J. Mol. Model. **1999**, *5*, 177 – 188

Journal of Molecular Modeling

FULL PAPER

Development and Validation of a Homology Model of Human Cathepsin H Including the Mini-Chain

Annett Fengler and Wolfgang Brandt

Department of Biochemistry and Biotechnology, Institute of Biochemistry, Martin-Luther-University Halle-Wittenberg, Kurt-Mothes-Str. 3, D-06120 Halle (Saale), Germany. Tel.: +49 (345) 5524859; Fax: +49 (345) 5527011;
E-mail: brandt@biochemtech.uni-halle.de

Received: 27 July 1999/ Accepted: 16 August 1999/ Published: 22 September 1999

Abstract Cathepsin H is involved in intracellular protein degradation and is implicated in a variety of physiological processes such as proenzyme activation, enzyme inactivation, hormone maturation, tissue remodeling, and bone matrix resorption. A model of the tertiary structure of the human lysosomal cysteine protease cathepsin H was constructed. The protein structure was built from its amino acid sequence and its homology to papain, actinidin, and cathepsin L for which crystallographic co-ordinates are available. The model was generated using the COMPOSER module of SYBYL.
The position and interaction behavior of the so called mini-chain, the octapeptide EPQNCSAT, to the active-site cleft of cathepsin H could be determined by docking studies. Refinement was achieved through interactive visual and algorithmic analysis and minimization with the TRIPOS force field. The model was found to correlate with observed empirical data regarding ligand specificity. The model defines possible steric, hydrophobic, and electrostatic interactions. We anticipate that the model will serve as a tool to understand substrate specificity and may be used for the development of new specific ligands.

Keywords Cathepsin H , COMPOSER, Mini-chain, Tertiary structure

Running title Structure of human cathepsin H including mini-chain

Introduction

Papain-like cysteine proteases (E.C. 3.4.22.) belong to the papain superfamily [1]. The family comprises proteases of plant, mammalian, parasite, and viral origin. Beside papain the family includes plant proteases, such as chymopapain, caricain, actinidin, aleurain, and the lysosomal cathepsins B, H, L, and S. The cysteine proteases cathepsins B, H, L, and S are involved in the lysosomal protein degradation of mammalian tissues [2] and are very important since they have been implicated in various disease states [3-6]. They are essential in the life cycle in a number of protozoan parasites [5], and are related to *Trypanosama brucei* (sleeping sickness), malaria, and dysentery. Mammalian cysteine proteases, such as cathepsins B, H, and L have been implicated in diseases that involve aberrant protein turnover, e. g.

Correspondence to: W. Brandt

muscular dystrophy [3], bone resorption [7], tumor invasiveness [6], arthritis, and inflammatory diseases [4, 8]. Cathepsin B acts predominantly as a carboxy dipeptidase [2, 9], whereas the cathepsins S and L work as endopeptidases [2]. Beside these well-known activities, the specificity of cathepsin H appears somewhat obscure. The aminopeptidase activity of cathepsin H was probably first studied by Fruton et al. who called the enzyme the thiol-dependent 'leucine aminopeptidase' [10]. Kirschke et al. described rat liver cathepsin H as an 'endoaminopeptidase' because it appears to have both aminopeptidase as well as endopeptidase activities on polypeptide substrates [11]. Takahashi et al. investigated porcine spleen cathepsin H activity in more detail [12]. It was found that substrate peptides are cleaved by aminopeptidase activity. They suggested that the specificity of the enzyme depends primarily on the S1 side chain recognition.

Mature cathepsin H (E.C. 3.4.22.16) consists of three fragments: the N-terminal heavy chain, the C-terminal light chain, and an octapeptide called the mini-chain (EPQNCSAT) [13]. Cleavage of the enzyme between the heavy and light chain is partial and can occur between residues Asn168B and Gly168C (papain numbering).

