BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

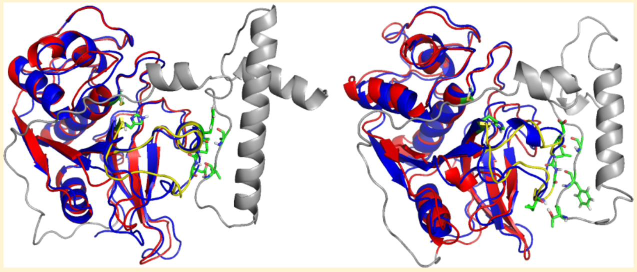
Terms of Use

Molecular Insight into Propeptide–Protein Interactions in Cathepsins L and O

Maria M. Reif,[†] Lukas Mach,[‡] and Chris Oostenbrink*,[†]

[†]Institute for Molecular Modeling and Simulation and [‡]Department of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences, 1190 Vienna, Austria

Supporting Information



ABSTRACT: Cathepsins are mammalian papain-like cysteine proteases that play an important role in numerous physiological and pathological processes. In the present study, various molecular dynamics (MD) simulations of pro- and mature human cathepsins L and O were performed. This study is the first to report MD simulations to complement the initial model structure of (pro-)cathepsin O through conformational sampling, thus offering insight into the maturation of procathepsin O, which to date has not been described experimentally. The overall fold of (pro-)cathepsin O appears very similar to that of (pro-)cathepsin L. The propeptide binding loop (PBL)–propeptide interface of both procathepsins is found to form a stable two-stranded β-sheet. Additional stabilization of the PBL–propeptide interface is provided by hydrophobic side chain contacts in procathepsin L, whereas this seems to be due to charge-dipole interactions in procathepsin O. Introduction of two mutations (L147P and G148P) into procathepsin O entails a significant loss of hydrogen bonding, disabling formation of the interfacial β-sheet. Simulations at different protonation states suggest that procathepsin L is more sensitive to a change in pH than procathepsin O. Potential differences between the maturation of procathepsin O and procathepsin L inferred from the MD simulations might be caused by (i) stronger PBL–propeptide interactions in procathepsin O due to salt-bridge formation across the interface, (ii) more limited entropic gain of the propeptide of procathepsin O upon release into the bulk solvent due to diverse conformational states sampled in the bound state, (iii) more pronounced entropic loss of the PBL in procathepsin O upon substrate binding caused by diverse conformational states sampled in the free, mature enzyme, and (iv) lower sensitivity of procathepsin O to pH change caused by the presence of fewer carboxylate groups at the PBL–propeptide interface.

Protease enzymes are divided into seven main families, namely, metallo, serine, threonine, asparaginyl, aspartic, glutamic, and cysteine proteases. The predominant subclass of cysteine proteases, occurring in a wide range of different organisms, consists of so-called papain-like cysteine proteases,[1,2] which, because of common evolutionary history, are structurally similar to the cysteine protease papain found in the papaya plant. Papain-like cysteine proteases expressed in mammals are termed cysteine cathepsins [Greek *κατά* ("down"), *ἐπιζεῖν* ("to boil")]. So far, 11 distinct human cathepsins have been described. These are cathepsins B, C, F, H, K, L, O, S, V, W, and X. Whereas the structures of cathepsins B, C, F, H, K, L, S, V, and X are known, only sequence data are available for cathepsins O and W.

The common fold of all structurally determined cathepsins consists of two domains forming a V-shaped active-site cleft containing a cysteine and a histidine residue (C25 and H163 or C25 and H159 according to numbering in cathepsin L or papain, respectively) as catalytic centers. The cysteine residue is located in the N-terminal, mostly α-helical domain (left-hand "L-domain" according to the standard orientation), at the N-terminal end of an α-helix, while the histidine residue is located in the C-terminal β-barrel domain (right-hand "R-domain" according to the standard orientation).[3]

Cathepsins B and H show both exo- and endopeptidase activity, while cathepsins F, K, L, S, and V are endopeptidases, and cathepsins C and X are exopeptidases. The peptidase behavior of cathepsins O and W is still unknown.[4] Cathepsins are, however, not synthesized as active enzymes to avoid uncontrolled self-digestion of the cell. Instead, they are synthesized as latent proproteins ("procathepsins") that differ from their mature counterparts by an additional N-terminal propeptide fragment, which runs through the active-site cleft along the direction opposite to that of normal substrates. The functions of the propeptide have been thoroughly discussed for cathepsin L and involve[5] (i) inhibition of proteolytic activity, (ii) assistance in folding of the proprotein, (iii) stabilization of the proprotein at neutral and basic pH values, and (iv) intracellular target signalization through covalent attachment of mannose 6-phosphate groups, lysosomes as destination. The

Received: June 15, 2012
Revised: September 12, 2012
Published: September 25, 2012

inhibition of enzyme activity through reverse binding of the propeptide within the active-site cleft is thought to be common to all papain-like proteases. In addition, it has been found for cathepsins B and L that although the propeptide fragments may differ in length and sequence, their overall fold appears to be similar.[5] Based on the presence of certain conserved sequence motifs in the propeptide, procathepsins may be divided into cathepsin L-like and cathepsin B-like species. In particular, cathepsin L-like species present variants of the so-called "ERFNIN" motif,[6] which is absent in cathepsin B-like species. The observed overall diversity in propeptides has been suggested to allow for selective protease inhibition during trafficking to different compartments in the lysosomes.[7] In cathepsin L, the propeptide consists of 96 residues, the first 79 forming a globular domain with three α-helices and the last 17 forming an extended chain that runs across the protein surface.[5] At the border region of these two domains, i.e., around residues M75p−K82p, the propeptide is contacting the active-site cleft (sites S1 and S1′ according to papain nomenclature).[5] Extensive interactions between the propeptide and the mature part of the protein are formed via the so-called propeptide binding loop (PBL, I136−D162) of the mature protein moiety, part of which (F145−I150) forms a two-stranded β-sheet with residues S55p−A59p.[5] Additional stabilization is provided by hydrophobic side chain contacts.[5] The PBL is also involved in the binding of certain classes of inhibitors other than the propeptide; e.g., it was shown that there are favorable binding interactions between negatively charged cathepsin L side chains residing in this loop and positively charged side chains of the p41 fragment of the major histocompatibility complex (MHC) class II-associated invariant chain, a known inhibitor of cathepsins L, V, K, and F.[8] According to molecular modeling investigations concerning the PBL in cathepsin O, such an inhibition is also considered possible for this cathepsin.[8] Note that the PBL is sometimes[8] referred to as the "cover loop".

The stability of cathepsins is crucially dependent on their protonation state. For instance, it has been found that procathepsin L is stable whereas the mature enzyme is unstable at neutral and slightly alkaline pH.[9] Intracellular cathepsins are active in the lysosomes. The acidic condition in these compartments (down to pH 3.8 in mature lysosomes)[9] facilitates cleavage of the propeptide, and subsequent dissociation of this fragment activates the enzymatic function. Cathepsin L is activated at pH ~5.5−6, shows optimal proteolytic activity in this slightly acidic environment, and is irreversibly inactivated at pH <4 due to substantial loss of helical structure,[9] whereas, e.g., cathepsin H has a pH optimum in the almost neutral range (pH 6.5−6.8).[10] The precise mechanism of the proteolytic cleavage of the propeptide in vivo is still unknown.[5] It has been suggested that protein−membrane interactions might influence protease activation in vivo.[11] The N-terminal propeptide region of procathepsin L might be involved in binding to the membranes of microsomes.[12,13] More recent studies[14,15] of cathepsin B draw autocatalytic activation back to enzymatic activity of the proenzyme that is only present after a conformational change in the propeptide, exposing the active-site cleft, has taken place. The conformational equilibrium between the inactive and active proenzyme forms is shifted toward the latter at acidic pH, in line with in vitro studies confirming the weakening of propeptide−protein interactions at low pH values in procathepsins L and B.[16,17] The propeptide conformational change allowing for enzymatic activity also appears to improve the substrate properties of the proprotein.[15] Dissociation of the propeptide from the active-site cleft may be considered a unimolecular step leading to an enzymatically active proenzyme,[7,15] whereas subsequent processing of other procathepsins by the latter active species is considered a bimolecular reaction.[7,18] Such an activation mechanism only works for endopeptidases. Propeptide dissociation is not sufficient to activate exopeptidases; i.e., endopeptidases must be available to activate exopeptidic cathepsins.[19]

The pH dependence of the propeptide conformation and its relation to proenzyme activation were studied in the case of procathepsin L, and it was found that a decrease in the pH to a value of 3 induces a loss of tertiary structure but leaves secondary structure almost unaffected, the resulting protein presenting features of a molten globule.[20] Notably, the propeptide is able to fold independently of the mature enzyme. At pH ~4−6, the isolated propeptide is thus a strong inhibitor of cathepsin L, while inhibitory potency is lost in the more strongly acidic regime.[20]

Although cathepsins play an important role in the management of intracellular protein turnover and necrosis- or autophagocytosis-associated protein breakdown, their function is by no means restricted to lysosomal protein degradation but also involves, e.g., MHC class II-mediated antigen presentation by degradation of antigens (e.g., cathepsin S in dendritic, macrophage, and B cells), collagenolysis in the context of bone resorption (e.g., degradation of type I collagen by cathepsin K in osteoclasts), the processing of protein prohormones (e.g., generation of β-endorphin, α-MSH, or ACTH by cathepsin L in pituitary secretory vesicles), or proneuropeptides (e.g., generation of enkephalin, neuropeptide Y, dynorphins, or cholecystokinin by cathepsin L in neuropeptide-containing secretory vesicles).[3,4,7,21,22] Cathepsin activity, besides being regulated by pH, is also controlled by a number of endogenous enzyme inhibitors, the most important ones being cystatins and thyropins. They act as so-called emergency inhibitors, i.e., stop detrimental protease action, e.g., in the case of cathepsins escaping the cell.[23] The propeptide is a so-called regulatory buffer-type inhibitor; i.e., it stops protease action if there is no substrate to be processed.[23] An imbalance in cathepsin protease action and inhibition is associated with a number of pathological conditions.[7] For instance, cathepsin L, being overexpressed in malignant tumor cells, is thought to advance tumor malignancy by degrading extracellular matrix (collagen, fibronectin, and laminin) and cell−cell adhesion (E-cadherin) proteins.[4] The corresponding action of cathepsin L in the extracellular environment is made possible by a decreased pH value in the vicinity of tumor cells.[4] While the contributions of cathepsin L to tumor progression (invasion and metastasis) have been confirmed by a number of studies, its involvement in tumorigenesis (proliferation and angiogenesis) is still unclear.[4] Similarly, cathepsin B overexpression has been linked to tumor progression.[24] According to experimental tumor models, cathepsin inhibitors have therapeutic anticancer potential but at present lack clinical evidence.[25] Elastolysis and collagenolysis by cathepsins are also considered to be involved in the development and rupture of cerebral aneurysms (cathepsins B, K, and S)[26] and the development of atherosclerosis (cathepsins K and S)[27] through the degradation of the extracellular matrix in artery walls. Cathepsins B and K present in synovial fluid were found to contribute to rheumatoid arthritis through collagenolysis.[28,29] Mutations in cathepsin genes or in the genes of their endogeneous inhibitors cause genetic disorders such as,

Table 1. Abbreviations for the Simulated Proteins Used throughout the Text Along with the Number of Residues, the Net Charge, the Propeptide and PBL Residues, and the Corresponding Simulations[a]

abbreviation	full protein specification	no. of residues	net charge (e)	propeptide	PBL	no. of simulations (length)
	Wild Type					
Lpro	procathepsin L	312	−10	D5p−E96p	I136−D162	5 (6 ns)
Lmat	mature cathepsin L	220	−12	−	I136−D162	5 (6 ns)
Opro	procathepsin O	298	0	D1p−S84p	V138−N161	5 (6 ns)
Omat	mature cathepsin O	214	−4	−	V138−N161	5 (6 ns)
	Mutant					
Opro-D145L	Opro with aspartate-145 mutated to leucine	298	1	D1p−S84p	V138−N161	5 (6 ns)
Opro-G148P	Opro with glycine-148 mutated to proline	298	0	D1p−S84p	V138−N161	5 (6 ns)
Opro-L147P,G148P	Opro with leucine-147 and glycine-148 each mutated to proline	298	0	D1p−S84p	V138−N161	5 (6 ns)
Opro-trunc	Opro with residues 1−38 deleted	260	−2	N39p−S84p	V138−N161	1 (6 ns)
	Alternative Protonation					
Lpro-pH4$^{\text{part1}}$	Lpro with additional protonation of H45p, E48p, E51p, H54p, D65p, E141, E148, E153, and D155	312	−1	D5p−E96p	I136−D162	1 (6 ns)
Lpro-pH4$^{\text{part2}}$	Lpro-pH4$^{\text{part1}}$ with additional protonation of E69p, E70p, D137, H140, E159, D160, and D162	312	6	D5p−E96p	I136−D162	1 (6 ns)
Opro-pH4$^{\text{part1}}$	Opro with additional protonation of H28p, E38p, D145, H153, and H154	298	5	D1p−S84p	V138−N161	1 (6 ns)
Opro-pH4$^{\text{part2}}$	Opro-pH4$^{\text{part1}}$ with additional protonation of E55p, E56p, D139, and E159	298	9	D1p−S84p	V138−N161	1 (6 ns)
Opro-pH4	Opro at pH <4 (Opro with additional protonation of all histidine, aspartate, and glutamate side chains)	298	35	D1p−S84p	V138−N161	1 (6 ns)

[a] Propeptide residue numbers are designated with an additional letter "p", and index 1 is reassigned to the first mature cathepsin residue.

e.g., early onset periodontal disease due to a point mutation in the cathepsin C gene[30] or a certain form of amyloid angiopathy due to a point mutation in the cystatin C gene.[31] See a recent review[7] for a detailed and thorough discussion of the role of cathepsins in physiological and pathological processes.

