Crystal structures of native and inhibited forms of human cathepsin D: Implications for lysosomal targeting and drug design

(aspatic protease/N-linked oligosaccharide/pepsatin A)

ERIC T. BALDWIN*, T. NARAYANA BHAT*, SERGEI GULNIK*, MADHUSOODAN V. HOSUR†, RAYMOND C. SOWDER II‡, RAUL E. CACHAU*, JACK COLLINS, ABELARDO M. SILVA*, AND JOHN W. ERICKSON*§

*Structural Biochemistry Program, Frederick Biomedical Supercomputing Center and ‡AIDS Vaccine Program, Program Resources Inc./DynCorp, National Cancer Institute–Frederick Cancer Research and Development Center, Frederick, MD 21702

Communicated by David R. Davies, March 24, 1993 (received for review February 4, 1993)

ABSTRACT Cathepsin D (EC 3.4.23.5) is a lysosomal protease suspected to play important roles in protein catabolism, antigen processing, degenerative diseases, and breast cancer progression. Determination of the crystal structures of cathepsin D and a complex with pepstatin at 2.5 Å resolution provides insights into inhibitor binding and lysosomal targeting for this two-chain, N-glycosylated aspartic protease. Comparison with the structures of a complex of pepstatin bound to rhizopuspepsin and with a human renin–inhibitor complex revealed differences in subsite structures and inhibitor–enzyme interactions that are consistent with affinity differences and structure–activity relationships and suggest strategies for fine-tuning the specificity of cathepsin D inhibitors. Mutagenesis studies have identified a phosphotransferase recognition region that is required for oligosaccharide phosphorylation but is 32 Å distant from the N-domain glycosylation site at Asn-70. Cathepsin D (EC 3.4.23.5) is an aspartic protease that is normally found in the lysosomes of higher eukaryotes where it functions in protein catabolism (1). This enzyme is distinguished from other members of the pepsin family (2) by two features that are characteristic of lysosomal hydrolases. First, mature cathepsin D is found predominantly in a two-chain form due to a posttranslational cleavage event (1, 3). Second, it contains phosphorylated, N-linked oligosaccharides that target the enzyme to lysosomes via mannose 6-phosphate (M6P) receptors (4, 5). Phosphorylation involves recognition of both sugar and protein structural determinants by a phosphotransferase enzyme (6, 7).

Interest in cathepsin D as a target for drug design results from its association with several biological processes of therapeutic significance including lysosomal biogenesis and protein targeting (4, 5), antigen processing and the presentation of peptide fragments to class II major histocompatibility complexes (8–10), connective tissue disease pathology (11), muscular dystrophy (12), degenerative brain changes (13, 14), and cleavage of amyloid precursor protein within senile plaques of Alzheimer brain (15). Recent studies of primary breast cancers demonstrated that elevated levels of cathepsin D were correlated with an increased risk of metastasis and shorter relapse-free survival (16). High levels of cathepsin D and other proteases such as collagenase, produced in the vicinity of the growing tumor, may degrade the extracellular matrix and thereby promote the escape of cancer cells to the lymphatic and circulatory systems and enhance the invasion of new tissues (17, 18). The design of potent and specific inhibitors of cathepsin D will aid the further elucidation of the roles of this enzyme in human disease. We previously described the purification and crystallization of human cathepsin D from liver (3); similar studies have been reported recently for cathepsin D isolated from bovine liver (19) and human spleen (20). We now describe the complete three-dimensional structure of native and pepstatin-inhibited forms of human cathepsin D.§

MATERIALS AND METHODS

Crystals of native human liver cathepsin D (space group P6_3; a = b = 125.9 Å, c = 104.1 Å) were prepared as described (3). The structure was solved by molecular replacement using the program XPLOR-3.0 (21). The crystal structure of porcine pepstatin pepsin was used as the search molecule (22). The final R factor was 18.8% for data from 10.0 to 2.5 Å resolution for 28,077 unique reflections. The rms deviations from ideality for bond lengths, angles, dihedral torsions, and improper torsions were 0.012 Å, 3.1°, 27.3°, and 1.14°, respectively. Cococrystals of cathepsin D with pepstatin were prepared by using conditions similar to those used for the crystallization of native enzyme. The structure of the complex was solved by using the refined native structure of cathepsin D and difference Fourier methods. After a single round of refinement, the resulting 2F_o − F_c and F_o − F_c maps showed substantial electron density in the active-site region. The inhibitor was built and the complex was refined to a final R factor of 17.9% with XPLOR. The rms deviations from ideality for bond lengths, angles, dihedral torsions, and improper torsions were 0.012 Å, 3.2°, 26.6°, and 1.14°, respectively.