From the amino acid sequences of rat [14], human [15], and mouse cathepsin H [16] it is evident that the mini-chain originates from the cathepsin H propeptide and is located between propeptide residue Glu76P and Thr83P (propeptide numbering). It has been shown previously that the mini-chain is bound *via* a disulfide bridge to Cys212 of the body of cathepsin H (cathepsin H numbering) [17]. The mini-chain has a definitive role in substrate recognition. The location of the C-terminal carboxyl group of the mini-chain defines the cathepsin H aminopeptidase function. Modeling of a substrate into the active site cleft suggests that the negatively charged C-terminus of the mini-chain acts as an anchor for the positively charged N-terminal amino group of a substrate [18].

An X-ray structure of human cathepsin H (hCatH) is unknown up to now. Recently, Guncar et al. determined the high resolution crystal structure of native porcine cathepsin H (pCatH) at 2.1 Å resolution [18]. However, the co-ordinates for this enzyme were not available over the period of the modeling of the protein structure of hCatH.

In agreement with our recently developed models of the tertiary structures of cathepsins K and S we intend to extend the availability of the three-dimensional structure to human cathepsin H including the so called mini-chain by using a knowledge-based approach incorporated in the COMPOSER suite of programs and refinement by interactive graphics and energy minimization [19, 20].

The knowledge of the three-dimensional structure of human cathepsin H is essential for the understanding of the function and it is a prerequisite for engineering of high specific substrates and inhibitors for this enzyme.

On the basis of the known X-ray structures of selected proteins from the Brookhaven Protein Data Bank (PDB) a model of the tertiary structure of cathepsin H was constructed [21]. In addition, we modeled the structure of the mini-chain situated into the active site cleft of hCatH. Based on the results described by Guncar et al. it was investigated if the orientation of this octapeptide in the site cleft is the same like determined for pCatH to avoid the use of an incorrect structure of the active site of hCatH for intended predictions of specific ligands for this enzyme. Furthermore, it was of high importance to analyze conformational differences and amino acid residue substitutions in the active sites between both hCatH and pCatH.

The calculation of force field based interaction energies for the enzyme-ligand complexes and the results obtained by LEAPFROG [22] should be tested as two relatively independent methods to find correlations between calculated values and experimentally observed affinities between the enzyme and the ligands.

The investigation of the interaction behavior of a specific ligand with cathepsin H with special consideration to the S1 subsite should help to understand the specificity of this enzyme.

Table 1 *Name of the proteins from the PDB used for the modeling of the tertiary structure of cathepsin H*

ID code	Protein
1aec	Actinidin with E-64
1cjl	human Procathepsin L
1fie	human Coagulation Factor XIII
1gec	Glycylendopeptidase with Z-Leu-Leu-Val-Gly
1huc	human Cathepsin B
1lyb	human Cathepsin D with Pepstatin
1pe6	Papain with E-64c
1pop	Papain with Leupeptin
1ppn	Papain
1ppo	Protease Ω
1sht	Trypsin
2act	Actinidin
9pap	Papain
1ppd	2-Hydroxyethylthiopapain

Methods

Knowledge-based model building of cathepsin H

The model of human cathepsin H was generated from its amino acid sequence [15] and currently known high-resolution crystal structures of the homologous enzymes using the COMPOSER program (see Table 1) [20]. The automated procedure permits manual interventions by determination of the structurally conserved regions (SCRs) (Table 2) and selection of the structurally variable regions (SVRs) (Table 3). The minimum sequence identity required as a homologous sequence to the target sequence was set to 30%. The number which provides a measure of the significance of the alignment was chosen to 4. The Needlemann-Wunsch algorithm

Table 2 *Modeling of the structure of human cathepsin H based on SCRs from cysteine proteases*

SCR	Amino acid length	Cathepsin H number	ID code	Start in source protein	Identity (%)
SCR 1	11	1-11	1cjl	1	50
2	47	13-59	1cjl	12	52
3	19	63-81	1cjl	62	53
4	18	83-100	1ppn	79	50
5	9	111-119	1gec	108	36
6	10	120-129	1cjl	118	50
7	11	131-141	1aec	130	46
8	13	142-154	1ppn	139	46
9	13	163-175	1gec	156	77
10	22	181-202	1cjl	182	59
11	14	205-218	1cjl	198	57

and the permutation homology matrix were used in the sequence alignment [23, 22]. The gap penalty was set to 8 [22].