The present study focuses on the proprotein and mature forms of cathepsins L and O. The tissue-specific expression pattern and localization, role in physiological and pathological processes, enzymatic activity properties, and structure of cathepsin L have been extensively investigated.[4,5,9,32] On the other hand, very little is known about cathepsin O. In 1994, it was cloned from a human breast tumor cDNA library and its amino acid sequence was determined. It seems to be expressed in a wide range of different tissues.[33] However, its function has not been characterized to date. The authors of the present study are aware of only one discussion of structural features of cathepsin O in the literature, where a molecular modeling technique was used to characterize the PBL.[8] This characterization was based on analysis of a single homology-based model structure created using cathepsin L from a cathepsin L−p41 inhibitor complex crystal structure as a template. However, an accurate description of the behavior of (bio-)molecular systems at non-zero temperature requires analysis of an ensemble of structures statistically representative of the phase space sampled by the system under the relevant thermodynamic boundary conditions. Atomistic molecular dynamics (MD) simulation is a powerful tool for investigating molecular structural ensembles, thermodynamics, and dynamics at a resolution often inaccessible to experiment and has previously been used to study, e.g., cathepsins L and K.[34,35]

The present study seeks to provide insight into the structural features of cathepsin O and, in particular, to rationalize at atomic resolution the noncovalent binding interactions of the mature part of the protein with the propeptide in procathepsin O. Its goal is fourfold, namely: (1) creating homology-based models for the proprotein and mature forms of cathepsin O and verifying their stability at a pressure P° = 1 bar, a temperature T^{-} = 298.15 K, and neutral pH by MD simulation, (2) analyzing and comparing structural features of the PBL in the proprotein and mature forms of cathepsins L and O as well as PBL−propeptide interactions in procathepsins L and O by MD simulation [(pro-)cathepsin L was chosen for comparison because experimental structures were determined for both its pro- and mature forms[5,36] and because the sequence of procathepsin O is similar with that of the subclass of cathepsin L-like enzymes], (3) investigating the effect of point mutations and deletions on the stability of PBL−propeptide interactions in procathepsin O using MD simulation, and (4) predicting the effect of pH on the stability of PBL−propeptide interactions in procathepsins L and O using MD simulation.

COMPUTATIONAL DETAILS

Simulated Proteins. All simulations are summarized in Table 1.

(Pro-)Cathepsin L. The initial structure for human procathepsin L (Lpro) was taken from entry 1CJL in the Protein Data Bank (PDB). It is an experimental structure that was determined[5] via X-ray crystallography at a resolution of 2.2 Å. Two modifications to this structure were introduced by the authors of the present study. (i) A missing loop (T175-E176-S177-D178-N179) was inserted from an experimental structure[36] for mature cathepsin L (PDB code 3IV2) after alignment of the two structures using PyMOL.[37] (ii) Residue S25 was changed to C25, E60 to Q60, and A110 to T110 to restore the wild-type sequence as given in ref 38 and as observed in the structure[36] for mature cathepsin L (PDB code 3IV2). Residues Q60 and T110 are located at the protein surface, at considerable distance from the PBL−propeptide interface and the catalytic center, and are not expected to be of immediate relevance to the present study. Furthermore,

structure 1CJL does not provide coordinates for the first four N-terminal propeptide residues. These residues were therefore neglected in the present simulations. The F78pL mutation in the propeptide of structure 1CJL near the active-site cleft was retained in the present simulations. Note also that the catalytic cysteine residue C25 was here represented with a terminal thiol group. Because the thiol pK_a value is in the range of 2.5−3.5, a thiolate group would actually be more appropriate at pH 7. The catalytic histidine H163 was neutral; i.e., N_δ was deprotonated, and N_ε was protonated.

The initial structure for mature human cathepsin L (Lmat) was taken from entry 3IV2 in the PDB.[36] It is an experimental structure that was determined via X-ray crystallography at a resolution of 2.2 Å. Two modifications to this structure were made. (i) The N-terminal missing A1 residue was inserted from the experimental structure[5] for procathepsin L (PDB code 1CJL) after alignment of the two structures using PyMOL.[37] (ii) Residue A25 was changed to C25.

(Pro-)Cathepsin O and Mutants. No experimental structures are available for (pro-)cathepsin O. Therefore, using the known amino acid sequence,[33] a homology-based model for wild-type human procathepsin O (Opro) was created with the SWISS-MODEL server in the "automatic mode".[39] The template was another cysteine protease proenzyme (procaricain from the papaya plant, PDB code 1PCI, structure determined via X-ray crystallography at a resolution of 3.2 Å)[40] which has 28.7% sequence identity with procathepsin O. A corresponding model for mature cathepsin O (Omat) was obtained after removal of the propeptide coordinates.

In the following, simulations of wild-type pro- and mature cathepsin O at pH 7 are denoted Opro and Omat, respectively (Table 1). Residues D145, L147, and G148 appear to be important for hydrogen bonding interactions at the Opro PBL−propeptide interface (PBL−Propeptide Interactions). Furthermore, considering PBL−propeptide interactions in wild-type procathepsin O, the initial ~40 residues encompassing propeptide stretch that does not form an interface with the mature part of the protein appears to be dispensable for noncovalent interfacial contacts. Therefore, to investigate the stability of the PBL−propeptide interface in procathepsin O, four mutant forms were created: (i) a D145L point mutant denoted Opro-D145L representing reduced side-chain hydrogen-bonding capacity, (ii) a G148P point mutant denoted Opro-G148P representing reduced backbone hydrogen-bonding capacity, limited backbone conformational flexibility, and increased steric hindrance within the interfacial β-sheet, (iii) an L147P,G148P double mutant denoted Opro-L147P,G148P representing a more severe loss of backbone noncovalent-binding capacity and conformational freedom than the G148P mutant, and (iv) a deletion mutant denoted Opro-trunc that lacks the first 38 N-terminal (propeptide) residues of wild-type procathepsin O. These systems were created on the basis of a wild-type procathepsin O configuration from simulation Opro (run 4, structure at 5.7 ns) that shows numerous hydrogen-bonding interactions between the PBL and the propeptide. For Opro-D145L, the mutation in this configuration was undertaken by appropriate renaming of atoms. For Opro-G148P and Opro-L147P,G148P, the concerned residues were replaced using the molecular modeling program MOE[41] and subsequent steepest-descent energy minimization[42] with GROMOS11.[43] Except for Opro-trunc, the corresponding solvent configurations were taken from the original wild-type procathepsin O configuration. Because the deletion of 38 residues in Opro-trunc generated a considerable cavity, the Opro-trunc structure was solvated anew in a water box (Molecular Dynamics Simulations). Five simulations of Opro-D145L, Opro-G148P, and Opro-L147P,G148P with different initial counterion positions (Opro-D145L only) and atom velocities (runs 1−5) were performed. Because of the limited relevance for experimental investigations in the context of this study, Opro-trunc was simulated only once.

Alternative Protonation States of Procathepsins L and O. If not indicated otherwise, amino acid protonation states were appropriate for pH 7. Throughout, the N- and C-terminal groups were represented by the charged functionalities NH_3^+ and COO^-. The choice to simulate at neutral pH was motivated by the stability of procathepsin L under such conditions.[9] Although it is known that neutral pH renders mature cathepsin L unstable,[9] this protein was nevertheless simulated at the same protonation state as procathepsin L to limit the change in the system to a single structural difference, namely, the presence or absence of the propeptide. This facilitates the investigation of the effect of the propeptide on the structure and dynamics of the mature part of the protein. The experimental dependence of the stability of (pro-)cathepsin O on pH is at present unknown. To further investigate the stability of the PBL−propeptide interface in procathepsins L and O, the proteins were simulated at three alternative protonation states: (i) an unphysical protonation state, denoted Lpro-pH4$^{\text{part1}}$ and Opro-pH4$^{\text{part1}}$, where all titratable side chains are represented according to pH 7, except for H45p, E48p, E51p, H54p, D65p, E141, E148, E153, and D155 (Lpro-pH4$^{\text{part1}}$) and H28p, E38p, D145, H153, and H154 (Opro-pH4$^{\text{part1}}$) residing in the vicinity of the PBL−propeptide interfacial β-sheet, which are protonated; (ii) an unphysical protonation state, denoted Lpro-pH4$^{\text{part2}}$ and Opro-pH4$^{\text{part2}}$, where E69p, E70p, D137, H140, E159, D160, and D162 (Lpro-pH4$^{\text{part2}}$) and E55p, E56p, D139, and E159 (Opro-pH4$^{\text{part2}}$) are protonated in addition to the ones in Lpro-pH4$^{\text{part1}}$ and Opro-pH4$^{\text{part1}}$ (these residues reside at the PBL−propeptide interface, albeit not in the vicinity of the PBL−propeptide interfacial β-sheet); and (iii) a protonation state appropriate for pH <4, i.e., all aspartate, glutamate, free cysteine, histidine, lysine, and arginine residues protonated, denoted Opro-pH4 (the C-terminus was kept deprotonated). Simulations at alternative protonation states were started from the same protein configurations as the simulations at pH 7 for cathepsin L, and the mutant simulations at pH 7 for cathepsin O. Because of the limited relevance for experimental investigations in the context of this study, all simulations at alternative protonation states were performed only once.

Molecular Dynamics Simulations. All simulations listed in Table 1 were performed using the GROMOS11 package of programs.[43] Proteins were described using the 45A4 parameter set;[44] ions were described using the parameters of ref 45, and water was represented by means of the three-site simple-point-charges (SPC) model.[46] The equations of motion were integrated using the leapfrog scheme[47] with a time step of 2 fs. The solute bond length distances and the rigidity of the water model were enforced by application of the SHAKE algorithm[48] with a relative geometric tolerance of 10^{-4}. The systems were simulated under periodic boundary conditions (PBC) based on rectangular computational boxes. A roto-translational constraint[49] was applied to the protein to avoid spurious contacts between protein copies along short box edges after possible rotational movement. Solute (protein and

Table 2. Parameters Characterizing the Stability of Secondary Structure Evaluated from 6 ns MD Simulations of the Indicated Proteins[a]

protein	run	rmsd$_{bb}$ (maximum) (nm)	N_{bb-hb}	$f_{3_{10}\text{-helix}}$ (%)	$f_{\alpha\text{-helix}}$ (%)	$f_{\pi\text{-helix}}$ (%)	$f_{\beta\text{-sheet}}$ (%)
Lmat	1	0.20 (0.25)	80.41	1.1	22.9	0.1	16.2
	2	0.22 (0.27)	80.82	0.6	22.3	0.4	13.9
	3	0.20 (0.25)	79.03	1.0	23.0	0.1	16.7
	4	0.19 (0.23)	80.35	1.1	23.5	0.1	14.4
	5	0.28 (0.35)	75.90	0.6	22.4	0.7	15.3
	ini.	0	92	5.5	25.0	0.0	18.6
Lpro	1	0.16 (0.22)	129.26	2.7	30.0	0.1	11.4
	2	0.19 (0.28)	126.78	1.9	30.8	0.2	12.7
	3	0.18 (0.24)	123.31	2.2	28.2	0.9	10.6
	4	0.20 (0.29)	125.97	1.9	29.3	0.1	11.7
	5	0.18 (0.23)	127.36	1.9	29.3	0.8	11.9
	ini.	0	148	4.5	31.1	0.0	15.7
Omat	1	0.25 (0.30)	74.64	0.7	19.8	0.7	16.6
	2	0.24 (0.31)	73.42	0.9	18.4	0.9	15.7
	3	0.22 (0.27)	74.54	0.3	19.2	0.0	16.3
	4	0.21 (0.26)	73.23	0.7	18.0	0.8	16.9
	5	0.24 (0.29)	73.70	1.1	19.1	1.2	17.2
	ini.	0	83	7.9	19.6	0.0	17.3
Opro	1	0.32 (0.38)	98.18	1.3	20.5	0.4	13.1
	2	0.34 (0.47)	103.29	1.2	20.6	0.6	12.1
	3	0.31 (0.36)	102.11	0.9	22.0	0.0	12.7
	4	0.32 (0.37)	101.92	1.6	20.4	2.0	13.0
	5	0.31 (0.36)	96.81	1.4	19.8	1.0	13.4
	ini.	0	111	6.7	21.8	0.0	15.1
Opro-D145L	1	0.23 (0.30)	105.43	1.8	20.9	1.3	13.3
	2	0.17 (0.24)	101.64	2.5	21.6	1.1	13.1
	3	0.20 (0.29)	98.90	1.4	19.4	1.1	12.9
	4	0.15 (0.20)	103.30	1.8	21.2	0.6	12.0
	5	0.16 (0.22)	105.69	2.5	21.5	1.8	12.6
	ini.	0	105	2.0	19.5	0.7	12.1
Opro-G148P	1	0.13 (0.17)	100.78	1.4	20.9	1.3	12.9
	2	0.17 (0.22)	96.67	1.9	19.5	1.9	11.9
	3	0.20 (0.27)	101.62	1.3	20.4	1.4	12.7
	4	0.16 (0.23)	99.87	1.6	20.1	1.8	12.8
	5	0.16 (0.23)	97.85	1.5	20.7	1.8	13.0
	ini.	0	106	2.0	19.5	0.7	11.4
Opro-L147P,G148P	1	0.15 (0.21)	99.12	2.1	19.4	2.8	11.7
	2	0.16 (0.20)	97.76	2.2	19.0	2.5	11.4
	3	0.16 (0.20)	98.13	2.2	18.9	1.1	11.6
	4	0.19 (0.26)	99.67	1.9	20.3	2.2	11.1
	5	0.17 (0.24)	98.96	1.8	19.6	1.8	12.0
	ini.	0	105	2.0	19.5	0.7	11.4
Opro-trunc	1	0.20 (0.25)	89.61	1.2	18.0	1.0	15.3
	ini.	0	96	2.3	16.9	0.8	13.8
Lpro-pH4^{part1}	1	0.25 (0.37)	123.68	1.3	29.8	0.3	11.3
	ini.	0	150	4.8	31.1	0.0	15.7
Lpro-pH4^{part2}	1	0.33 (0.45)	126.64	1.9	29.8	0.2	12.0
	ini.	0	150	4.8	31.1	0.0	15.7
Opro-pH4^{part1}	1	0.19 (0.28)	101.68	2.0	19.8	1.9	12.8
	ini.	0	105	2.0	19.5	0.7	12.1
Opro-pH4^{part2}	1	0.34 (0.46)	104.64	1.4	20.5	2.4	12.9
	ini.	0	105	2.0	19.5	0.7	12.1
Opro-pH4	1	0.37 (0.56)	104.29	1.9	19.4	1.5	14.4
	ini.	0	104	2.0	19.5	0.7	12.1