The inhibition constants for pepstatin were determined by using a fluorometric assay at 37°C with the substrate Ac-Glu-Glu-Glu-Glu(Edans)-L-lys-Pro-Ile-Cys-Phe-Phe-Arg-Leu-Gly-Lys(Dabcyl)-Ser-Asp-Dabcyl-NH_2, which was kindly supplied by Grant Krafft (Abbott) (Edans is 5-(2-aminoethyl)amino)napthalene-1-sulfonic acid and Dabcyl is p-(dimethylaminophenylazo)benzoic acid). The K_i for this substrate was 5 ± 0.6 μM and 0.3 ± 0.07 μM for cathepsin D and rhizopuspepsin, respectively. In the case of cathepsin D, the inhibitor concentration was

Abbreviations: M6P, mannose 6-phosphate; Nag, N-acetylgalactosamine.

§To whom reprint requests should be addressed.

The atomic coordinates and structure factors have been deposited in the Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton NY 11973 (reference 1LYA, 1LYB).
less than or equal to the enzyme concentration, and kinetic data were analyzed by using a model for tight-binding inhibitors (23). Michaelis–Menten kinetics were assumed with rhizopuspepsin. The experimental data were fitted by nonlinear regression analysis.

RESULTS AND DISCUSSION

The 2.5 Å resolution structures of native cathepsin D and of its isomorphous complex with pepstatin, a naturally occurring aspartic protease inhibitor, were solved by molecular replacement. The crystallographic asymmetric unit of the P6_3 unit cell contains two molecules that are identical and related by a pseudo-twofold rotation, \( \kappa = 160.9^\circ \), and by a relative translation of 36 Å. Cathepsin D contains three topologically distinct regions that are typical of aspartic proteases (Fig. 1): an N-terminal domain (residues 1-188), a C-terminal domain (residues 189-346), and an interdomain, anti-parallel \( \beta \)-sheet composed of the N terminus (residues 1-7), the C terminus (residues 330-346), and the interdomain-linking residues (160-200). The latter region links the pseudo-twofold-related N and C domains (\( \kappa = 179.3^\circ \) for the superposition of 19 \( C^\alpha \) atoms with a rms of 1.5 Å), each of which contributes an aspartic acid, Asp-33 and Asp-231, to the active site. The overall structural agreement for the main chain of cathepsin D with the known structures of other mammalian aspartic proteases, human renin (24), porcine pepsin (Protein Data Bank identification code, 4PEP; ref. 25), and bovine chymosin (1CMS; ref. 25), ranged between 0.9 and 1.0 Å rms.

Several structural features distinguish cathepsin D from other aspartic proteases. The cleavage and excision of an external loop between residues 98 and 106 in the N domain results in the noncovalent association of light and heavy chains (1). Electron density maps of cathepsin D show main-chain density for the light chain from Gly-1 through Gln-97. Continuous density for the heavy chain extends from Gly-106 to the C-terminal residue, Ala-346. Peptide sequencing of the protein preparation used for our crystallization experiments indicated that the light chain terminates with Ser-98 and that the heavy chain usually begins with Gly-106 (59%) and less frequently with Leu-105 (27%) or Gly-107 (13%). Electron density maps also indicated the presence of a disulfide bridge in cathepsin D between Cys-27 and Cys-96, which has not been observed for other aspartic proteases.

This disulfide bridge, which had been predicted from modeling studies of porcine cathepsin D (26), is within the light chain region and may stabilize the C terminus of the light chain against further proteolytic degradation. There are three proline residues in cathepsin D that refined in a cis-peptide conformation. A cis-peptide bond involving Pro-24 is conserved among all the aspartic proteases except for penicillopepsin, which has a deletion in this region. The other proline residues with cis-peptide bonds are localized within a proline-rich segment, termed the "proline loop," that consists of residues Pro-312, Pro-313, cis-peptide Pro-314, Ser-315, Gly-316, and cis-peptide Pro-317. An analogous structure has been described recently for human renin (24).

Binding of pepstatin to cathepsin D induced small structural changes in the "flap" region (the \( \beta \)-hairpin structure composed of residues 72-87; Fig. 1). Residues 79 and 80 at the tip of the flap moved in toward the inhibitor by about 1.7 Å. The flexibility of this \( \beta \)-bend decreased upon inhibitor binding as indicated by a drop in the mean B factor from 64 Å^2 to 25 Å^2 for the 16 main chain atoms of residues 78-81. Small flap movements due to inhibitor binding have been observed for other aspartic proteases (2). No subdomain displacements were observed in cathepsin D upon inhibitor binding, unlike the case of pepsin (27) and endothiapepsin (28).