In brief, the modeling procedure contained the following steps: tertiary structures of similar cysteine proteases were superimposed and SCRs were identified, a framework of conserved regions was determined as the mean positions of structurally equivalent Cα atoms. Based on the sequence similarity with cathepsin H the SCRs were selected from the tertiary structures. We used the following parameters to define the SCRs: the maximal distance between equivalent Cα atoms was set to 3.5 Å since the average virtual bond distance between neighboring Cα atoms in a peptide chain is 3.8 Å. If the parameter is changed to a higher value, this may cause a Cα atom from a neighboring residue in a homologue to be included in the construction of the framework for a SCR. The minimum number of residues in a SCR was 3 and the maximal iterations before the SCRs were set to 50. The framework must be redefined with the new sets of equivalent Cα positions if the equivalencies have changed after defining average co-ordinates for a SCR that satisfies residual difference criteria. The limit of the number of such equivalencies was set to 20.

The SVRs were chosen from the database of peptide fragments extracted from the protein data bank in correlation to the end-to-end distance of the SCRs which were already positioned in the framework of the cathepsin H structure [24]. Besides, the program also examined for spatial overlaps (i.e. a validation was performed to check if Cα atoms of the loop fragments are too close to Cα atoms in the non-loop (SCR) region of the model). Subsequently, hydrogen atoms were added and the model of the generated structure was minimized to a convergence of the energy gradient less than 0.01 kcal·mol^{-1}·Å^{-1} using the TRIPOS force field included in the SYBYL / MAXIMIN2 module [22]. The minimization included electrostatic interactions based on Gasteiger-Marsili partial charge distributions using a dielectric constant with a distance dependent function ε = 4r and a non-bonded interaction cut-off of 8 Å [25, 26].

The geometry of the minimized structure was inspected with the program PROSA [27, 28]. The method can be used to identify misfolded structures as well as faulty parts of structural models. PROSA calculates a score for the modeled structure that indicates the quality of the protein structure. A polyprotein was used for the z-score calculation that comprises 230 proteins of known structures with a total length of about 50,000 residues. The conformations of these proteins have a good stereochemistry and many features of protein folds. The set of conformations derived from the polyprotein represents a sample of the conformation space of a given protein. The amino acid sequence of cathepsin H was combined with all conformations in the polyprotein and the energies were calculated. The z-score is derived from the resulting energy distribution.

Moreover, the method constructs an energy graph for the energetic architecture of the protein folds as a function of the amino acid sequence position. It represents the energy distribution of the sequence structure pair in terms of sequence position. In this energy graph positive values point to strained sections of the chain and negative values correspond to stable parts of the molecule.

Finally, the tertiary structure model of cathepsin H was checked with PROCHECK [29]. It produces a Ramachandran diagram and allows the examination of various structural features such as bond lengths and angles, secondary structures, and exposure of residues to the solvent.

Docking procedure

The mini-chain, the octapeptide EPQNCSAT, was docked in the minimized modeled structure of cathepsin H using the program FLEXIDOCK [22].

Furthermore, the docking studies of specific substrates for cathepsin H (Arg-NMec; NMec: N-methylcoumarinylamide) and cathepsin L (Z-Phe-Arg-NMec; Z: benzyloxycarbonyl) [2, 30] were also performed with the program FLEXIDOCK.

