor	0.197
free *R* (based on 7386 reflections)	0.211
RMS deviations from ideal geometry	
bonds (Å)	0.008
angles (deg)	1.18

[a] $R_{sym} = \Sigma|I - \langle I\rangle|/\Sigma\langle I\rangle$.

Structure Determination and Refinement. The positions of the two molecules in the asymmetric unit were determined by molecular replacement with the program EPMR (*12*) with a search model that consisted of the mature part of the procathepsin L structure (PDB accession code 1CS8) (*13*). The initial electron density maps showed the inhibitor covalently attached to the active-site cysteine and clearly showed areas of difference between cathepsins V and L. The refinement was carried out by iterating between manual rebuilding with Quanta (MSI, Inc.) and automated refinement with CNX 99.0 (MSI, Inc.) (*14*). During the initial refinement, the conformations of the two molecules in the asymmetric unit were constrained to be identical, and only torsional angle refinement was employed (*15*). After several rounds of refinement and after addition of the ligand to the model, the noncrystallographic symmetry constraint was removed and Cartesian refinement was initiated. Toward the end of the refinement, individual isotropic *B*-factors were refined and water molecules were added to the model. All measured data $[|F| > 0.0\sigma(F)]$ to 1.6 Å resolution were used during the final refinement, and a bulk solvent mask was employed. The progress of the refinement was monitored by calculating a free *R*-factor, based on 7386 reflections (approximately 9% of the total reflections) and by evaluating the model's deviation from ideal geometry, by use of CNX 99.0 and PROCHECK (*16*).

Analysis of the Sequence and Structure. Sequence comparisons between cathepsin V and related enzymes were carried out with the BLAST program (NCBI) (*17*). Structural alignments and comparisons were done with Insight II (MSI, Inc.). The assignment of secondary structural elements was carried out with the algorithm described by Kabsch and Sander (*18*), as implemented by the program DSSP. Figure 1 was prepared with CS ChemDraw Pro (CambridgeSoft Corp.); Figures 2, 3, 4A, and 5 were prepared with Molscript (*19*) or Bobscript (*20*), followed by Raster3D (*21*); and Figure 4B was prepared with InsightII (MSI, Inc.).

RESULTS AND DISCUSSION

Assessment of the Quality of the Cathepsin V Model. In general, the structure of cathepsin V is clearly defined by the electron density maps. The only exceptions to this are residues 168D through 168G of one of the monomers (in the papain numbering system), which form a surface loop, and a few of the long, hydrophilic side chains at the surface of the enzyme. The current model consists of all 221 residues for each monomer and a total of 287 water molecules. There is a strand of density, near the active site of the protease, that remains unaccounted for in the cathepsin V model. This density might correspond to a water channel, or it might correspond to a peptide fragment, possibly a remnant of the enzyme's pro region. The crystallographic *R*-factor for the current model is 19.7% (for all data between 50.0 and 1.6 Å resolution), with a free *R*-factor of 21.1%. The root-mean-square deviations from ideal bond lengths and angles are 0.008 Å and 1.18°, respectively. The ϕ and ψ angles for 85.9% of the residues are within the most favorable region of the Ramachandran plot, and no residues lie within the disallowed region. The statistics for the current model are summarized in Table 1.

The precision of the cathepsin V model can be assessed by comparing the two molecules that make up the asymmetric unit. During the initial refinement the two molecules were constrained to behave identically. However, a considerable amount of manual rebuilding and simulated annealing refinement was done once the noncrystallographic symmetry constraint was removed. Therefore, the two molecules in the asymmetric unit should represent reasonably independent views of the structure. A superposition of the two monomers shows that the backbones of the two monomers are very similar, with a root-mean-square deviation of 0.24 Å for all 221 pairs of α-carbons. The most significant main-chain deviations between the two molecules occur in residues 168C−168H and in residues 137−143. Residues 168C−168H form a loop at the surface of the protein. In monomer 1 the electron density for this loop was poorly defined, suggesting that these residues are flexible. In monomer 2, the corresponding loop is stabilized by interactions with a symmetry-related molecule. Presumably the crystal contact and the flexibility in this region account for the differences seen in this area. Residues 137−143 also lie at the surface of the protein. The discrepancy between the two molecules in this region is explained by the observation that, in both monomers, these residues are involved in crystal packing contacts with a neighboring molecule.