[a] The reported data include the average backbone heavy atom-positional rmsd values from the (minimized) starting structure (rmsd$_{bb}$), evaluated after the backbone heavy atoms had been fit to this structure, as well as the maximal instantaneous rmsd values in parentheses, the average numbers of backbone hydrogen bonds (N_{bb-hb}), and the average fractions $f_{3_{10}\text{-helix}}$, $f_{\alpha\text{-helix}}$, $f_{\pi\text{-helix}}$, and $f_{\beta\text{-sheet}}$ of protein residues occurring in corresponding secondary structure elements. The labels "run 1"−"run 5" pertain to simulations initiated from the same protein starting configurations, albeit with

Table 2. continued

different initial counterion atom positions and different initial atom velocities. Table entries "ini." denotes data corresponding to the initial (minimized) structures. Abbreviations for the protein names are defined in Table 1.

counterions) and solvent degrees of freedom were independently coupled to a heat bath[50] at 298.15 K with a relaxation time of 0.1 ps. The box dimensions were isotropically coupled to a pressure bath[50] at 1 atm, with a relaxation time of 0.5 ps and an isothermal compressibility of 4.575 × 10^{-4} kJ^{-1} mol nm^3.

Van der Waals and electrostatic interactions were handled using the Lennard-Jones potential[51−53] and the Barker-Watts reaction field scheme[54] within a triple-range cutoff approach[55] applied on the basis of distances between charge group centers,[42] with short- and long-range cutoff radii of 0.8 and 1.4 nm, respectively, and an update frequency of five time steps for the short-range pair list and intermediate-range interactions. The reaction field scheme was applied with $\varepsilon_{BW} = 66.6$ for the relative permittivity of the dielectric continuum surrounding the cutoff sphere, as appropriate[56] for the SPC water model.

Prior to the beginning of any simulation, each protein structure was solvated in a rectangular computational box composed of periodically replicated cubic water boxes containing 216 SPC water molecules at the equilibrated density of the SPC water model (972 kg/m^3),[56] allowing for a minimal protein−wall distance of 0.8 nm and a minimal protein−solvent distance of 0.23 nm. The resulting box sizes were on average 5.90 nm × 8.02 nm × 8.80 nm and 5.48 nm × 6.04 nm × 7.39 nm for pro- and mature wild-type cathepsins, respectively, and similar for corresponding protonated or mutant species. After the systems had been relaxed using steepest-descent energy minimization,[42] a neutralizing amount of sodium or chloride counterions (Table 1) was added by replacing randomly chosen water molecules within the bulk solvent. The absence of ions in the vicinity of the protein solute was verified by visual inspection. All simulations were initiated from a random set of atom velocities satisfying a Maxwell−Boltzmann distribution at 50 K, and seven equilibration steps were used to obtain a starting configuration appropriate for a simulation in the *NPT* ensemble at a pressure of 1 atm and a temperature of 298.15 K. In a first simulation period of 20 ps in the *NVT* ensemble at a temperature of 50 K, a harmonic potential with a force constant of 2.5 × 10^4 kJ mol^{-1} nm^{-2} was applied to all solute atom positions to restrain them to the initial configuration. During five succeeding simulation periods of 20 ps in the *NVT* ensemble, the temperature was consecutively increased by 50 K and the restraint force constant was consecutively reduced by a factor of 10 a restraint force constant of 0. Here, the sixth simulation step involved a temperature of 298.15 K rather than 300 K. In the seventh simulation step, the roto-translational constraint and constant-pressure restraint were introduced and kept for the following 6 ns production run, during which the sampled configurations and energetic data were written to file every 0.5 ps for subsequent analysis.

To gain improved statistical certainty of the simulation results, for most of the investigated proteins, five different simulations, denoted runs 1−5, were performed. The equilibration period of the different runs was initiated from the same starting configuration, albeit with different initial counterion positions and atom velocities.

Analysis of Simulation Results. The analyses of simulations described in this section were used to investigate the overall protein secondary structure stability and interactions between the PBL and the propeptide. The definition of the PBL in Lpro and Opro is reported in Table 1.

The stability of a protein during the simulation was assessed by (i) monitoring the time series of the atom-positional root-mean-square deviation (rmsd) of all backbone heavy atoms (N, C_α, and C) from the (minimized) starting structure after a roto-translational fit of all backbone heavy atoms to the reference structure, (ii) evaluating the time average of the number of backbone hydrogen bonds, based on defining a hydrogen bond via a geometric criterion (a maximal hydrogen atom−acceptor distance of 0.25 nm combined with a minimal donor−hydrogen atom−acceptor angle of 135°), and (iii) evaluating the time average of fractions of protein residues occurring in secondary structure elements 3_{10}-helix, α-helix, π-helix, or β-sheet, based on secondary structure definitions according to the Kabsch−Sander rules.[57]

Specific investigations concerning the PBL−propeptide interface were undertaken by (i) assessing the atom-positional root-mean-square fluctuation (rmsf) for all C_α atoms, after a roto-translational fit of all backbone heavy atoms to the first trajectory frame, and focusing in particular on the flexibility of the PBL−propeptide interface, (ii) monitoring the occurrence of hydrogen bonds between the involved residues (Table 1), (iii) evaluating time averages of electrostatic and van der Waals interaction energies between all atoms of system components of interest, and (iv) conducting a cluster analysis based on a conformation-based similarity matrix as described in ref 58, where conformational similarity is estimated by means of the atom-positional rmsd between given sets of backbone heavy atoms. The clustering was conducted for PBL conformations of wild-type cathepsins L and O. Cathepsin configurations for the whole mature part were extracted every 2.5 ps from the trajectories of Lmat, Omat, Lpro, and Opro (runs 1−5). Thus, each cathepsin was characterized by a pool of 24000 conformations sampled from 10 different simulation runs. The rmsd of backbone heavy atoms residing in the PBL was evaluated for all pairs of conformations. The rmsd cutoff for the conformational clustering was set to the lowest rmsd value exhibiting a local probability minimum, i.e., 0.15 and 0.14 nm for clustering of PBL configurations of cathepsins L and O, respectively (Figure S.12 of the Supporting Information), and clusters were defined such that all configurations pertaining to a particular cluster be within the given rmsd cutoff distance of the central member structure of the cluster.

All the analyses described above were conducted with the gromos++ package of programs.[59] PyMOL[37] was used for visualizations.

■ RESULTS AND DISCUSSION

MD Simulations of (Pro-)Cathepsin L. The secondary structure of proteins Lmat and Lpro was largely maintained during all five simulation runs. Table 2 reports the time averages and maximal instantaneous values of the backbone heavy atom-positional rmsd from the (minimized) starting structure. For Lmat and Lpro, the maximal values do not exceed 0.35 and 0.29 nm, respectively, while the corresponding time averages over 6 ns trajectories are in the range of 0.19−

0.28 and 0.16−0.20 nm, respectively. For both proteins, the rmsd value appears to be converged after ∼2 ns (Figure S.1 of the Supporting Information). The (minimized) starting structures of Lmat and Lpro present 92 and 148 hydrogen bonds between backbone atoms, respectively, with 49.1 and 51.3%, respectively, of all residues being involved in 3_{10}-helix, α-helix, π-helix, and β-sheet, while the corresponding time averages over 6 ns trajectories are in the range of ∼37.2−40.8 and ∼41.9−45.6%, respectively (Table 2). The time evolutions of the secondary structure elements are illustrated graphically in Figures S.2 and S.3 of the Supporting Information. The main contributions to the loss of secondary structure in comparison to the starting structures are decreases of 1.9−5.1% in β-sheet content (recurrent disappearance of short β-strands in Lmat and Lpro) and a decrease of 1.8−4.9% in 3_{10}-helix content (disappearance or recurrent disappearance of short helices in Lmat and Lpro).

Homology Model and MD Simulations of (Pro-)-Cathepsin O. *Stability of MD Simulations.* Prior to attempting any structural characterization of (pro-)cathepsin O on the basis of the generated homology-based models, the authors of the present study used MD simulation to verify the structures and to complement them with dynamical information. The secondary structure of proteins Omat and Opro was largely maintained during all five simulation runs. For Omat and Opro, the maximal observed backbone heavy atom-positional rmsds from the (minimized) starting structure do not exceed 0.31 and 0.47 nm, respectively, while the corresponding time averages over 6 ns trajectories are in the range of 0.21−0.25 and 0.31−0.34 nm, respectively (Table 2). The sampled Omat configurations thus show a high degree of conformational similarity to the initial model structure, whereas the deviations observed for Opro are somewhat larger. Significantly lower maximal and average rmsd values are found for the mature part of Opro (the roto-translational fit still being based on all backbone heavy atoms), namely, 0.31 and 0.21−0.23 nm, respectively, suggesting that the difference between Opro and Omat in atom-positional rmsd is caused by the flexibility of the propeptide (data not shown). The structural features of the latter may not be accurately represented by the initial model structure, because it is known that cysteine proteases, although sharing high sequence and structural similarity in the mature part of the enzyme, differ more considerably in terms of length, sequence, and structure of the propeptide.[5] The present Opro simulations suggest in addition that the predominant contributions to the increased rmsd arise from the first 40 residues of the propeptide. These N-terminal residues do not have any interface, i.e., no direct nonbonded interactions, with the mature part of the protein.

Except for run 2 of Opro, where an increase in rmsd is observed after 3 ns, the rmsd of Omat and Opro appears to be converged after ∼2 ns (Figure S.1 of the Supporting Information). The rmsd increase of run 2 of Opro is due to the first 12 N-terminal (no defined secondary structure) propeptide residues, which, after approximately 3 ns, move closer to the protein. This movement is illustrated in Figure S.10 of the Supporting Information and underlines the importance of verifying (homology-based) model structures by molecular dynamics simulation, i.e., of investigating the dynamical behavior of a molecule, which can give valuable insight into possible conformational variability. When residues 1−12 are omitted from the rmsd calculation, the rmsd appears to be converged after ∼2 ns for all Opro runs (Figure S.1 of the Supporting Information). Conceding the high flexibility of the first 12 N-terminal residues, the observation of rmsd convergence during the 6 ns period provides a reasonably trustworthy indication of stable simulations.

The (minimized) starting structures of Omat and Opro present 83 and 111 hydrogen bonds between backbone atoms, respectively, with 44.8 and 43.6%, respectively, of all residues being involved in 3_{10}-helix, α-helix, π-helix, or β-sheet secondary structure elements, while the corresponding time averages over 6 ns trajectories are in the range of ∼35.8−38.6 and ∼34.5−37.0%, respectively (Table 2). The time evolutions of the secondary structure elements are illustrated graphically in Figures S.4 and S.5 of the Supporting Information. The main contributions to the loss of secondary structure in comparison to the (minimized) starting structures are due to a decrease of 5.1−7.6% in 3_{10}-helix content in (disappearance or recurrent disappearance of short helices in Omat and Opro) and a decrease of 0.1−3.0% in β-sheet content (recurrent disappearance of short β-strands in Omat and Opro).

Procathepsin O mutants Opro-D145L, Opro-G148P, and Opro-L147P,G148P show secondary structure stability similar to that of Opro (Table 2 and Figures S.6−S.8 of the Supporting Information). Note that the number of backbone hydrogen bonds and the fractions of helical secondary structure elements in their starting configurations are close to the resulting averages. This is because the starting configurations were constructed from an already equilibrated Opro configuration. The initial β-sheet content of Opro-G148P and Opro-L147P,G148P is lower by 3.7% than that of the Opro starting structure, due to disruption of PBL−propeptide interactions through mutations in the PBL [PBL−Propeptide Interface in (Pro-)Cathepsins L and O], and is also close to the resulting averages. The rmsd of the mutant simulations appears to be converged after ∼2−3 ns (Figure S.1 of the Supporting Information).