Genetic algorithm based Flexible Docking (FLEXIDOCK) provides alternative arrangements of docked ligands into the protein active site. The program allows the ligand as well as the receptor binding pocket to ‘flex’ during docking so that an induced fit can be explored. Briefly, the method contains a genetic algorithm (GA) for altering the ligand, the receptor

Table 3 *The resultant SVRs in the structure model of human cathepsin H*

SVR	Amino acid length	Cathepsin H number	ID code	Start in source protein	Loop homology score (%)
SVR 1	1	12	-	-	-
2	3	60-62	1aec	56	44
3	1	82	1aec	78	30
4	10	101-110	1cjl	97	59
5	1	130	9pap	113	100
6	8	155-162	1cjl	125	33
7	5	176-180	1fie	124	59
8	2	203-204	1aec	169	72
9	2	219-220	1huc	233	82

binding site, and their relative fit and an energy evaluation function for scoring the resulting interactions [31]. In all cases, the ligand is prepositioned into the binding pocket of the enzyme. We marked the following bonds as rotatable: non-terminal single bonds that do not belong to rings or amide bonds as well as the bonds of the amino acid side chains of the binding pocket. The electrostatic interactions between the ligand and the receptor were taken into considerations based on Gasteiger-Marsili partial charge distributions using a dielectric constant of ε = 4r with a distance dependent function. The cut-off distance for non-bonded interactions between the residues was set to 16 Å.

The obtained enzyme-ligand complexes were minimized with the TRIPOS force field for the calculation of the non-bonded interaction energies and for the binding energies between the cathepsins and their ligands.

Calculation of the binding energy

The binding energies of the enzyme-ligand complexes were calculated in two relatively independent ways. First, the non-bonded interaction energies between the enzyme and their ligands within the optimized complex were calculated using the TRIPOS force field. It is clear, it is only an estimation of the real interaction energy between the ligands and the enzyme particularly due to neglected of solvation and desolvation effects. Energy contributions of these effects are approximately calculated using the LEAPFROG program as a second method [22].

Therewith, the binding energy is calculated on a per atom basis and includes a solvation contribution as well as conventional steric and electrostatic terms. Although LEAPFROG was designed to rapidly approximate binding energies to design new ligands, comparisons of binding energy values to experimental measures of activity are encouraging.

In general, the procedure of calculating binding energies using the LEAPFROG program is similar in spirit to that of Goodford's GRID program [32]. It has three major components: the steric, electrostatic, and hydrogen bonding enthalpies of ligand-cavity binding, calculated using the TRIPOS force field, a cavity desolvation energy, and a ligand desolvation energy. The LEAPFROG cavity desolvation energy measures the solvation of the ligand by the cavity. This energy term might be considered as the energy to desolvate the cavity, plus the lipophilic component of the ligand-receptor interaction.

A ligand desolvation energy is provided, as an estimate of the free energy required to remove a ligand from aqueous solution to the vapor phase. Ligand aqueous desolvation is related to an experimental observable, the aqueous activity coefficient or log of the concentration of a compound between its vapor phase and its dilute aqueous solution. The value used in LEAPFROG comes from QSAR measurements. The experimental data underlying this QSAR were taken from literature compilations [33].

For the calculation of the binding energies we carried out the OPTIMIZE mode. Over a number of MOVE cycles the average LEAPFROG binding energy tends to improve. Considering that we performed this method to calculate the binding score and not to generate new ligands we used the following different MOVES kinds: TWIST, FLY, and SAVE. The intent of the FLY move is to seek alternative minimum energy orientations for a ligand. The conformation of the ligand is chosen at random. Each of six rigid body translations and rotations are perturbed by random values uniformly distributed between –2.0 Å and +2.0 Å for translations and –45° and +45° for rotations. With the SAVE option the currently considered ligand will be compared with all previously saved ligands in respect of their energy values. If the binding energy has improved by more than 0.001 kcal·mol^{-1}, the existing database entry file is replaced. The TWIST option resembles a conventional minimizer and as a second step, up to three torsional settings are relaxed along with the six rigid body degrees of freedom, so that internal bump checking as well as docking of the ligand must be a part of the energy calculation.