For the purpose of this discussion it was especially important to evaluate the model of the active site. An examination of the crystal lattice showed that were no crystal packing contacts in the immediate vicinity of the active site of either monomer. A comparison of 44 residues that comprise the active sites of the two monomers showed very good agreement, with a root-mean-square deviation of 0.09 Å for the α-carbons and a 0.49 Å deviation for all atoms. These observations indicate that this structure should provide a reliable model for the cathepsin V active site.

Overall Cathepsin V Structure. At low pH, procathepsin V self-activates to form the mature enzyme. The results of N-terminal sequencing showed that the first five amino acids of the activated protein are Leu-Pro-Lys-Ser-Val, indicating that the activation process led to a 221-residue mature protein that starts with Leu 114 of the preproenzyme.

Cathepsin V consists of two domains that are approximately equal in size (Figure 2). One of the domains is structurally homologous to the R-domain of papain, and consists primarily of a twisted β-sheet. The other domain is similar to the papain L-domain and contains three α-helices. The active site of cathepsin V is located in a wedge-shaped cleft at the interface between the R- and L-domains. The structure of cathepsin V is generally similar to the consensus structure for the papain family of cysteine proteases. Figure 3 shows a superposition of cathepsin V with papain (*22*) and with the closely related cathepsins L (*23*), K (*24*), and S (*25*). A comparison of cathepsin V with papain shows that the core of these two enzymes is well conserved, although there some variation in a number of the surface loops. Similarly, a comparison of cathepsin V with the human cathepsins S, K, and L shows that the overall fold of these

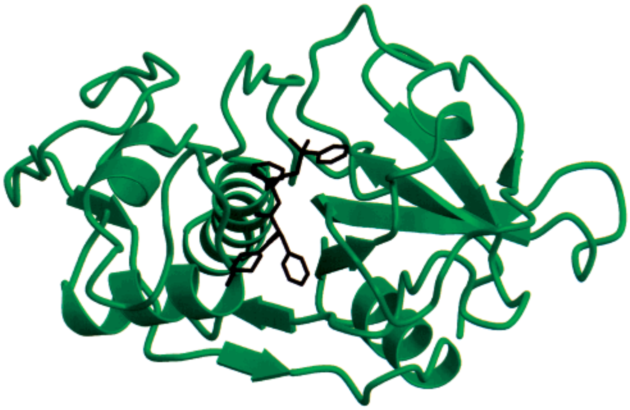

FIGURE 2: Schematic view of the structure of human cathepsin V. The coils represent α-helices and the arrows represent β-strands.

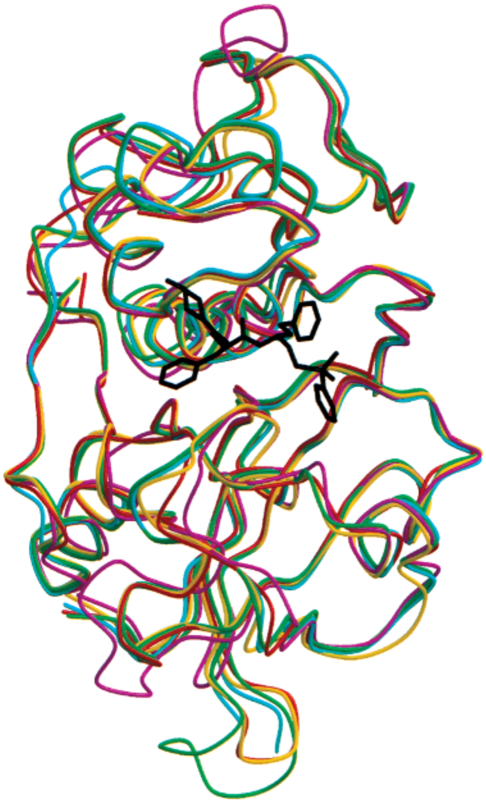

FIGURE 3: Cathepsin V is shown in green superimposed on the structures of papain (magenta) (9PAP) (*22*), cathepsin S (red) (*25*), cathepsin K (yellow) (*24*), and the mature part of procathepsin L (blue) (1CJL) (*23*). The overlay is based on a superposition of the residues that form the binding sites of these enzymes.