After an initial increase in rmsd, simulation Opro-trunc shows a stable rmsd value around 0.2 nm (Figure S.1 of the Supporting Information). The simulations involving the partially protonated Opro species, Opro-pH4$^{\text{part1}}$ and Opro-pH4$^{\text{part2}}$, show slight and considerable increases in rmsd, respectively, throughout the whole 6 ns period. However, simulation Opro-pH4 shows a drastic increase in rmsd after ∼3.5 ns (Figure S.1 of the Supporting Information). The simulations involving the partially protonated Lpro species, Lpro-pH4$^{\text{part1}}$ and Lpro-pH4$^{\text{part2}}$, also show increased rmsd values in comparison to those of Lpro. Simulation Lpro-pH4$^{\text{part1}}$ shows increased rsmd values and fluctuations in comparison to Opro-pH4$^{\text{part1}}$, whereas the rmsd values of Lpro-pH4$^{\text{part2}}$ and Opro-pH4$^{\text{part2}}$ evolve in a similar fashion (Figure S.1 of the Supporting Information). Opro-trunc, Opro-pH4$^{\text{part1}}$, Opro-pH4$^{\text{part2}}$, Lpro-pH4$^{\text{part1}}$, and Lpro-pH4$^{\text{part2}}$ show secondary structure stability similar to that of the corresponding wild-type proteins at pH 7, while Opro-pH4 undergoes a (limited) loss of helical structure in the propeptide (Table 2 and Figure S.9 of the Supporting Information).

Structural Characterization of (Pro-)Cathepsin O. After the stability of Omat and Opro simulations initiated from the homology-based model structures generated for (pro-)-cathepsin O have been confirmed, the protein structures can be characterized. The structure of mature cathepsin O appears to be extremely similar to that of mature cathepsin L. This finding is illustrated in Figure 1, which depicts an overlay of configurations obtained after equilibration of Lmat and Omat.

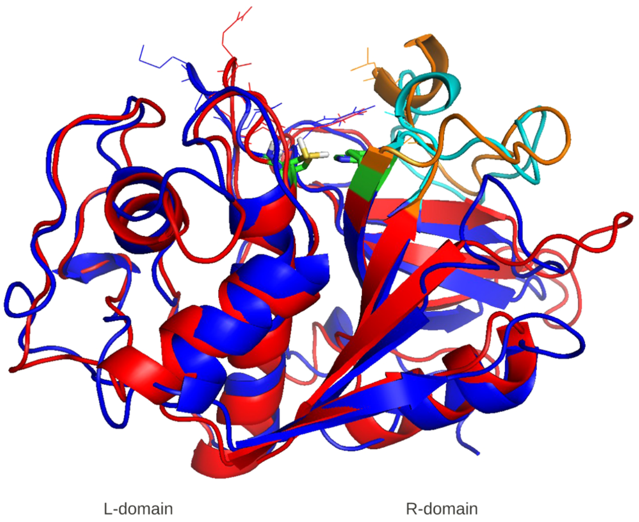


Figure 1. Overlay of the structures of mature cathepsins L (red) and O (blue). The proteins are displayed in cartoon representation and according to the standard orientation, illustrating the left-hand α-helical L-domain and the right-hand β-barrel R-domain.[3] The PBL is colored orange (cathepsin L) and cyan (cathepsin O). The structures correspond to the configuration obtained after the equilibration period of 140 ps of run 1. The catalytic cysteine (C25) and histidine (H163 and H162 in procathepsins L and O, respectively) residues are shown in stick representation, and residues pertaining to the S1 (Q19, Q21, C22, and G23 for cathepsin L and Q19, M21, C22, and G23 for cathepsin O)[21] and S1′ (A138, L144, and W189 for cathepsin L and A140 and W184 for cathepsin O)[21] sites of the active-site cleft are shown in line representation.

The propeptides of procathepsins L and O differ more significantly than the mature parts (Figure 2). While the immediate N-terminal residues in Lpro are involved in an α-helix, a very flexible secondary structure-less N-terminal segment of ~12−15 residues is observed in Opro. The dominating α-helical motif found in procathepsin L [residues E27p−E51p (Figure S.3 of the Supporting Information)] is shorter in procathepsin O (Figure S.5 of the Supporting Information). The remaining secondary structural features of the propeptide, namely, formation of a β-sheet with the PBL [PBL−Propeptide Interface in (Pro-)Cathepsins L and O] and a short helical motif in the spatial vicinity of the active-site cleft, are again very similar in both procathepsins. Obviously, there is pronounced structural conservation between procathepsins L and O in the mature parts of the proteins and in those propeptide parts that are directly contacting the protein. Considering the important biological function of the propeptide, namely, blocking of the active-site cleft, it seems to make sense that the entirely solvent-exposed propeptide region encompassing the N-terminal region before the start of the PBL−propeptide β-sheet (residues D5p−H54p and D1p−S40p in procathepsins L and O, respectively) involves a greater variety in length, sequence, and structure. Presumably, mutations in this initial propeptide stretch that may have accumulated during evolutionary history do not influence biological function. This is also suggested by the fact that the deletion mutant Opro-trunc is not less stable than Opro. Indeed, the first exon of procathepsin O, encoding the first 22 propeptide residues, is believed to have evolved differently from that of other cathepsin L-like enzymes.[60]

Such a variation in propeptide sequences can impair the quality of the homology-based model because a poor sequence alignment can propagate and introduce errors into adjacent regions. Therefore, a multiple-sequence alignment using the MultAlin server[61] in the default mode of human procathepsins K, L, O, S, and V and murine procathepsin O was performed to investigate whether the propeptide stretch forming an interface with the PBL is reasonably positioned in the employed homology-based model (data not shown). The sequence alignment confirmed analogy of residues T57p, M58p, and A59p in human procathepsin L and residues F43p, Y44p, and G45p in human procathepsin O. In addition, analogy of residues F145, Y146, K147, and E148 in human procathepsin L and residues D145, Y146, L147, and G148 in human procathepsin O was confirmed. Thus, the residues forming

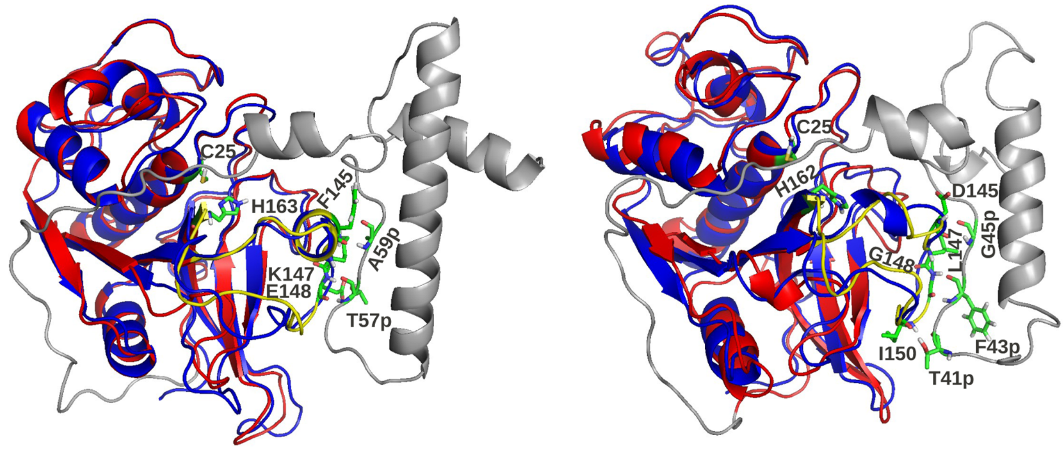


Figure 2. Overlay of the structures of pro- (red) and mature (blue) cathepsins L (left) and O (right). The proteins are depicted in cartoon representation, and their orientations differ from the standard one[3] by a clockwise in-plane rotation of ~45° to better display PBL−protein interactions. The propeptide is colored gray, and the PBL of the procathepsins is colored yellow. The backbones of residues forming a two-stranded antiparallel β-sheet between the PBL and the propeptide in the procathepsins are shown in stick representation (residues T57p, A59p, F145, K147, and E148 in Lpro and residues T41p, F43p, G45p, D145, L147, G148, and I150 in Opro). The catalytic cysteine (C25) and histidine (H163 and H162 in mature cathepsins L and O, respectively) residues are also shown in stick representation. The structures correspond to the configurations obtained after the equilibration period of 140 ps of run 1. Note the absence of hydrogen bonding interactions between the cysteine and histidine side chains. This is a peculiarity of the depicted trajectory snapshot.

Table 3. Occurrence of Hydrogen Bonds between the Propeptide and the PBL in Procathepsins L and O during 6 ns MD Simulations of the Indicated Proteins[a]

	donor		acceptor		occurrence (%)				
protein	residue	atoms	residue	atom	run 1	run 2	run 3	run 4	run 5
Lpro	T57p	N H	E148	O	96.15	96.67	96.92	96.93	96.93
	A59p	N H	F145	O	98.39	96.45	98.51	98.68	98.32
	K147	N H	T57p	O	94.33	93.16	94.26	93.87	93.34
	G77p	N H	D162	O	68.79	77.23	–	–	–
Opro	F43p	N H	G148	O	91.82	86.91	66.00	90.95	81.98
	G45p	N H	D145	O	84.49	53.53	87.18	84.78	51.10
	L147	N H	F43p	O	62.77	94.85	69.58	–	85.09
	G148	N H	F43p	O	72.95	–	84.32	75.25	87.48
	I150	N H	T41p	O	60.49	73.53	–	73.19	–
	S24p	OG HG	D145	OD1	65.43	–	–	–	–
	S24p	OG HG	D145	OD2	55.88	–	–	–	–
	S50p	OG HG	D145	OD2	54.17	–	–	–	–
	N161	ND2 HD22	R63p	O	71.88	–	–	–	–
	S50p	OG HG	D145	OD1	–	93.34	–	70.28	–
	N47p	N H	D145	OD1	–	95.97	–	–	–
	Q48p	N H	D145	OD1	–	75.50	–	–	–
	S50p	N H	D145	OD1	–	71.96	–	72.69	80.44
	S50p	N H	D145	OD2	–	61.88	–	–	–
	Q48p	N H	D145	OD2	–	–	–	71.59	53.63
	F49p	N H	D145	OD2	–	–	–	67.32	58.25
	N47p	N H	D145	OD2	–	–	–	–	55.57
Opro-D145L	F43p	N H	G148	O	77.99	92.33	82.16	77.18	87.04
	G45p	N H	L145	O	87.86	95.02	83.52	78.75	82.75
	L147	N H	F43p	O	81.09	87.02	66.97	96.25	84.15
	G148	N H	F43p	O	54.74	–	77.46	–	62.99
	I150	N H	T41p	O	87.75	69.54	61.93	–	65.25
	R63p	NH1 HH12	N161	OD1	64.59	81.94	77.79	–	–
Opro-G148P	G45p	N H	D145	O	88.89	83.38	85.49	–	–
	L147	N H	F43p	O	94.97	77.04	93.05	77.55	97.12
	I150	N H	T41p	O	–	–	71.64	–	51.73
	N47p	N H	D145	OD1	94.51	97.78	–	74.06	97.36
	Q48p	N H	D145	OD2	96.92	99.45	65.48	74.32	98.62
	F49p	N H	D145	OD2	95.80	96.54	–	79.93	95.71
	S50p	N H	D145	OD1	99.46	99.62	–	83.53	99.54
	S50p	OG HG	D145	OD1	83.94	98.56	–	83.17	98.52
	R63p	NH1 HH12	N161	OD1	82.46	78.62	66.18	91.82	–
	Q48p	N H	D145	OD1	–	–	55.33	–	–
Opro-L147P,G148P	G45p	N H	D145	O	50.21	73.30	79.56	87.28	69.39
	I150	N H	T41p	O	62.96	90.99	81.24	–	54.04
	N47p	N H	D145	OD1	80.86	97.97	99.22	60.45	97.16
	Q48p	N H	D145	OD2	94.23	97.29	99.30	80.28	92.38
	F49p	N H	D145	OD2	96.92	96.47	94.78	61.27	96.09
	S50p	N H	D145	OD1	99.05	99.54	99.45	63.92	99.68
	S50p	OG HG	D145	OD1	99.10	96.52	98.08	54.87	94.77
	R63p	NH1 HH12	N161	OD1	62.34	61.28	–	78.92	–
Opro-trunc	F43p	N H	G148	O	90.17				
	G45p	N H	D145	O	93.19				
	L147	N H	F43p	O	82.88				
	N47p	N H	D145	OD1	80.49				
	Q48p	N H	D145	OD2	80.88				
	F49p	N H	D145	OD2	75.06				
	S50p	N H	D145	OD1	79.71				
	S50p	OG HG	D145	OD1	83.57				
	R63p	NH1 HH12	N161	OD1	67.27				
Lpro-pH4^{part1}	T57p	N H	E148	O	59.03				
	A59p	N H	F145	O	66.65				
	K147	N H	T57p	O	85.25				
Lpro-pH4^{part2}	T57p	N H	E148	O	93.72				

Table 3. continued

protein	donor		acceptor		occurrence (%)				
	residue	atoms	residue	atom	run 1	run 2	run 3	run 4	run 5
	A59p	N H	F145	O	87.74				
	K147	N H	T57p	O	92.97				
	I150	N H	S55p	O	87.57				
Opro-pH4^{part1}	F43p	N H	G148	O	76.63				
	G45p	N H	D145	O	65.21				
	L147	N H	F43p	O	77.76				
	G148	N H	F43p	O	57.20				
	I150	N H	T41p	O	66.87				
	R63p	NH1 HH12	N161	OD1	83.53				
Opro-pH4^{part2}	F43p	N H	G148	O	87.66				
	G45p	N H	D145	O	85.83				
	L147	N H	F43p	O	89.54				
	G148	N H	F43p	O	61.29				
	I150	N H	T41p	O	64.59				
	R63p	NH1 HH12	N161	OD1	83.85				
Opro-pH4	F43p	N H	G148	O	73.41				
	G45p	N H	D145	O	73.55				
	L147	N H	F43p	O	69.27				
	G148	N H	F43p	O	54.73				
	I150	N H	T41p	O	74.02				

[a] Amino acid residues are given in the one-letter code, and donor and acceptor atom specifications are given according to the IUPAC-IUB nomenclature.[64] For donor atoms, the electronegative hydrogen-bound atom is also reported. Here, only backbone hydrogen bonds and hydrogen bonds involving (pro-)cathepsin O residues D145 (if deprotonated) and R63p are reported, the complete data being provided in Table S.1 of the Supporting Information. Hydrogen bonds that occur in <50% of the simulation time are not reported. The labels "run1"−"run5" pertain to simulations initiated from the same protein starting configurations, albeit with different initial counterion atom positions and different initial atom velocities. Abbreviations for the protein names are defined in Table 1.