The pepstatin–cathepsin D complex is stabilized by numerous hydrogen bonds between backbone atoms of the inhibitor and both main chain and side chain atoms of the enzyme (Fig. 2A). The central statine hydroxyl group occupies the position of a water molecule that interacts with the two active-site aspartate residues in the native enzyme structure. In both the native and inhibited structures of cathepsin D, the active-site carboxylate groups are nearly coplanar (rms deviation from the plane < 0.1 Å for six atoms). Surprisingly, the P3' statine residue of the inhibitor makes no obvious hydrogen bond with the enzyme despite the presence of several donor and acceptor substituents. His-77 is within 4 Å of the statine carboxylate group and may make an electrostatic interaction with the C terminus of the inhibitor.

The inhibitor side chains make extensive van der Waals contacts with the enzyme subsites (Fig. 2B). In the pepstatin complex, the S3, S2, S1, and S2' subsites are only partially filled and could accommodate bulkier inhibitor substituents at the corresponding positions on the inhibitor. These con-
conclusions are consistent with structure–activity relationship studies with bovine cathepsin D in which leucine was favored over alanine at P3, 4-amino-3-hydroxy-5-phenylpentanoic acid was favored over statine at P1, and valine was favored over alanine at P2 for synthetic analogues of pepstatin (30). Studies with human cathepsin D showed that homocysteine was preferred over cysteine at P2 (31). The S1' and S3' subsites interact with the C-terminal statine substituent. The central statine residue of pepstatin can be considered to be a P1-P1' dipeptide isostere of Leu-Gly for which there is no P1' side chain substituent. In the structure of the pepstatin–cathepsin D complex, the leucine moiety of the P3' statine residue folds back in order to partially occupy the S1' subsite (Fig. 2B). The (3S)-hydroxyl group on the P1 statine residue contributes substantially to the binding affinity of cathepsin D for pepstatin since the deoxy and (R)-hydroxy analogues were much weaker inhibitors (30). The structure of the complex reveals that the (S)-OH configuration permits the formation of favorable hydrogen bonds to both Asp-33 and Asp-231, while the R configuration would allow interaction with only Asp-33. Addition of a methyl group to the C-3 position also leads to less favorable binding (30). Examination of the structure indicates that the binding of a 3-methylstatine analogue would require conformational alterations of the enzyme or inhibitor structures to alleviate bad contacts.

The crystal structure of a complex of pepstatin with the fungal aspartic protease, rhizopuspepsin, has been reported recently (32) and enables us to evaluate the structural basis of binding affinity differences for the same inhibitor with different enzymes. Inhibition constants (K_i) for pepstatin with human cathepsin D and rhizopuspepsin were determined to be 3.8 ± 1.0 and 150 ± 14 pM, respectively. In both structures, pepstatin adopts a very similar conformation; the rms difference for 41 nonhydrogen atom pairs is 0.43 Å. Cathepsin D makes two hydrogen bonds to the inhibitor that are not observed in the rhizopuspepsin complex (Table 1). The OH group of Tyr-205 is hydrogen bonded to the P2' carbonyl oxygen, and Ser-235 O' forms a hydrogen bond to the P4 carbonyl oxygen. In rhizopuspepsin, Tyr-205 is replaced by Trp-194, which points toward the P2' carbonyl

![Diagram](image_url)