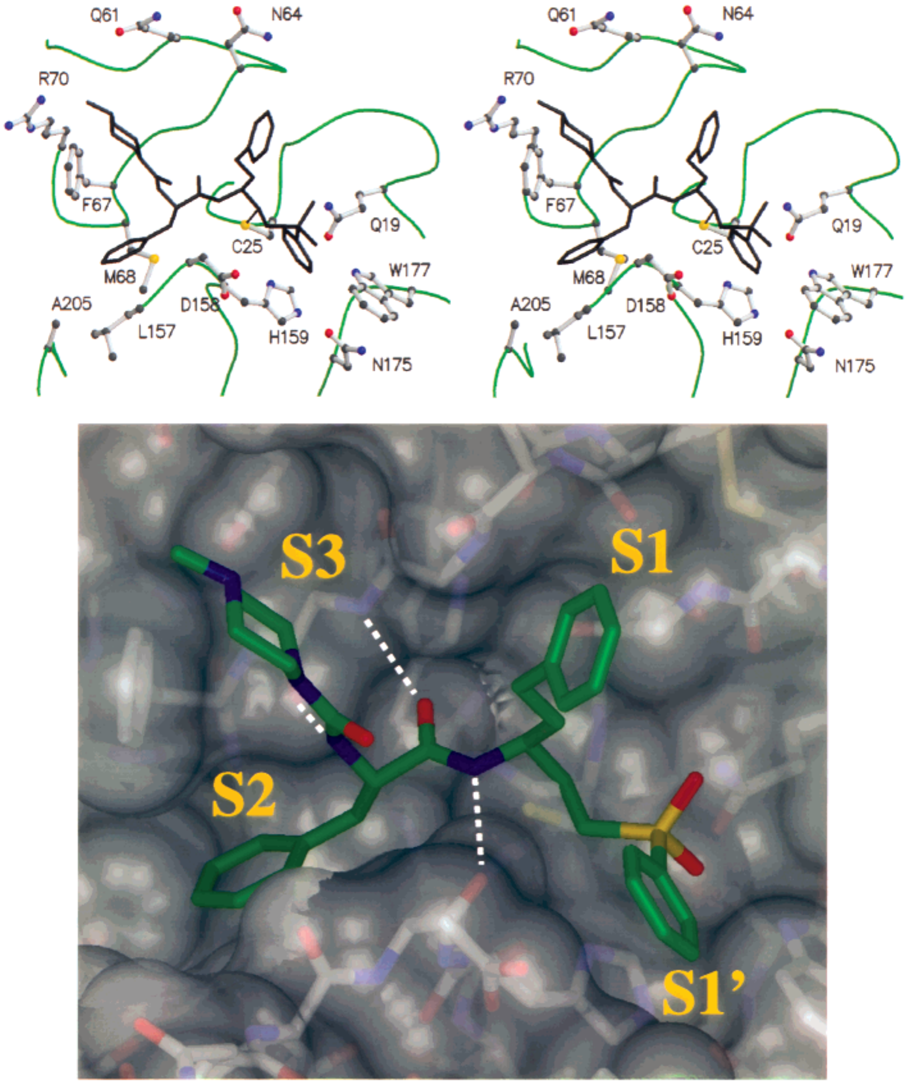

FIGURE 4: (A) Stereo diagram of the cathepsin V binding site. The side chains that are shown are those that are most important for catalysis or for inhibitor binding. (B) View of the solvent-accessible surface of the cathepsin V binding site with the structure of APC-3316 shown superimposed in green. The three hydrogen bonds between main-chain atoms of the inhibitor and the enzyme are shown as dotted white lines.

proteases is strikingly well conserved. The only area where the cathepsin V structure differs significantly from the other cathepsins is the 10-residue surface loop formed by residues 168A through 169 (this is the loop at the bottom of Figure 3). A sequence comparison of the papain-like cysteine proteases shows considerable variation in the length and sequence of insertions in this area of the structure (*4*).