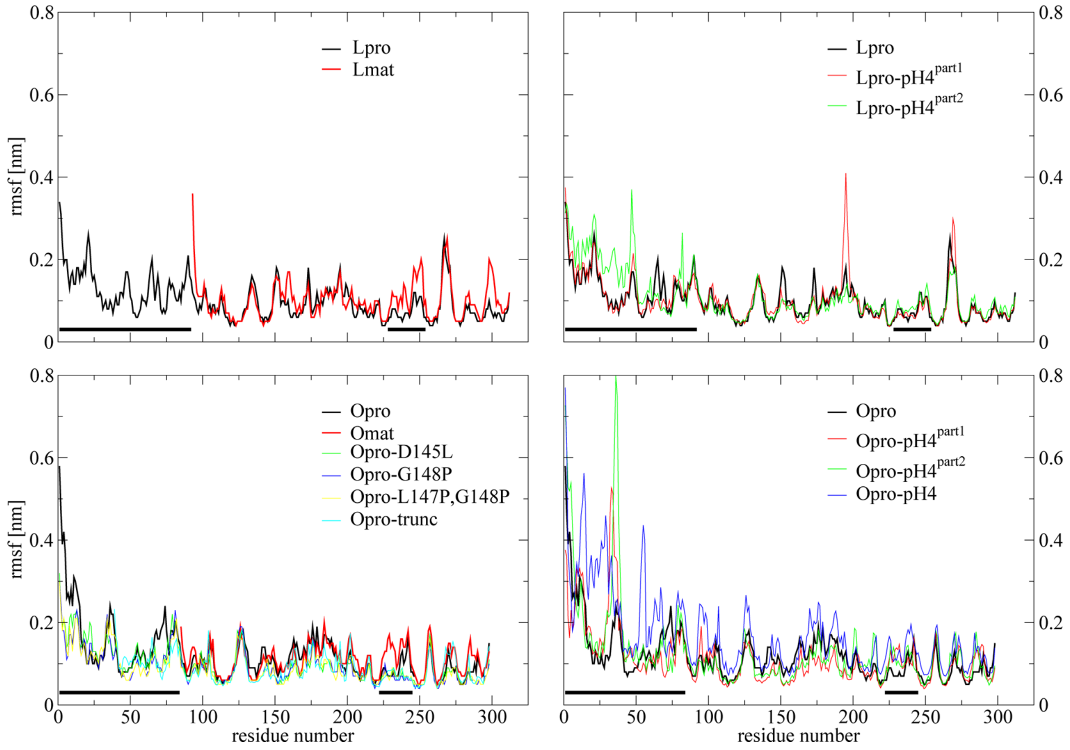


Figure 3. rmsf of C_α atom positions evaluated from 6 ns MD simulations of the indicated proteins after the backbone heavy atoms had been fit to the first trajectory frame. The data are averaged over runs 1−5 for simulations Lpro, Lmat, Opro, Omat, Opro-D145L, Opro-G148P and Opro-L147P, G148P. Black bars denote residues of the propeptide and the PBL, the corresponding residue numbers being reported in Table 1. Abbreviations for the protein names are defined in Table 1.

the interfacial β-sheet in human procathepsins L and O appear at matching positions in the multiple-sequence alignment, the PBL−propeptide hydrogen bonding pattern being exactly equivalent (Table 3), which supports the adequacy of the

this work, multiple sequence alignment of human procathepsins K, L, O, S, V and murine procathepsin O using the MultAlin server[61]																																					
Lpro/Lmat	P	I	S	V	A	I136	D	A	G	H	E	S	F	L	F	Y	K	E	G	I	Y	F	E	P	D	C	S	S	E	D	M	D162	H*	G	V	L	V
Opro/Omat	P	L	V	V	I	V138	D	A	-	-	V	S	W	Q	D	Y	L	G	G	I	-	I	Q	H	H	C	S	S	G	E	A	N161	H*	A	V	L	I
this work, sequence alignment of procathepsins L and O only, using the ClustalW server[62,63]																																					
Lpro/Lmat	P	I	S	V	A	I136	D	A	G	H	E	S	F	L	F	Y	K	E	G	I	Y	F	E	P	D	C	S	S	E	D	M	D162	H*	G	V	L	V
Opro/Omat	P	L	V	V	I	V138	D	A	-	-	V	S	W	Q	D	Y	L	G	G	I	I	Q	H	-	H	C	S	S	G	E	A	N161	H*	A	V	L	I
Ref.[8], multiple sequence alignment of the PBLs of human cathepsins F, K, L, O, S, V, W																																					
Lpro/Lmat	P	I	S	V	A	I136	D	A	G	H	E	S	F	L	F	Y	K	E	G	I	Y	F	E	P	D	C	S	S	E	D	M	D162	H*	G	V	L	V
Opro/Omat	P	L	V	V	I	V138	D	A	V	S	W	Q	D	Y	L	G	G	-	-	-	I	I	Q	H	H	C	S	S	G	E	A	N161	H*	A	V	L	I

Figure 4. Sequence alignment of procathepsins L and O as obtained in this work from a multiple-sequence alignment of human procathepsins K, L, O, S, and V and murine procathepsin O using the MultAlin server[61] or from a sequence alignment of procathepsins L and O only, using the ClustalW server,[62,63] and as reported by ref 8 based on a multiple-sequence alignment of the PBLs of human cathepsins F, K, L, O, S, V, and W. Only residues pertaining to the PBL (gray shaded table entries; residues I136–D162 and V138–N161 in procathepsins L and O, respectively) along with five preceding and five trailing residues are displayed. The full sequence alignment of procathepsins L and O obtained using the ClustalW server[62,63] is displayed in Figure S.13 of the Supporting Information. The catalytic histidine residue (H163 and H162 in procathepsins L and O, respectively) is indicated with an asterisk. Abbreviations for the protein names are defined in Table 1.

employed homology-based model for investigations concerning the PBL–propeptide interface. A second alignment of human procathepsins L and O only, using the clustalw2 program of the ClustalW server in the fast mode,[62,63] delivered the same match in hydrogen bonding residues at the PBL–propeptide interface (Figure S.13 of the Supporting Information).

Note, finally, that the mature parts of each procathepsin do not undergo structural rearrangement upon removal of the propeptide, which justifies construction of a model for mature cathepsin O on the basis of a homology-based model for the proenzyme (Figure 2).

A deficiency in the homology-based initial Opro structure, presumably arising from a histidine to glycine mutation in the active-site cleft of the template protein,[40] is the absence of hydrogen bonding between the side chains of the catalytic cysteine (C25) and histidine (H162) residues. Such hydrogen bonding can, however, be expected to occur in the current simulations, with the cysteine thiol group sulfur atom and histidine N_δ atom functioning as donor and acceptor species, respectively. Indeed, C25–H162 hydrogen bonding, while not present in the trajectory frame depicted in Figure 2, and barely present during Omat runs 1, 3, and 4 and Opro runs 1–4, is regularly established during Omat runs 2 and 5 and Opro run 5, where the thiol hydrogen atom binds to the histidine N_δ atom for 10.0, 11.7, and 24.3%, respectively, of the simulation time. Fractions of similar magnitude are observed in Lmat Runs 1–3 and 5 and Lpro runs 4 and 5, namely 26.2, 16.6, 12.1, 11.1, 14.2, and 19.3%, respectively.

In summary, the structure of the mature part of Opro seems to be very comparable to that of Lpro, as are the regions of the propeptide interacting with the PBL. However, the remaining parts of the Opro propeptide seem to be more conformationally diverse.

PBL–Propeptide Interface in (Pro-)Cathepsins L and O.

PBL and Propeptide Structure. The central constituent of the interface between the propeptide and the PBL is a short two-stranded antiparallel β-sheet, and the involved residues are highlighted in stick representation in Figure 2. Because of tight interactions with the propeptide, the positional root-mean-square fluctuation (rmsf) of PBL C_α atoms in the procathepsins is reduced in comparison to that of the mature cathepsins. For example, the average of C_α rmsfs over residues pertaining to the PBL (Table 1) amounts to 0.08 and 0.09 nm in Lpro and Opro, respectively, and 0.12 and 0.13 nm in Lmat and Omat, respectively (average over runs 1–5). A graphical illustration of C_α atom-positional rmsf values is provided in Figure 3, from which it can also be seen that the PBL of procathepsins appears to be a region relatively more rigid than the remainder of the protein. The first ~21 (Lpro) and ~15 (Opro) N-terminal propeptide residues appear to be especially flexible, in agreement with the observations described in Structural Characterization of (Pro-)Cathepsin O. The prominent α-helical propeptide motif following this flexible stretch is stable in both procathepsins, and the PBL and extended terminal propeptide regions also show similar structural features in both procathepsins (Figure 2).

Figure 4 reports the sequence alignments of the PBLs of procathepsins L and O, as suggested by the MultAlin server,[61] the ClustalW server,[62,63] and ref 8. The residues of the PBL (Table 1) are highlighted with yellow and gray shading in Figures 2 and 4, respectively. Despite a limited degree of sequence similarity (5–9 identical amino acid pairs for a total of 27 and 24 PBL residues in Lpro and Opro, respectively), the overall PBL structure is similar. It involves the central β-strand whose C-terminal end connects after a short bent stretch to the catalytic histidine, and whose N-terminal end connects after a bend that includes a short helical motif to the β-barrel domain of the cathepsin. In (pro-)cathepsin L, this helical motif occurs around residues E141–L144 as a very stable α-helix, with recurrent extensions to residue F145. In (pro-)cathepsin O, the initial model structure shows a 3_{10}-helix for residues A140–S142. This helix is less stable and disappears initially but reappears, however, occasionally as a 3_{10}-helix for residues V141–W143 (Omat, runs 1–3), as well as a rather unstable (Omat, run 4) or very stable α-helix around residues V141–D145 (Omat, run 5) with recurrent extensions to residue A140. It is lost in all five Opro runs (Figures S.4 and S.5 of the Supporting Information). Visual inspection of the homology-based model of ref 8 (Figure 3A therein) also suggests a short helixlike stretch between (in approximately) residues V141 and D145, which is likewise present in the analogous region of the template species (cathepsin L–p41 inhibitor complex).

A conformational clustering of the PBL configurations sampled in (pro-)cathepsin L and O simulations gives similar results (Figures 5 and 6). In all Lpro runs, the PBL appears to be restricted to a cluster with a central member structure that shows an α-helical conformation (E141–F145). The corresponding cluster is, however, also accessed occasionally during Lmat runs 2–4 (Figure 5a). Opro runs 1 and 4 sample mostly the same cluster, whereas the other Opro runs show little overlap in sampled clusters. Only occasionally do Opro runs 1–4 access the predominant cluster of Opro run 5. The latter cluster also appears transiently during Omat runs 1–5 (Figure 5b). None of the Opro runs samples clusters with central

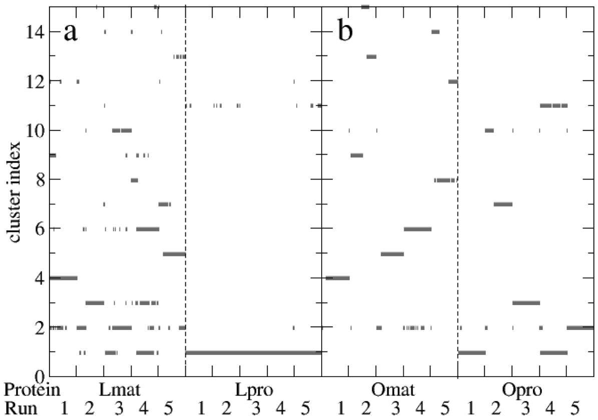


Figure 5. Occurrence of the 15 most populated clusters PBL conformations sampled during 6 ns MD simulations of proteins Lmat, Lpro, Omat, and Opro can be assigned to. The underlying configurations were extracted from both mature cathepsin and procathepsin simulations, with the latter data (Lpro and Opro) appended to the former (Lmat and Omat). The conformational clustering was performed on the basis of PBL rmsd distributions shown in Figure S.12 of the Supporting Information. The labels "run1"–"run5" pertain to simulations initiated from the same protein starting configurations, albeit with different initial counterion atom positions and different initial atom velocities. (a) Cathepsin L (runs 1–5 of simulations Lmat and Lpro). (b) Cathepsin O (runs 1–5 of simulations Omat and Opro).