**Table 1. Hydrogen bonds between pepstatin and cathepsin D or rhizopuspepsin**

<table>
<thead>
<tr>
<th>Residue</th>
<th>Atom</th>
<th>Residue</th>
<th>Atom</th>
<th>Distance (Å)</th>
<th>Residue</th>
<th>Atom</th>
<th>Distance (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P4</td>
<td>O</td>
<td>Ser-235</td>
<td>O</td>
<td>2.6</td>
<td>Thr-222</td>
<td>O</td>
<td>3.1</td>
</tr>
<tr>
<td>P3</td>
<td>N</td>
<td>Ser-80</td>
<td>O</td>
<td>3.5</td>
<td>Thr-222</td>
<td>N</td>
<td>3.0</td>
</tr>
<tr>
<td>P2</td>
<td>N</td>
<td>Ser-80</td>
<td>O</td>
<td>2.7</td>
<td>Asp-79</td>
<td>O</td>
<td>3.1</td>
</tr>
<tr>
<td>P1</td>
<td>N</td>
<td>Gly-233</td>
<td>O</td>
<td>2.9</td>
<td>Asp-79</td>
<td>N</td>
<td>3.1</td>
</tr>
<tr>
<td>OH</td>
<td>Asp-33</td>
<td>O[1]</td>
<td>2.8</td>
<td>Asp-35</td>
<td>O[2]</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>OH</td>
<td>Asp-33</td>
<td>O[2]</td>
<td>2.6</td>
<td>Asp-35</td>
<td>O[2]</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>OH</td>
<td>Asp-231</td>
<td>O[2]</td>
<td>2.8</td>
<td>Asp-218</td>
<td>O[2]</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>OH</td>
<td>Asp-231</td>
<td>O[2]</td>
<td>2.6</td>
<td>Asp-218</td>
<td>O[2]</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>OH</td>
<td>Gly-233</td>
<td>O</td>
<td>3.4</td>
<td>Gly-220</td>
<td>O</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Gly-79</td>
<td>N</td>
<td>2.9</td>
<td>Gly-78</td>
<td>N</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>P2'</td>
<td>N</td>
<td>Gly-35</td>
<td>O</td>
<td>2.9</td>
<td>Gly-37</td>
<td>O</td>
<td>3.1</td>
</tr>
<tr>
<td>O</td>
<td>Tyr-205</td>
<td>O</td>
<td>2.9</td>
<td>Trp-194</td>
<td>N*</td>
<td>(3.6)</td>
<td></td>
</tr>
</tbody>
</table>

The hydrogen bonds are defined as having donor-acceptor distances within 3.4 Å from the protein to the inhibitor. Distances greater than 3.4 Å are indicated in parentheses.
group but does not form a hydrogen bond with the inhibitor. Although Ser-235 is replaced by Thr-222 in rhizopuspepsin, the P4 carbonyl oxygen in pepstatin points toward the opposite side of the substrate-binding cleft due to a difference in the backbone torsion angle for the inhibitor. Pepstatin makes ≈50% more van der Waals contacts in cathepsin D than in rhizopuspepsin, as measured by the number of heavy atom contacts within 4.2 Å. In the S3' subsite, the presence of His-77 in cathepsin D affords more contacts with pepstatin. There is a serine at this position in rhizopuspepsin. In cathepsin D, the presence of the smaller Ser-80 side chain (Asp-79 in rhizopuspepsin) permits the flap to move closer to the inhibitor and results in an increase in the number of van der Waals contacts between enzyme and inhibitor in the S2 subsite. The increased number of hydrogen bonds and van der Waals contacts in the cathepsin D complex is consistent with the lower Kᵢ observed with cathepsin D for pepstatin.

Differences in the subsite structures of human cathepsin D and related human aspartic proteases may be exploited for drug design. Comparison of the crystal structures of cathepsin D with human renin (24, 33) reveals that the substrate-binding cleft in cathepsin D is wider overall than in renin. Cathepsin D has smaller residues in the S2 (Gly-79 vs. Ser-76) and S4 (Leu-236 vs. Tyr-220) subsites, which leads to larger subsite volumes, which is consistent with the observation that bulky P4 substituents contribute more to the potency of inhibitors for cathepsin D than for pepsin and renin (30, 31). In renin, the proline loop constitutes part of the S' portion of the substrate-binding cleft. The analogous segment in cathepsin D (residues 312–317) does not make contact with pepstatin (Fig. 1), despite the fact that this inhibitor extends considerably farther into the S' half of the active site than do the shorter renin inhibitors. The explanation appears to be that renin and cathepsin D exhibit different relative displacements of the small subdomains, which results in about a 3 Å wider binding cleft for the latter enzyme.

Phosphotransferase recognition of protein structural determinants and the N-linked oligosaccharide chains at Asn-70 and Asn-199 has been shown to play a key role in the targeting of cathepsin D to lysosomes (4, 5). Mutagenesis studies using chimeric pepsinogen–cathepsin D molecules defined a minimal protein recognition domain for phosphotransferase consisting of Lys-203 and residues 265–292 (6, 7). The structure of the phosphotransferase recognition region formed by residues 265–292 in cathepsin D is seen clearly in our electron density maps, and it is topologically similar to a homologous region in pepsinogen that served as the basis for a previous cathepsin D model that has been used to evaluate lysosomal targeting experiments (6, 7). Residues 265–292 in cathepsin D can be superposed onto residues 250–277 of pepsinogen (1PSG; ref. 25) to within 0.5 Å rms for all equivalent Cα atom pairs. However, when the active-site regions for the two