Catalytic Machinery of Cathepsin V. Not all of the events that lead to catalysis in the cysteine proteases are well understood. However, there is general agreement on many of the basic events that take place during catalysis, and there is a wealth of studies that have probed the parts played by specific active-site residues (for reviews see refs *26* and *27*). In the cathepsin V structure, the active site is occupied by the covalent inhibitor APC-3316. This compound was developed as part of an effort to explore mechanism-based inhibition of cysteine proteases (*9*). It contains a vinyl sulfone that functions as a Michael acceptor. Binding of the inhibitor to the protease leads to a nucleophilic attack by the active-site cysteine Sγ on the inhibitor's vinyl carbon, much in the same way as the enzyme attacks the P1 carbonyl of a natural substrate.

The centerpiece of the catalytic machinery is formed by Cys 25 and His 159 (Figure 4A). Before substrate binding, these amino acids exist as a thiolate−imidazolium ion pair (*28*−*31*). The thiolate initiates the hydrolysis reaction through a nucleophilic attack on the carbonyl carbon of the substrate's scissile peptide bond. The histidine has a number of functions in catalysis, including interacting with the cysteine to help maintain its deprotonated state and protonating the leaving amine. It is also likely that the histidine functions as a general base in the deacylation reaction (*26*).

Adjacent to the catalytic cysteine is Gln 19. This residue is strictly conserved across the papain family of cysteine proteases and is probably involved in transition-state stabilization. The nucleophilic attack that initiates hydrolysis leads to a tetrahedral intermediate in which the negative charge has been transferred to the substrate's carbonyl oxygen. On the basis of structural and mutagenesis studies, it seems clear that this oxygen is stabilized through interactions with the side-chain amine of Gln 19 and with the backbone amide of Cys 25 (*32*). It is likely that the deacylation step also proceeds through a tetrahedral intermediate that is similarly stabilized by these two residues.

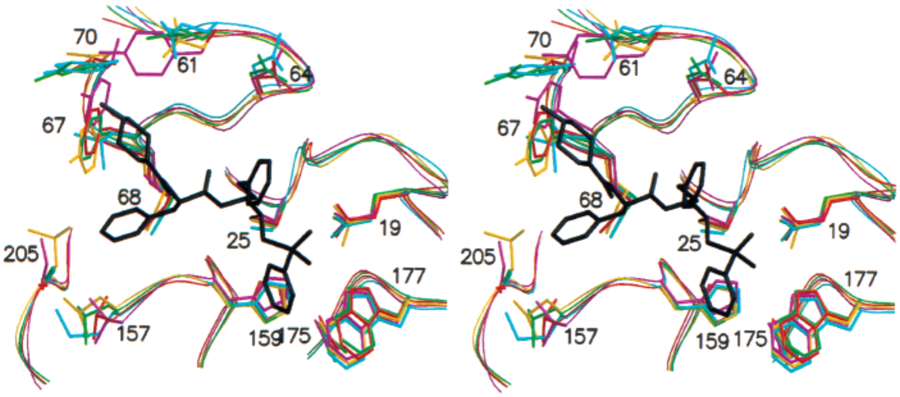

FIGURE 5: Superposition of the cathepsin V binding site (green) with papain (magenta) (9PAP) (*22*), cathepsin S (red) (*25*), cathepsin K (yellow) (*24*), and the mature part of procathepsin L (blue) (1CJL) (*23*). The side chains shown are most of those involved in substrate recognition and catalysis. Please note that Phe 205 and Lys 61 of cathepsin S are not shown because of problems with the original electron density (*25*). The superposition is based on α-carbons from residues that form the S1′ through S3 pockets of the binding site (residues 19−26, 63−68, 131−133, 157−159 and 176−177 of cathepsin V). All solvent molecules are excluded from this view, as are all inhibitors with the exception of the inhibitor bound to cathepsin V.

In cathepsin V there is a hydrogen bond between a side-chain oxygen of Asn 175 and the Nϵ2 of His 159. Although Asn 175 is conserved across the papain family, Vernet et al. (*33*) showed that this residue is not essential to the catalytic ability of papain. It is likely that Asn 175 plays an indirect role in catalysis by positioning the histidine correctly, and possibly by helping to maintain the structural integrity of the protease (*33*).