Results

The three-dimensional structure of human cathepsin H was modeled using the program COMPOSER [20]. In a first step a database of several proteins with known high-resolution crystal structures from the PDB database was created for homology modeling (see Table 1). The identify scores of the primary structure similarity of the selected proteins to cathepsin H are listed in Table 4. The final model contains 11 SCRs of 9-47 residues and 9 SVRs of 1-10 residues (see Tables 2 and 3). A multiple sequence alignment including the SCRs identified by COMPOSER is shown in Figure 1. In addition, the Figure presents the resulting secondary structure elements of the structural model of cathepsin H (cp. Figure 2).

The root-mean square (rms) deviation to the framework for the known three-dimensional structures is less than 0.40 Å. Similarity of the SCRs with the amino acid sequence of cathepsin H was above 46 % with exception of the region of the amino acid residues 111-119, where the similarity was only 36 %. Three disulfide bridges were formed between the cysteine residues Cys23-Cys66, Cys57-Cys99, and Cys157-Cys207. The tertiary structure of cathepsin H including the secondary structure elements of the final model obtained by sequential application of the various procedures described in the Methods section is shown in Figure 2. All these secondary structure elements are included in SCRs. Of all SCRs, 38% amino acid residues adopt conformations of defined secondary structures (cp. Figure 1).

The resulting model of human cathepsin H was checked with PROSA. Figure 3 shows the energy graph calculated from the modeled structure of this enzyme. This energy graph of the observed sequence structure pairs has negative values which corresponds to stable parts of this molecule. The z-score of this structure is –7.79.

The Ramachandran plot for the model of cathepsin H calculated with PROCHECK, shown in Figure 4, revealed good quality stereochemistry. The Φ, Ψ torsion angles of 73 % of the residues had values within the most favored areas and 27 % of the residues had values within additionally allowed regions of the Ramachandran plot. The energy graph and the results from the PROCHECK procedure support the principal correctness of the model of human cathepsin H.

The position of the mini-chain into the active-site cleft of cathepsin H

The octapeptide EPQNCSAT was docked into the binding cleft of the calculated model using FLEXIDOCK.

Structures of related zymogens of procathepsin B [34-36] and procathepsin L [37] have revealed that the propeptide of these cathepsins binds along the active-site cleft in the direction opposite to that of the substrate. However, Guncar et al. described that the mini-chain in porcine cathepsin H (PDB entry: 8pch) binds in the active-site cleft in the direction of a bound substrate with negatively charged carboxylic group of its C-terminal Thr83P attracting the positively charged N-terminus of a substrate (propeptide numbering) (orientation **1**) [18]. The Thr83P binds in the place which is in the related enzyme the S2 binding site, thereby mimicking a substrate P2 residue. For the validation of this approach the mini-chain was also docked rotated to 180° into the active site of cathepsin H whereas the Glu76P occupied the S2 subsite (orientation **2**). In both cases, the mini-chain binds via a disulfide bridge to Cys212 of the mature cathepsin H (cathepsin H numbering). Several complexes of both possibilities were formed with different structures and energetic contents. After minimization of these complexes, a different docking and interaction behavior can be observed.

Table 4 *Identify score of the cathepsin H to the used proteins of the database*

ID code	Identify score %
1cjl	44.1
1gec	40.5
1ppn	41.4
9pap	40.5
1ppd	40.5
1aec	40.0
2act	38.6

Resultant by FLEXIDOCK, the complexes of cathepsin H with the mini-chain with Thr83P in the S2 subsite are mostly about 8 kcal·mol^{-1} energetically preferred compared to the complex with an inverse position of the mini-chain. The resulting most favorable non-bonded interaction energies are -42.87 kcal·mol^{-1} **(1)** and –34.49 kcal·mol^{-1} in the case the mini-chain is rotated 180° **(2)**, respectively. The differences of binding energies for both possibilities obtained using LEAPFROG are in correlation to the calculated non-bonded interaction energies (binding scores: -32.94 kcal·mol^{-1} for **(1)** and –25.61 kcal·mol^{-1} for **(2)**).