![Fig. 3. Glycosylation sites and phosphotransferase recognition regions for cathepsin D. (A) Stereoview showing the modeled oligosaccharide structure (blue) extending from Asn-70. The core oligosaccharide residues, N-acetyllactosamine (Nag)-401, Nag-402, Man-403, and Man-408, were fitted initially into 2F₀ − F₁ electron density contoured at 1.0 σ (transparent white surfaces) by using the oligosaccharide conformation found in the crystal structure of Ecor lectin (34). The remaining sugar residues were modeled beyond the ordered electron density. In this model, the phosphate group of M6P-406 makes an electrostatic interaction with the amino group of Lys-203. Protein portions are shown in purple. (B) Stereoview showing the disposition of N- and C-domain glycosylation sites, the proposed oligosaccharide structure, and the phosphotransferase recognition region on cathepsin D. Both sugar chains and Lys-203 extend from the same side of cathepsin D. The sugar chain on Asn-199 (lower right) projects into the solvent and is absent or disordered in our structure, whereas the sugar chain at Asn-70 (far left) extends along the protein surface and places the M6P of the terminal sugar residue near Lys-203 (bottom center). Residues 265–292, which constitute the phosphotransferase recognition region (6, 7), are shown in purple.](image-url)
proteins are first superimposed, the above segments are displaced positionally by about 3.0 Å rms. This analysis reflects the different relative subdomain displacements for cathepsin D and pepsinogen, which could not have been predicted \textit{a priori} from the latter structure alone.

Continuous electron density extended from the side chain N\textsubscript{64} atom of Asn-70, which corresponds to the N-domain glycosylation site. Four sugar residues that correspond to the core oligosaccharide structure for cathepsin D (6, 7) were fitted to the density (Fig. 3A). The side chain O\textsubscript{4} atom of Asn-70 is hydrogen bonded to the hydroxyl group of Thr-72 in an "asparagine turn" conformation (35, 36). This interaction stabilizes the Asn-70 side chain and may help to orient the N-linked oligosaccharide. Evidence for C-domain glycosylation at Asn-199 was weak, and a single N\textsubscript{6} residue was modeled into the initial $F_{o} - F_{c}$ map. We conclude that the oligosaccharide structure at this position is either degraded or disordered.

The crystal structure reveals that Lys-203 and both N-linked oligosaccharides extend from the same face of the protein molecule (Fig. 3B). Lys-203 is found just below the active-site cleft on the S\textsuperscript{2} side between Asn-70 and Asn-199, which are separated by 36 Å. Based on the proposed chemical structure of the phosphorylated oligosaccharide (4, 5), the Asn-70-linked sugar chain was extended beyond the branch point at Man-403 (Fig. 3A). In the resulting model structure, the oligosaccharide extends along the protein surface towards Lys-203, and the phosphate oxygens of M\textsubscript{Q}P-406 lie within 5.0 Å of the ε-amino group of this residue. Our model provides an explanation as to how sites that are 28 Å distant on the enzyme surface, Asn-70 and Lys-203, can both contribute to recognition via a flexible oligosaccharide chain. This model is consistent with recent experiments using pepsinogen–cathepsin D chimeras, which demonstrated that oligosaccharides located at considerable distances from the protein recognition domain can still be phosphorylated (37–39).

The present results provide a structural basis for further investigations of the targeting of cathepsin D to lysosomes and to the design and synthesis of specific and potent cathepsin D inhibitors. Such inhibitors will help to clarify the role of cathepsin D in a variety of important disease processes and could lead to the development of a new class of therapeutic agents.

\textbf{Note Added in Proof:} Our crystallographic results are in general accord with predictions of subsite preferences that were based on a recent modeled structure of human cathepsin D (40).

We wish to thank Karen Friel for assistance with the manuscript preparation, T. J. O'Donnell for Fig. 3A using the program GRAMPS, David Davies for the rhizopusspepsin coordinates, Tom Blundell for renin coordinates, Boaz Shaanan for providing oligosaccharide topology files, and Alex Wlodawer and Stan Burt for helpful discussion. We acknowledge the excellent assistance of the operations staff at the Frederick Biomedical Supercomputing Center. This research was supported by the National Cancer Institute, Department of Health and Human Services, under contract no. N01-CC-74102 with Program Resources Inc./DynCorp.