The cathepsin V structure suggests that there are a number of other residues that are likely to play a role in catalysis. For example, the indole ring of Trp 177 lies above His 159 and Asn 175 and seems to shield these residues from the solvent. There is also some evidence suggesting that the Trp−His interaction stabilizes the protonated form of the histidine (*34*, *35*). The importance of the Trp 177 is supported by its conservation across the papain family.

Substrate Binding Subsites. Whereas the cysteine proteases make use of a common catalytic apparatus, each protease has a unique substrate specificity that is determined by the identity and conformation of the residues that form the substrate-binding cleft. Information about the structural differences between binding clefts has led to an understanding of the unique specificities of each enzyme and has been helpful in identifying inhibitors that are able to discriminate between closely related proteases. As described above, the binding site of cathepsin V contains the irreversible inhibitor APC-3316. The backbone of this inhibitor is well tethered to the enzyme through a set of three hydrogen bonds (Figure 4B). The vinyl sulfone inhibitors are substratelike in their interactions with cysteine proteases. The hPhe, Phe, and *N*-methylpiperazine moieties of the inhibitor are structurally homologous to the P1, P2, and P3 residues of the natural substrate and lie in the S1, S2 and S3 pockets of the protease, respectively. The sulfone phenyl group binds in the S1′ region.

The S1 pocket of cathepsin V is occupied by the inhibitor's hPhe moiety. This pocket is formed by Asn 64, Gly 23, Gly 65, and the carbonyl oxygen of Cys 63. The pocket is very shallow, little more than an indentation in the L-domain. A comparison with the structures of some related proteases (Figure 5) shows that this pocket is well conserved, although a sequence comparison reveals that there is some variation in the identity of residue 64 over the papain superfamily (*4*). The structural conservation, along with the relatively shallow features of the S1 site, suggests that it will be difficult to exploit this region for the design of specific inhibitors.

The S2 pocket is the most clearly defined of the binding sites and is the major specificity determinant in the papain family of cysteine proteases. S2 is a deep cleft in the R-domain (Figure 4B). In cathepsin V, the floor of this cleft is formed primarily by Ala 133, Gly 160, and the side chain of Met 68, while the sides are formed by Lys 155, Asn 156, Leu 157, Phe, the main chain of Asp 158, and the carbonyl oxygen of Gly 66. The far side of the pocket is created by the methyl group of Ala 205. Interestingly, this alanine does not completely block the pocket, resulting in a passage to another pocket. While this extension of the S2 pocket is probably not important in the binding of the enzyme's natural substrates, it might be relevant to inhibitor design.

In this structure, the S2 pocket is filled with the phenylalanine at the P2 position of the inhibitor. This moiety is well defined by the electron density maps and is tightly packed between Phe 67 and Ala 133. A comparison with the structures and sequences of other cysteine proteases reveals considerable variation in the identity and conformation of residues that form the S2 pocket (Figure 5) (*4*). One of the primary determinants of the width of the S2 pocket is residue 67. In cathepsin V this is a phenylalanine, but its identity differs in other cysteine proteases (*4*). Aside from the differences in the identity of this residue, there is striking variability in the position of its side chain (Figure 5). This variability is due to differences in the local main-chain conformation, side-chain conformation, and interactions between this residue and residues 59, 60, 70, and in some cases with residue 69. Differences in the position of the side chain of 67 lead to differences in the width of S2 that, in turn, translate into differences in the size of the residue at P2 that can be accommodated in the binding site. Residue 157 lies opposite Phe 67. This residue is a leucine in cathepsins V and K, a methionine in cathepsin L, and a valine in cathepsin S. The variation in this amino acid's position and identity is also a source of differences in the width and shape of S2 that may also be useful in identifying specific inhibitors. Residue 205, at the end of the S2 pocket, also shows considerable variability. This residue plays an important part in defining the length of the pocket and in determining whether passage to a distal pocket will be blocked. In cathepsin V this residue is an alanine, making the pocket longer than in cathepsins S and K (which have a phenylalanine and leucine, respectively). As mentioned above, the presence of a small side chain at this position in cathepsin V also results in a passage between the S2 pocket and another pocket deeper in the enzyme. Overall, the depth of the S2 site makes it likely that it is a major determinant of substrate specificity in cathepsin V, as it is in many of the members of the papain family. The variability in this subsite makes it a tempting target in the design of specific inhibitors.