member structures having a helical conformation in the PBL (considering the first 15 most populated clusters, representing 89.0% of the analyzed configuration data). On the other hand, the predominant cluster sampled during run 5 of Omat shows an α-helix along PBL residues A140–D145. According to the definition of the PBL adopted in this study (Table 1), this short helix starts at the sixth (Lmat and Lpro) or third (Omat) PBL position. The cluster central member structures in Figure 6 also illustrate this shift in position, and all three sequence alignments reported in Figure 4 confirm analogy of the first PBL residues (I136 in Lpro and V138 in Opro), which suggests that the employed PBL definition appears to be reasonable. One may be tempted to conclude that the PBL of (pro-)cathepsin O is conformationally more promiscuous than that of (pro-)cathepsin L. However, such a conclusion is not rigorously justified, because the initial model structure for (pro-)cathepsin O might have provided a PBL conformation that does not adequately correspond to reality, i.e., that does not pertain to the ensemble of conformationally similar PBL conformations that is sampled preferentially under the given thermodynamic boundary conditions (if such a dominant configurational cluster existed). In this view, the ability of the MD simulation to explore different conformational clusters (rather than focusing on a single structure as provided by, e.g., X-ray crystallography) appears to be especially beneficial for the validation of (bio-)molecular model structures. Observation of PBL conformations in (pro-)cathepsin O that were not represented in the initial model structure during the performed MD simulations underlines the importance of conformational sampling. Sampling even more extensive than in the present study would be necessary to draw quantitative conclusions with respect to sampled cluster populations. However, the purpose of the present cluster analysis is rather to qualitatively supplement other indications of increased PBL flexibility in

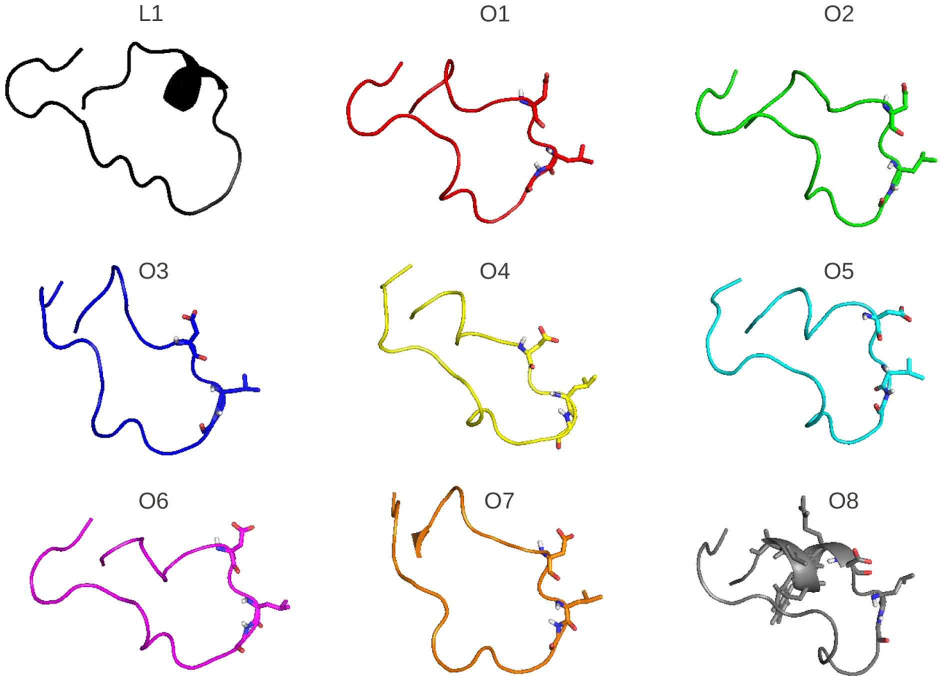


Figure 6. Cartoon representation of the PBL structure of the central member of the most populated cluster (cathepsin L, denoted L1) or of the first eight most populated clusters (cathepsin O, denoted O1–O8). The former cluster encompasses 56.0% of the analyzed configuration data for cathepsin L, and the latter eight clusters encompass 70.5% of the analyzed configuration data for cathepsin O. The first six most populated clusters of cathepsin O encompass 60.8% of the analyzed configuration data, i.e., a fraction similar to the first cluster of cathepsin L. Note the α-helices in the PBL of cathepsin L, central member structure of cluster 1 (E141–F145), and cathepsin O, central member structure of cluster 8 (A140–D145). Residues in the latter helix are shown in stick representation. Cathepsin O residues D145, L147, and G148 are also shown in stick representation. The orientation of the PBL is close to that adopted in Figure 2.

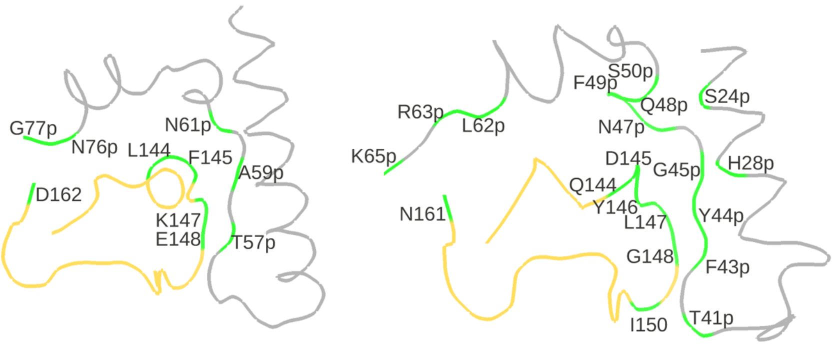


Figure 7. Illustration of important hydrogen bonding residues involved in the PBL−propeptide interface of procathepsins L (left) and O (right). Only the PBL (yellow; residues I136−D162 and V138−N161 in procathepsins L and O, respectively) and a short propeptide stretch (gray; residues W30p−G77p and F21p−K65p in procathepsins L and O, respectively) are shown for the sake of clarity. Amino acid residues are reported in one-letter code. The structures are shown in ribbon representation, the orientation of the PBL being close to that adopted in Figure 2. Residues involved in hydrogen bonding are colored green.

Table 4. Average Electrostatic (E_{ele}) and van der Waals (E_{vdW}) Interaction Energies between All Atoms of the Indicated System Moieties, Evaluated as Averages over Time and the Number of Residues Based on 6 ns MD Simulations of the Indicated Proteins[a]

protein	$E_{\text{ele}}^{\text{PBL−PP}}$ (kJ/mol)	$E_{\text{vdW}}^{\text{PBL−PP}}$ (kJ/mol)	$E_{\text{ele}}^{\text{PBL−S}}$ (kJ/mol)	$E_{\text{vdW}}^{\text{PBL−S}}$ (kJ/mol)	$E_{\text{ele}}^{\text{PP−APO}}$ (kJ/mol)	$E_{\text{vdW}}^{\text{PP−APO}}$ (kJ/mol)	$E_{\text{ele}}^{\text{PP−S}}$ (kJ/mol)	$E_{\text{vdW}}^{\text{PP−S}}$ (kJ/mol)
Lmat	–	–	−176.7	−7.5	–	–	–	–
Lpro	−5.2	−2.9	−157.9	−1.4	−20.1	−10.0	−117.1	−9.5
Omat	–	–	−98.7	−11.4	–	–	–	–
Opro	−6.2	−3.0	−68.2	−6.6	−16.5	−10.6	−105.6	−11.7
Opro-D145L	−4.6	−3.4	−54.7	−6.2	−13.8	−10.8	−100.4	−10.8
Opro-G148P	−8.0	−3.1	−56.2	−6.7	−18.1	−10.4	−101.6	−10.7
Opro-L147P,G148P	−7.9	−3.1	−56.6	−6.4	−18.5	−10.6	−100.9	−10.8
Opro-trunc	−12.6	−3.9	−56.2	−7.7	−33.8	−17.0	−98.3	−13.2
Lpro-pH4$^{\text{part1}}$	−4.4	−3.1	−95.8	−4.9	−24.4	−10.2	−108.7	−10.0
Lpro-pH4$^{\text{part2}}$	−1.4	−3.7	−49.5	−6.4	−23.1	−10.9	−108.1	−9.6
Opro-pH4$^{\text{part1}}$	−4.8	−3.3	−82.4	−5.0	−14.5	−10.4	−98.2	−11.1
Opro-pH4$^{\text{part2}}$	−3.3	−3.0	−58.8	−6.0	−12.0	−10.3	−91.6	−12.6
Opro-pH4	−1.6	−3.3	−48.4	−6.1	−5.1	−11.1	−88.9	−12.8

[a] The reported data include interaction energies between the PBL and the propeptide (PBL−PP), the PBL and the solvent (PBL−S), the propeptide and the remainder (mature part) of the protein (PP−APO), and the propeptide and the solvent (PP−S). Because the propeptide is absent in Lmat and Omat, only PBL−S interaction energies are shown for these proteins. The averaging was done over the number of runs and the number of PBL (PBL−S) or propeptide (PP−APO and PP−S) residues, or the sum thereof (PBL−PP). Residues pertaining to the different system moieties are reported in Table 1. Abbreviations for the protein names are defined in Table 1.

(pro-)cathepsin O as compared to (pro-)cathepsin L [Structural Characterization of (Pro-)Cathepsin O and PBL−Propeptide Interactions]. The fact that a considerable fraction [Omat, run 5 (Figure 5)] of the sampled configurations shows a PBL fold similar to that adopted in the experimental structures of (pro-)cathepsin L should not be surprising because of the overall similar fold, common evolutionary history, and important biological function of the PBL. The absence of the helical conformation in the Opro runs is interesting. It might be purely accidental (lack of sampling) or point to distinct conformational preferences of the PBL in the pro- and mature forms of cathepsin O. Possibly, only particular conformations (e.g., helical) of the PBL in the vicinity of the active-site cleft allow successful substrate binding or/and processing, the required restriction of the PBL to a conformationally less diverse ensemble being entropically unfavorable. Overall, autocatalytic maturation of procathepsin O may therefore not occur as easily as for procathepsin L.

PBL−Propeptide Interactions. As explained in the introductory section, the PBL−propeptide interface of procathepsin L is stabilized by formation of a two-stranded β-sheet and hydrophobic side chain contacts.[5] Interfacial backbone hydrogen bonds and selected side chain-mediated hydrogen bonds of interest to the present study, which are present for at least 50% of the simulation time of the individual 6 ns runs, are reported in Table 3 and illustrated graphically in Figure 7. A complete list of interfacial hydrogen bonds present for at least 50% of the simulation time is provided in Figure S.13 of the Supporting Information. Note that because of the additional T41p−I150 interaction, the β-sheet is on average longer and putatively more stable in procathepsin O than in procathepsin L. The former protein also has more side chain-mediated hydrogen bonds across the PBL−propeptide interface, a very important hydrogen bond involving the carboxylate oxygen atoms of D145, which accepts various hydrogen bonds from propeptide residues (backbone and side chains) with an average occurrence of 69.0%. Procathepsin L does not form such strong electrostatics-mediated interfacial contacts. According to the sequence alignments reported in Figure 4, it has a phenylalanine instead of the deprotonated aspartate (D145) present in procathepsin O, namely, F145 (this work) or F143 (ref 8).

F145 lies at the transition between the short PBL α-helix and the PBL strand of the interfacial β-sheet, and its side chain protrudes between the side chains of propeptide residues N38p and N61p, the average distances between the center of geometry of the aromatic ring carbon atoms and the asparagine nitrogen atoms being 0.48 and 0.47 nm, respectively. F143 is part of the short PBL α-helix, and its side chain protrudes through the center of the loop toward the β-barrel domain of the mature part of the protein rather than toward the propeptide. However, it has been noted before that several other aromatic residues of procathepsin L contribute to PBL−propeptide interactions through formation of a hydrophobic core in the center of the loop.[5] Of particular importance are Y146 and Y151, whose aromatic rings are positioned in the central loop area and face the aromatic ring of propeptide residue F56p,[5] this aromatic clustering also being observed in the present simulation study. The average distances between the centers of geometry of the aromatic ring carbon atoms for the F56p−Y146 and F56p−Y151 interactions are 0.59 and 0.49 nm, respectively. Opro has only one tyrosine residue in the PBL, Y146, which is equivalent to Lpro residue Y146 according to the sequence alignments performed in the present study (Figure 4). Aromatic propeptide residues in the spatial neighborhood of Y146 are F43p and Y44p. The average distances between the centers of geometry of the aromatic ring carbon atoms for the F43p−Y146 and Y44p−Y146 interactions are 0.85 and 0.67 nm, respectively, the larger distances in comparison to Lpro implying weaker interaction energies and greater flexibility of these aromatic side chains.

Although the PBL of procathepsin L involves a considerably higher number of carboxylic acid side chains (D137, E141, E148, E153, D155, E159, D160, and D162) than that of procathepsin O (D139, D145, and E159), none of the former seems to be involved in hydrogen bonding as long-living as observed in procathepsin O. Table 3 also illustrates that several interfacial hydrogen bonds are only present in a subset of runs 1−5. This is especially prominent for Opro and points toward an increased side chain flexibility in the PBL−propeptide interface of Opro in comparison to that of Lpro, which is possibly due to the absence of a hydrophobic core and a reduced charge density.

The interaction energies between different constituents of the system are averaged over the five simulation runs and over the number of residues in the PBL or/and propeptide in Table 4. The stronger PBL−propeptide interaction energies in Opro compared to those in Lpro reflect the stronger interfacial hydrogen bonding in the first protein.