Based on the interactions of the mini-chain with the model of the tertiary structure of cathepsin H we can explain the different affinity of this octapeptide with the enzyme when the position of this molecule in the binding pocket of cathepsin H changes. The mini-chain **(1)** binds into the active-site cleft through the side chains of Gln78P, Cys80P, Ser81P, and Thr83P (see Figure 5a). Attractive interactions can be formed between the methylen groups of the Gln78P side chain to the residue Ile118 of hCatH as well as between the Thr83P side chain and Val164 (cathepsin H numbering). The side chains of the residues Ala82P, Asn79P, and Pro77P point away from the bottom of the active-site cleft of cathepsin H. Moreover, several hydrogen bonds between the residues of the mini-chain Glu76P, Gln78P, the negatively charged C-terminus Thr83P and the enzyme contribute to the energetic stabilization of the mini-chain into the cleft of cathepsin H. No main chain interactions with the underlying enzyme surface, other than Thr83P are observed.

The interaction behavior to the active site is clearly diminished if the mini-chain in cathepsin H is rotated 180° **(2)** (see Figure 5b). Attractive hydrophobic interactions can only be observed between Pro77P and the side chain of Val164. The tendency to hydrogen bonds formation between the cathepsin and the mini-chain is reduced. Only one hydrogen bond can be detected between the carbonyl oxygen atom of Ser81P and the side chain of Asn115.

Comparative investigations to pCatH

During our studies on human cathepsin H the data of the X-ray structure of porcine cathepsin H were not yet available so that we could not use this enzyme for the generation of the model. The sequence homology of hCatH to pCatH is 91%. There are 18 different amino acid residues. Mostly these residues are on the surface of the cathepsin and have no influence on the active site characteristics.