The S3 site of cathepsin V is occupied by the *N*-methylpiperazine moiety at the P3 position of the inhibitor. The bottom of S3 is formed by glycines 65 and 66 and part of the Asn 60 side chain. The sides of the pocket are created by the side chains of Phe 67, Arg 70, and Gln 61, as well as by the main-chain carbonyls of Asn 64 and Gly 59. The shape of the S3 pocket in cathepsin V differs substantially from that of other cysteine proteases. This is due to differences in each of the three side chains that form the sides of the subsite as well as to differences in the position of the carbonyl oxygen of residue 59. The variability in the side chain at position 67 was discussed in the context of the S2 site. The identity and conformation of this side chain also affect the shape and depth of S3. The characteristics of the rest of the S3 wall are largely defined by Gln 61 and Arg 70. There is little conservation in the position or identity of either of these residues across the papain superfamily (*4*), leading to a range of differences in the characteristics of this subsite. From the point of view of inhibitor design, the presence the guanidino group of Arg 70 in the S3 pocket of cathepsin V is especially interesting. None of the related enzymes has an arginine at this position, so inhibitors that interact with this residue might be selective for cathepsin V. Finally, the end of the site is influenced by the main chain of Gly 59, with its carbonyl oxygen pointing into the binding site. A comparison of papain and cathepsins V, S, K, and L shows that the position of the main-chain atoms of this residue is variable, leading to differences in the depth of S3. For example, the carbonyl oxygen of cathepsin K is about 2 Å further from the entrance to S3 than the corresponding atom of cathepsin S. The reason for this main-chain variation can be traced to extra residues in this region of the sequence relative to papain. For example, cathepsins V and L have an insertion of two residues (compared to papain) in this region, while cathepsin S has a three-residue insertion. On the basis of the structures currently available, it appears that the larger the insertion, the closer the main chain is pushed toward the opening of the S3 pocket, thereby making the pocket smaller.

The sulfone phenyl group of the inhibitor occupies the S1′ pocket of cathepsin V. One of the sulfone oxygens lies 3.4 Å from the side chain of the Gln that forms part of the oxyanion hole, and the other oxygen interacts with the Nϵ1 of Trp 177. The phenyl ring binds in a pocket formed by the side chain of Trp 177, the side chain of His 159, the side chain and the carbonyl carbon of Asp 158, and Ala 136. This pocket, and the S′ region in general, is fairly well conserved across the much of the papain family. This is not surprising since many of the residues that form this area also form part of the highly conserved catalytic machinery. From a drug design perspective, it should be possible to exploit the S′ region for achieving potency, but interactions in this area are unlikely to lead to more selective compounds.

CONCLUSIONS

The sequence and structure of cathepsin V clearly place it within the papain superfamily of cysteine proteases. This work provides a framework for understanding the structural basis for cathepsin V's activity and should help in the design of cathepsin V inhibitors. The comparison of this structure with the structures of some of the closely related proteases revealed a number of differences that could be useful in the identification of selective, therapeutically relevant protease inhibitors.

ACKNOWLEDGMENT

We thank the Mass Spectrometry Center, Department of Medicinal Chemistry at the University of Washington, for their help with the mass spectrometry measurements. We also thank Dr. Robert McDowell for his support of this project.