Interactions between the PBL and the solvent differ considerably between (pro-)cathepsins L and O. Because of the absence of the shielding propeptide, for both proteins, PBL−solvent interactions are more pronounced in the mature enzyme. The PBL in (pro-)cathepsin L shows remarkably more favorable solvent interactions than (pro-)cathepsin O, which is certainly due to the abundance of carboxylate groups in the PBL of the former protein (Figure 4). This observation may help to explain the difficulty in maturing procathepsin O under experimental conditions. Fewer titratable groups in the PBL imply a smaller effect of changes in pH on the interaction with the propeptide (see below).

Interactions between the propeptide and the mature part of the protein or with the solvent are of comparable magnitude in procathepsins L and O.

Because of the importance of residues D145, L147, and G148 for hydrogen bonding interactions at the PBL−propeptide interface in Opro, mutants lacking hydrogen bonding capability at these sites were created [(Pro-)Cathepsin O and Mutants]. In particular, mutant Opro-D145L lacks the negatively charged carboxylate group while negligibly affecting side chain packing, and mutants Opro-G148P and Opro-L147P,G148P lack backbone amide hydrogen atoms, the bulky proline ring additionally introducing steric hindrance into the interfacial β-sheet motif. Clearly, the structure of the PBL−propeptide interface is considerably affected in the Opro-L147P,G148P mutant (Figure S.8 of the Supporting Information). Here, the β-strand in the propeptide is absent, whereas that in the PBL still occurs transiently. On the other hand, the β-strand in the propeptide appears entirely intact in most Opro-D145L and Opro-G148P simulations. It seems only slightly destabilized in runs 1 and 4 of Opro-D145L and run 2 of Opro-G148P (Figures S.6 and S.7 of the Supporting Information), which is, however, comparable to the stability of this β-strand in the wild-type simulations, where it is likewise not present continuously in run 2 (Figure S.5 of the Supporting Information).

Electrostatic PBL−propeptide interactions differ between the wild-type and the mutant procathepsin O proteins (Table 4 and Table S.2 of the Supporting Information). The reduced interaction energy for Opro-D145L illustrates the dominant contribution of charge−dipole in comparison to dipole−dipole interactions. The strengthening of electrostatic interactions in the Opro-G148P and Opro-L147P,G148P mutants by ∼1.7−1.8 kJ/mol in comparison to those in Opro might be due to an increased involvement of the charged R63p side chain in interfacial hydrogen bonding. None of the Opro runs exhibits R63p side-chain hydrogen bonding with an occurrence of at least 50% of the simulation time. In contrast, the guanidinium hydrogen atoms of R63p form very stable hydrogen bonds with the side chain carbonyl oxygen atom of N161 in Opro-G148P runs 1−4 and Opro-L147P,G148P runs 1, 2, and 4. However, this interaction is also present in Opro-D145L runs 1−3 (Table 3 and Table S.1 of the Supporting Information).

The deletion mutant Opro-trunc shows stable secondary structure throughout the whole simulation. Most importantly, the formation of the interfacial β-sheet is not affected (Figure S.9 of the Supporting Information), which confirms that the initial ∼40 residues encompassing propeptide stretch does not markedly affect PBL−propeptide binding and thus biological function. Electrostatic PBL−propeptide interactions are enhanced (Table 4 and Table S.2 of the Supporting Information). This is probably caused by an increased presence of interfacial hydrogen bonds involving charged side chains (Table 3).

Secondary structural elements are also stable in the partially protonated procathepsin O species Opro-pH4$^{\text{part1}}$ and less stable in Opro-pH4$^{\text{part2}}$ and the fully protonated procathepsin O species Opro-pH4, which lose helical structure in the propeptide (Figure S.9 of the Supporting Information). Electrostatic PBL−propeptide and PBL−solvent interactions are only marginally altered in Opro-pH4$^{\text{part1}}$, appreciably weakened in Opro-pH4$^{\text{part2}}$, and considerably weakened in Opro-pH4 in comparison to those in Opro, with corresponding van der Waals interactions being largely unaffected (Table 4 and Table S.2 of the Supporting Information). In the case of limited interfacial protonation (Lpro-pH4$^{\text{part1}}$ and Opro-pH4$^{\text{part1}}$), the magnitude of PBL-propeptide interactions is still similar in procathepsins L and O, but more extensive protonation of the PBL—propeptide interface (Lpro-pH4^{part2} and Opro-pH4^{part2}) has a more severe impact on procathepsin L than on procathepsin O. For Lpro-pH4^{part2}, electrostatic PBL—propeptide interactions decrease to 26.9% of the value observed for Lpro, whereas the corresponding fraction amounts to 53.2% for Opro-pH4^{part2}, the larger pH sensitivity of the PBL—propeptide interface in procathepsin L being caused by the increased number of carboxylate groups at the PBL—propeptide interface of procathepsin L (Table 1). Electrostatic interactions between the propeptide and the mature part of the protein differ significantly between the protonated procathepsin L and O species and show a contrary trend in comparison to electrostatic PBL—propeptide interactions. For Lpro-pH4^{part1} and Lpro-pH4^{part2}, they increase to 121.4 and 114.9% of the value observed for Lpro, whereas for Opro-pH4^{part1} and Opro-pH4^{part2}, they decrease to 87.9 and 72.7% of the value observed for Opro. The onset of the strengthening of electrostatic interactions in the protonated procathepsin L species occurs at ~0.85—1.0 ns and can tentatively be correlated with the formation of a salt bridge between the positively charged H54p and negatively charged E176 side chains, which is accompanied by a movement of the loops these residues pertain to.

Although electrostatic propeptide—solvent interactions are almost unaffected by interface protonation, amounting to 92.8 and 92.3% of the values observed for Lpro or 93.0 and 86.7% of the values observed for Opro for limited and extensive interfacial protonation, respectively, the structure of the N-terminal propeptide domain changes notably in Lpro-pH4^{part2}. The main α-helical propeptide motif and the successive short looplike stretch connecting the helix with the interfacial β-strand remain close to the initial structure in Opro-pH4^{part1} and Opro-pH4^{part2} [also reflected by the backbone heavy atom-positional rmsd values (Table 5)]. However, the two regions have average rmsd values of 0.35 and 0.48 nm in Lpro-pH4^{part1} and 0.63 and 0.91 nm in Lpro-pH4^{part2}. In Lpro-pH4^{part2}, after ~0.75 ns, the C-terminal end of the helix (approximately residues Y49p—E51p) starts to unfold and the looplike stretch connecting this end with the interfacial β-strand performs a drastic movement (Figure S.11 of the Supporting Information). In the time period between ~0.73 and ~1.01 ns, the rmsd values of residues E27p—E51p and G52p—H54p increase by ~0.2 and ~0.5 nm, respectively. During the remaining simulation time, the rmsd of residues E27p—E51p increases further by ~0.15 nm because of disruptions or/and movement of the helix, the most striking event (occurring after around 3 ns) being a splitting of the helix into two helices around a small kink formed by residues K40p and M41p (Figures S.9 and S.11 of the Supporting Information). Despite the severe structural rearrangement between the main α-helix of the propeptide and the interfacial β-strand, the overall α-helical character of the N-terminal domain is not lost (Table 2 and Figure S.9 of the Supporting Information). The small first α-helix of the propeptide remains structurally intact but moves with respect to the initial position (Figure S.11 of the Supporting Information).

Experimentally, procathepsin L is activated at slightly acidic pH values, while mature cathepsin L is irreversibly inactivated at pH <4 due to the substantial loss of helical structure.[9] During the simulation of Opro-pH4, a significant increase in the atom-positional rmsd is observed, and it is likely that further structural disruptions occur even after the 6 ns simulation (Figure S.1 of the Supporting Information). Predictions concerning the stability of procathepsin O under these pH conditions would require longer simulations.

Table 5. Average Backbone Heavy Atom-Positional rmsd of Different Propeptide Regions from the (Minimized) Starting Structure after All Backbone Heavy Atoms Had Been Fit to This Structure, Evaluated from 6 ns MD Simulations of the Indicated Proteins[a]

protein	residues	rmsd (nm)
Lpro	D5p—H54p	0.27
	S55p—E96p	0.20
	E27p—E51p	0.17
	G52p—H54p	0.20
Opro	D1p—S40p	0.58
	T41p—S84p	0.36
	E17p—N32p	0.30
	S33p—S40p	0.47
Lpro-pH4^{part1}	D5p—H54p	0.39
	S55p—E96p	0.24
	E27p—E51p	0.35
	G52p—H54p	0.48
Lpro-pH4^{part2}	D5p—H54p	0.64
	S55p—E96p	0.28
	E27p—E51p	0.63
	G52p—H54p	0.91
Opro-pH4^{part1}	D1p—S40p	0.38
	T41p—S84p	0.17
	E17p—N32p	0.23
	S33p—S40p	0.42
Opro-pH4^{part2}	D1p—S40p	0.80
	T41p—S84p	0.22
	E17p—N32p	0.28
	S33p—S40p	0.64

[a] Residues D5p—H54p, S55p—E96p, E27p—E51p, and G52p—H54p of procathepsin L and residues D1p—S40p, T41p—S84p, E17p—N32p, and S33p—S40p of procathepsin O form (i) the N-terminal part of the propeptide until the start of the PBL—propeptide interface, (ii) the C-terminal part of the propeptide after the start of the PBL—propeptide interface, (iii) the main α-helical motif of the propeptide; and (iv) the looplike short propeptide stretch between the main α-helical motif and the start of the PBL—propeptide interface, respectively. For Lpro and Opro, the reported data are averaged over runs 1—5. Abbreviations for the protein names are defined in Table 1.

■ CONCLUSION

The aim of this work was fourfold (points 1—4 in the introductory section), and the conclusions reached in this study can be summarized as follows.

(1) Until now, (pro-)cathepsin O lacked any structural characterization except suggestions concerning the conformation of the PBL on the basis of a single homology-based model structure.[8] To the authors' knowledge, the present study is the first to create a homology-based model for (pro-)cathepsin O and to perform extensive (30 ns) MD simulation to validate and complement the initial model structure through conformational sampling. It was found that the overall fold of (pro-)cathepsin O is very similar to that of (pro-)cathepsin L, except for conformational differences in the initial (~40 residues) stretch of the propeptide that does not interface the mature part of the protein and in a short (~5—6 residues) stretch of the PBL between the interfacial β-strand and the interface with the active-site cleft. The mature parts of both

proteins remain very similar in simulations of the proenzyme and mature species.

The main α-helical propeptide motif is stable in both procathepsins. However, in contrast to Lpro, whose N-terminal propeptide region appeared to sample structurally similar conformations throughout all simulations, the corresponding region of Opro showed more diversity with respect to both the structure (presence or absence of a short N-terminal α-helix) and the orientation to the mature part of the protein. Similarly, a more diverse conformational PBL ensemble was observed for (pro-)cathepsin O than for (pro-)cathepsin L. These results suggest that the propeptide in procathepsin O samples different and possibly more conformations, which may suggest that releasing the free, detached propeptide into the bulk solvent leads to a smaller entropic gain than for procathepsin L. Moreover, the PBL might lose more conformational entropy upon substrate binding in cathepsin O than in cathepsin L, which could disfavor the successful enzymatic action of the former.

(2) The PBL−propeptide interfaces of both procathepsins L and O form a stable two-stranded β-sheet. The β-sheet of the latter protein is on average longer and hence presumably more stable. Additional stabilization of the PBL−propeptide interface seems to be caused by hydrophobic side chain contacts in procathepsin L and charge−dipole interactions in procathepsin O. In particular, the carboxylate oxygen atoms of D145 form numerous long-living hydrogen bonds with hydroxyl hydrogen atoms of serine residues (S24p and S50p) and backbone amide hydrogen atoms (N47p, Q48p, F49p, and S50p) in the propeptide. The stronger interactions between the propeptide and the PBL in procathepsin O could result in less efficient autocatalytic maturation of the protein compared to that of procathepsin L.

(3) Mutations affecting charge−dipole (D145L) and dipole−dipole (L147P and G148P) interactions across the PBL−propeptide interface of procathepsin O suggest that the stability of the interfacial β-sheet is not directly affected by mutations involving side chain-mediated hydrogen bonding in its vicinity (D145L) and likewise not affected upon introduction of a single mutation disrupting the hydrogen bond pattern in the interfacial β-sheet (G148P). However, introduction of two of the latter mutations (L147P and G148P) entails a significant loss of hydrogen bonding, disabling formation of the interfacial β-sheet.

(4) The present simulations suggest that both procathepsin O and the mature enzyme are stable at pH 7. Furthermore, procathepsin O appears to be stable upon protonation of titratable residues at the PBL−propeptide interface, whereas procathepsin L appears to undergo a more pronounced weakening of electrostatic PBL−propeptide interactions and a structural rearrangement at the C-terminal end of the main α-helix in the propeptide and the looplike stretch connecting this helix with the interfacial β-strand. Interfacial protonation thus clearly has larger energetic and structural implications in procathepsin L than in procathepsin O, which is to be expected on the basis of the larger number of carboxylate groups at the PBL−propeptide interface in the former protein.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures and tables illustrating protein stability and PBL−propeptide interactions during the performed simulation runs, a full sequence alignment of procathepsins L and O, and PDB files of procathepsin O and mature cathepsin O structures corresponding to the last trajectory frame of run 5 of Opro and Omat, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*University of Natural Resources and Life Sciences, Institute for Molecular Modeling and Simulation, Muthgasse 18, 1190 Wien, Austria. Phone: +43 1 476548302. Fax: + 43 1 476548309. E-mail: chris.oostenbrink@boku.ac.at.