REFERENCES

1. Santamaría, I., Velasco, G., Cazorla, M., Fueyo, A., Campo, E., and López-Otín, C. (1998) *Cancer Res. 58*, 1624−1630.
2. Adachi, W., Kawamoto, S., Ohno, I., Nishida, K., Kinoshita, S., Matsubara, K., and Okubo, K. (1998). *Invest. Ophthalmol. Visual Sci. 39*, 1789−1796.
3. Brömme, D., Li, Z., Barnes, M., and Mehler, E. (1999) *Biochemistry 38*, 2377−2385.
4. Berti, P. J., and Storer, A. C. (1995) *J. Mol. Biol. 246*, 273−283.
5. Itoh, R., Kawamoto, S., Adachi, W., Kinoshita, S., and Okubo, K. (1999) *DNA Res. 6*, 137−140.
6. Schilling, D. M., and Ahlquist, D. A. (1999) *Gastroenterology 116*, A370.
7. Nakagawa, T., Roth, W., Wong, P., Nelson, A., Farr, A., Deussing, J., Villadangos, J. A., Ploegh, H., Peters, C., and Rudensky, A. Y. (1998) *Science 280*, 450−453.
8. Fehrentz, J. A., and Castro, B. (1983) *Synthesis 8*, 676−678.
9. Palmer, J. T., Rasnick, D., Klaus, J. L., and Brömme, D. (1995) *J. Med. Chem. 38*, 3193−3196.
10. Nowick, J. S.; Powell, N. A.; Nguyen, T. M., and Noronha, G. (1992) *J. Org. Chem. 57*, 7364−7366
11. Otwinowski, Z., and Minor, W. (1993) in *Data collecting and processing*, SERC Daresbury Laboratory, Warrington, U.K.
12. Kissinger, C. R., Gehlhaar, D. K., and Fogel, D. B. (1999) *Acta Crystallogr. D55*, 484−491.
13. Groves, M. R., Coulombe, R., Jenkins, J., and Cygler, M. (1998) *Proteins: Struct., Funct., Genet. 32*, 504.
14. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. D54*, 905−921.
15. Rice, L. M., and Brünger, A. T. (1994) *Proteins: Struct., Funct., Genet. 19*, 277−290.
16. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) *J. Appl. Crystallogr. 26*, 283−291.
17. Altschul, S. F., Madden, T. L. Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res. 25*, 3389−3402.
18. Kabsch, W., and Sander, C. (1983) *Biopolymers 22*, 2577−2637.
19. Kraulis, P. J. (1991) *J. Appl. Crystallogr. 24*, 946−950.
20. Esnouf, R. M. (1997) *J. Mol. Graphics 15*, 132−134.
21. Merritt, E. A., and Murphy, M. E. P. (1994) *Acta Crystallogr. D50*, 869−873.
22. Kamphuis, I. G., Kalk, K. H., Swarte, M. B. A., and Drenth, J. (1984) *J. Mol. Biol. 179*, 233−256.

23. Coulombe, R., Li, Y., Takebe, S., Ménard, R., Mason, P., Mort, J. S., and Cygler, M. (1996) *EMBO J. 15*, 5492−5503.
24. McGrath, M. E., Klaus, J. L., Barnes, M. G., and Brömme, D. (1997) *Nat. Struct. Biol. 4*, 105−109.
25. McGrath, M. E., Palmer, J. T., Brömme, D., and Somoza, J. R. (1998) *Protein Sci. 7*, 1294−1302.
26. Storer, A. C., and Ménard, R. (1994) *Methods Enzymol. 244*, 486−500.
27. Otto, H.-H., and Schirmeister, T. (1997) *Chem. Rev. 97*, 133−171.
28. Polgar, L. (1974) *FEBS Lett. 47*, 15−18.
29. Lewis, S. D., Johnson, F. A., and Shafer, J. A. (1976) *Biochemistry 15*, 5009−5017.
30. Sluyterman, L. A. Æ., and Wijdenes, J. (1976) *Eur. J. Biochem. 71*, 383−391.
31. Lewis, S. D., Johnson, F. A., and Shafer, J. A. (1981) *Biochemistry 20*, 48−51.
32. Ménard, R., Carriere, J., Laflamme, P., Plouffe, C., Khouri, H. E., Vernet, T., Tessier, D. C., Thomas, D. Y., and Storer A. C. (1991) *Biochemistry 30*, 8924−8928.
33. Vernet, T., Tessier, D. C., Chatellier, J., Plouffe, C., Lee, T. S., Thomas, D. Y., Storer, A. C., and Ménard, R. (1995) *J. Biol. Chem. 270*, 16645−16652.
34. Loewenthal, R., Sancho, J., and Fersht, A. R. (1992) *J. Mol. Biol. 224*, 759−770.
35. Brömme, D., Bonneau, P. R., Purisima, E., Lachance, P., Hajnik, S., Thomas, D. Y., and Storer, A. C. (1996) *Biochemistry 35*, 3970−3979.

BI000951P