Funding

Financial support from Project M1281-N17 of the Austrian Science Fund (FWF), Vienna Science and Technology Fund (WWTF) Grant LS08-QM3, and European Research Council (ERC) Grant 260408 is gratefully acknowledged.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank John S. Mort for critical reading of the manuscript.

■ REFERENCES

(1) Barrett, A. J., Rawlings, N. D., and Woessner, J. F., Jr. (1998) *Handbook of proteolytic enzymes*, Academic Press, London.

(2) Brömme, D. (2001) Papain-like cysteine proteases. *Curr. Protoc. Protein Sci. 21.2*, 1−14.

(3) Turk, V., Turk, B., Gunčar, G., Turk, D., and Kos, J. (2002) Lysosomal cathepsins: Structure, role in antigen processing and presentation, and cancer. *Adv. Enzyme Regul. 42*, 285−303.

(4) Lankelma, J. M., Voorend, D. M., Barwari, T., Koetsveld, J., van der Spek, A. H., de Porto, A. P. N. A., van Rooijen, G., and van Noorden, C. J. F. (2010) Cathepsin L, target in cancer treatment? *Life Sci. 86*, 225−233.

(5) Coulombe, R., Grochulski, P., Sivaraman, J., Ménard, R., Mort, J. S., and Cygler, M. (1996) Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *EMBO J. 15*, 5492−5503.

(6) Karrer, K. M., Peiffer, S. L., and Di Tomas, M. E. (1993) Two distinct gene subfamilies within the family of cysteine protease genes. *Proc. Natl. Acad. Sci. U.S.A. 90*, 3063−3067.

(7) Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., and Turk, D. (2012) Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochim. Biophys. Acta 1824*, 68−88.

(8) Mihelič, M., Doberšek, A., Gunčar, G., and Turk, D. (2008) Inhibitory fragment from the p41 form of invariant chain can regulate activity of cysteine cathepsins in antigen presentation. *J. Biol. Chem. 283*, 14453−14460.

(9) Turk, B., Dolenc, I., Lenarčič, B., Križaj, I., Turk, V., Bieth, J. G., and Björk, I. (1999) Acidic pH as a physiological regulator of human cathepsin L activity. *Eur. J. Biochem. 259*, 926−932.

(10) Kirschke, H., Barrett, A. J., and Rawlings, N. D. (1995) Lysosomal cysteine proteases. In *Protein profile* (Sheterline, P., Ed.) Vol. 2, pp 1587−1643, Academic Press, London.

(11) Mason, R. W., and Massey, S. D. (1992) Surface activation of pro-cathepsin L. *Biochem. Biophys. Res. Commun. 189*, 1659−1666.

(12) McIntyre, G. F., and Erickson, A. H. (1993) The lysosomal proenzyme receptor that binds procathepsin L to microsomal membranes at pH 5 is a 43-kDa integral membrane protein. *Proc. Natl. Acad. Sci. U.S.A. 90*, 10588−10592.

(13) McIntyre, G. F., Godbold, G. D., and Erickson, A. H. (1994) The pH-dependent membrane association of procathepsin L is mediated by a 9-residue sequence within the propeptide. *J. Biol. Chem. 269*, 567−572.

(14) Quraishi, O., and Storer, A. C. (2001) Identification of internal autoproteolytic cleavage sites within the prosegments of recombinant procathepsin B and procathepsin S. *J. Biol. Chem. 276*, 8118−8124.

(15) Pungerčar, J. R., Caglič, D., Sajid, M., Dolinar, M., Vasiljeva, O., Požgan, U., Turk, D., Bogyo, M., Turk, V., and Turk, B. (2009) Autocatalytic processing of procathepsin B is triggered by proenzyme activity. *FEBS J. 276*, 660−668.

(16) Fox, T., de Miguel, E., Mort, J. S., and Storer, A. C. (1992) Potent slow-binding inhibition of cathepsin B by its propeptide. *Biochemistry 31*, 12571−12576.

(17) Carmona, E., Dufour, E., Plouffe, C., Takebe, S., Mason, P., Mort, J. S., and Ménard, R. (1996) Potency and selectivity of the cathepsin L propeptide as an inhibitor of cysteine proteases. *Biochemistry 35*, 8149−8157.

(18) Rozman, J., Stojan, J., Kuhelj, R., Turk, V., and Turk, B. (1999) Autocatalytic processing of recombinant human procathepsin B is a bimolecular process. *FEBS Lett. 459*, 358−362.

(19) Dahl, S. W., Halkier, T., Lauritzen, C., Dolenc, I., Pedersen, J., Turk, V., and Turk, B. (2001) Human recombinant pro-dipeptidyl peptidase I (cathepsin C) can be activated by cathepsins L and S but not by autocatalytic processing. *Biochemistry 40*, 1671−1678.

(20) Jerala, R., Žerovnik, E., Kidrič, J., and Turk, V. (1998) pH-induced conformational transitions of the propeptide of human cathepsin L. *J. Biol. Chem. 273*, 11498−11504.

(21) Turk, B., Turk, D., and Turk, V. (2000) Lysosomal cysteine proteases: More than scavengers. *Biochim. Biophys. Acta 1477*, 98−111.

(22) Hook, V., Funkelstein, L., Wegrzyn, J., Bark, S., Kindy, M., and Hook, G. (2012) Cysteine cathepsins in the secretory vesicle produce active peptides: Cathepsin L generates peptide neurotransmitters and cathepsin B produces β-amyloid of Alzheimer's disease. *Biochim. Biophys. Acta 1824*, 89−104.

(23) Turk, B., Turk, D., and Salvesen, G. S. (2002) Regulating cysteine protease activity: Essential role of protease inhibitors as guardians and regulators. *Curr. Pharm. Des. 8*, 1623−1637.

(24) Sloane, B. F., Dunn, J., and Honn, K. V. (1981) Lysosomal cathepsin B: Correlation with metastatic potential. *Science 212*, 1151−1153.

(25) Palermo, C., and Joyce, J. A. (2008) Cysteine cathepsin proteases as pharmacological targets in cancer. *Trends Pharmacol. Sci. 29*, 22−28.

(26) Aoki, T., Kataoka, H., Ishibashi, R., Nozaki, K., and Hashimoto, N. (2008) Cathepsin B, K and S are expressed in cerebral aneurysms and promote the progression of cerebral aneurysms. *Stroke 39*, 2603−2610.

(27) Sukhova, G. K., Shi, G. P., Simon, D. I., Chapman, H. A., and Libby, P. (1998) Expression of the elastolytic cathepsins S and K in human atheroma and regulation of their production in smooth muscle cells. *J. Clin. Invest. 102*, 576−583.

(28) Hou, W. S., Li, Z., Gordon, R. E., Chan, K., Klein, M. J., Levy, R., Keysser, M., Keyszer, G., and Brömme, D. (2001) Cathepsin K is a critical protease in synovial fibroblast-mediated collagen degradation. *Am. J. Pathol. 159*, 2167−2177.

(29) Hashimoto, Y., Kagegawa, H., Narita, Y., Hachiya, Y., Hayakawa, T., Kos, J., Turk, V., and Katunuma, N. (2001) Significance of cathepsin B accumulation in synovial fluid of rheumatoid arthritis. *Biochem. Biophys. Res. Commun. 283*, 334−339.

(30) Toomes, C., James, J., Wood, A. J., Wu, C. L., McCormick, D., Lench, N., Hewitt, C., Moynihan, L., Roberts, E., Woods, C. G., Markham, A., Wong, M., Widmer, R., Ghaffar, K. A., Pemberton, M., Hussein, I. R., Temtamy, S. A., Davies, R., Read, A. P., Sloan, P., Dixon, M. J., and Thakker, N. S. (1999) Loss-of-function mutations in the cathepsin C gene result in periodontal disease and palmoplantar keratosis. *Nat. Genet. 23*, 421−424.

(31) Olafsson, I., and Grubb, A. (2000) Hereditary cystatin C amyloid angiopathy. *Amyloid 7*, 70−79.

(32) Mason, R. W., Green, G. D. J., and Barrett, A. J. (1985) Human liver cathepsin L. *Biochem. J. 226*, 233−241.

(33) Velasco, G., Ferrando, A. A., Puente, X. S., Sánchez, L. M., and López-Otín, C. (1994) Human cathepsin O. *J. Biol. Chem. 269*, 27136−27142.

(34) Shenoy, R. T., Chowdhury, S. F., Kumar, S., Joseph, L., Purisima, E. O., and Sivaraman, J. (2009) A combined crystallographic and molecular dynamics study of cathepsin L retrobinding inhibitors. *J. Med. Chem. 52*, 6335−6346.

(35) Ma, S., Devi-Kesavan, L. S., and Gao, J. (2007) Molecular dynamics simulations of the catalytic pathway of a cysteine protease: A combined QM/MM study of human cathepsin K. *J. Am. Chem. Soc. 129*, 13633−13645.

(36) Adams-Cioaba, M. A., Krupa, J. C., Xu, C., Mort, J. S., and Min, J. (2011) Structural basis for the recognition and cleavage of histone H3 by cathepsin L. *Nat. Commun. 2*, 1−8.

(37) DeLano, W. L. (2002) *The PyMOL molecular graphics system* (http://www.pymol.org).

(38) Joseph, L. J., Chang, L. C., Stamenkovich, D., and Sukhatme, V. P. (1988) Complete nucleotide and deduced amino acid sequences of human and murine preprocathepsin L. An abundant transcript induced by transformation of fibroblasts. *J. Clin. Invest. 81*, 1621−1629.

(39) Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006) The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics 22*, 195−201.

(40) Groves, M. R., Taylor, M. A. J., Scott, M., Cummings, N. J., Pickersgill, R. W., and Jenkins, J. A. (1996) The prosequence of procaricain forms an α-helical domain that prevents access to the substrate-binding cleft. *Structure 4*, 1193−1203.

(41) Chemical Computing Group, Inc. (2009) *The molecular operating environment (MOE)*, version 2009.10 (http://www.chemcomp.com).

(42) van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P., and Tironi, I. G. (1996) *Biomolecular simulation: The GROMOS96 manual and user guide*, Verlag der Fachvereine, Zurich.

(43) Schmid, N., Christ, C. D., Christen, M., Eichenberger, A. P., and van Gunsteren, W. F. (2012) Architecture, implementation and parallelisation of the GROMOS software for biomolecular simulation. *Comput. Phys. Commun. 183*, 890−903.

(44) Schuler, L. D., Daura, X., and van Gunsteren, W. F. (2001) An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem. 22*, 1205−1218.

(45) Reif, M. M., and Hünenberger, P. H. (2011) Computation of methodology-independent single-ion solvation properties from molecular simulations. IV. Optimized Lennard-Jones parameter sets for the alkali and halide ions in water. *J. Chem. Phys. 134*, 144104/1−144104/25.

(46) Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., and Hermans, J. (1981) Interaction models for water in relation to protein hydration. In *Intermolecular Forces* (Pullman, B., Ed.) pp 331−342, Reidel, Dordrecht, The Netherlands.

(47) Hockney, R. W. (1970) The potential calculation and some applications. *Methods Comput. Phys. 9*, 136−211.

(48) Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of *n*-alkanes. *J. Comput. Phys. 23*, 327−341.

(49) Amadei, A., Chillemi, G., Ceruso, M. A., Grottesi, A., and di Nola, A. (2000) Molecular dynamics simulations with constrained roto-translational motions: Theoretical basis and statistical mechanical consistency. *J. Chem. Phys. 112*, 9−23.

(50) Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., di Nola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys. 81*, 3684−3690.

(51) Mie, G. (1903) Zur kinetischen Theorie der einatomigen Körper. *Ann. Phys. 316*, 657−697.

(52) Jones, J. E. (1924) On the determination of molecular fields. I. From the variation of the viscosity of a gas with temperature. *Proc. R. Soc. London, Ser. A 106*, 441−462.

(53) Jones, J. E. (1924) On the determination of molecular fields. II. From the equation of state of a gas. *Proc. R. Soc. London, Ser. A 106*, 463−477.

(54) Barker, J. A., and Watts, R. O. (1973) Monte Carlo studies of the dielectric properties of water-like models. *Mol. Phys. 26*, 789−792.

(55) Berendsen, H. J. C., van Gunsteren, W. F., Zwinderman, H. R. J., and Geurtsen, R. G. (1986) Simulations of proteins in water. *Ann. N.Y. Acad. Sci. 482*, 269−285.

(56) Glättli, A., Daura, X., and van Gunsteren, W. F. (2002) Derivation of an improved simple point charge model for liquid water: SPC/A and SPC/L. *J. Chem. Phys. 116*, 9811−9828.

(57) Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577−2637.

(58) Daura, X., van Gunsteren, W. F., and Mark, A. E. (1999) Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Proteins: Struct., Funct., Genet. 34*, 269−280.

(59) Eichenberger, A. P., Allison, J. R., Dolenc, J., Geerke, D. P., Horta, B. A. C., Meier, K., Oostenbrink, C., Schmid, N., Steiner, D., Wang, D., and van Gunsteren, W. F. (2011) The GROMOS++ software for the analysis of biomolecular simulation trajectories. *J. Chem. Theory Comput. 7*, 3379−3390.

(60) Santamaría, I., Pendás, A. M., Velasco, G., and López-Otín, C. (1998) Genomic structure and chromosomal localization of the human cathepsin O gene (CTSO). *Genomics 53*, 231−234.

(61) Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res. 16*, 10881−10890.

(62) Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) ClustalW and ClustalX version 2. *Bioinformatics 23*, 2947−2948.

(63) Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res. 38*, W695−W699.

(64) IUPAC-IUB Commission on Biochemical Nomenclature (1970) Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemistry 9*, 3471−3479.