The backbone of the catalytic triad is found at the positions usual for a papain-like enzyme. But there is one key difference between the crystal structure and the modeled structure of cathepsin H. The imidazole ring of the active His159 (pCatH) does not form the thiolat-imidazolium ion pair with Cys25 typical for all other known structures of cysteine proteases. Its Nε proton is involved in a salt bridge with the C-terminal carboxyl group of Val212 from a neighboring molecule in the crystal. In the case of our model of human cathepsin H the orientation of the active site residues is suitable to form the right ion pair state (see Figure 1). The dihedral angle χ_1 of His159 is rotated about 90° in our model in comparison to that one in the X-ray structure of porcine cathepsin H. The distances Cys26(S$^\gamma$)···His166(Im) (3.78 Å) and His166(ImH$^+$)···Asn186(C=O) (3.37 Å) in the modeled structure are in the range observed in X-ray structures of cathepsins B, K, papain, and actinidin (PDB entries: 1huc, 1mem, 9pap, 1aec).

```
hCatH YPPSVDWRKKGNFVSPVKNQGACGSCWTFSTTGALESAIAIATGKMLSLAEQQLVDCAQDFNNYGCQGGL
1cjl  APRSVDWREKG-YVTPVKNQGQCGSSWAFSATGALEGQMFRKTGRLISLSEQNLVDCSGPEGNEGCNGGL
1gec  LPESVDWRAKG-AVTPVKHQGYCESCWAFSTVATVEGINKIKTGNLVELSEQELVDCDLQ--SYGCNRGY
1ppn  IPEYVDWRQKG-AVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNEYSEQELLDCDRR--SYGCNGGY
9pap  IPEYVDWRQKG-AVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRR--SYGCNGGY
1ppd  IPEYVDWRQKG-AVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRR--SYGCNGGY
1aec  LPSYVDWRSAG-AVVDIKSQGECGGCWAFSAIATVEGINKIVTGVLISLSEQELIDCGRTQNTRGCNGGY
2act  LPSYVDWRSAG-AVVDIKSQGECGGCWAFSAIATVEGINKITSGSLISLSEQELIDCGRTQNTRGCDGGY
      -------HHHH--------------HHHHHHHHHHHHHHHHHH-------HHHHHHH-------HHH---

hCatH PSQAFEYILYNKGIMGEDTYPYQGKDGYCKFQPGKAIGFV-KDVANITIYDEEAMVEAVALYNPVSFAFE
1cjl  MDYAFQYVQDNGGLDSEESYPYEATEESCKYNPKYSVAND-AGFVDIPK-QEKALMKAVATVGPISVAID
1gec  QSTSLQYVAQN-GIHLRAKYPYIAKQQTCRANQVGGPKVKTNGVGRVQSNNEGSLLNAIA-HQPVSVVVE
1ppn  PWSALQLVAQY-GIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNEGALLYSIA-NQPVSVVLE
9pap  PWSALQLVAQY-GIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGALLYSIA-NQPVSVVLQ
1ppd  PWSALQLVAQY-GIHYRNTYPYEGVQRYCRSREKGPYAAKTDGVRQVQPYNQGALLYSIA-NQPVSVVLQ
1aec  ITDGFQFIINNGGINTEENYPYTAQDGECNVDLQNEKYVTIDTYENVPYNNEWALQTAVT-YQPVSVALD
2act  ITDGFQFIINDGGINTEENYPYTAQDGDCDVALQDQKYVTIDTYENVPYNNEWALQTAVT-YQPVSVALD
      --HHHHHHHHH-----------------------------------------HHHHHHHHH---EEEE--

hCatH VT-QDFMMYRTGIYSSTSCHKTPDKVNHAVLAVGYGEKNGIPYWIVKNSWGPQWGMNGYFLIERGK----
1cjl  AGHESFLFYKEGIYFEPDCSSE--DMDHGVLVVGYGFESNK-YWLVKNSWGEEWGMGGYVKMAKDRR---
1gec  SAGRDFQNYKGGIFEGSCGT----KVDHAVTAVGYGKSGGKGYILIKNSWGPGWGENGYIRIRRASGNSP
1ppn  AAGKDFQLYRGGIFVGPCGN----KVDHAVAAVGYGPN----YILIKNSWGTGWGENGYIRIKRGTGNSY
9pap  AAGKDFQLYRGGIFVGPCGN----KVDHAVAAVGYGPN----YILIKNSWGTGWGENGYIRIKRGTGNSY
1ppd  AAGKDFQLYRGGIFVGPCGN----KVDHAVAAVGYGPN----YILIKNSWGTGWGENGYIRIKRGTGNSY
1aec  AAGDAFKQYSSGIFTGPCGT----AIDHAVTIVGYGTEGGIDYWIVKNSWDTTWGEEGYMRILRNVGGA-
2act  AAGDAFKQYASGIFTGPCGT----AVDHAIVIVGYGTEGGVDYWIVKNSWDTTWGEEGYMRILRNVGGA-
      ----HHHH---------------------EEE--EEEE--EEEE--------------------------

hCatH NMCGLAACASYPIPLV
1cjl  NHCGIASAASYPTV--
1gec  GVCGVYRSSYYPIKN-
1ppn  GVCGLYTSSFYPVKN-
9pap  GVCGLYTSSFYPVKN-
1ppd  GVCGLYTSSFYPVKN-
1aec  GTCGIATMPSYPVKY-
2act  GTCGIATMPSYPVKY-
      -------EEEE-----
```

Figure 1 *Multiple sequence alignment of hCatH with cathepsin L (1cjl), glycylen-dopeptidase (1gec), papain (1ppn, 9pap), 2-hydroxy-ethylthiopapain (1ppd), and actinidin (1aec, 2act). The amino acid sequences have been extracted from the Swiss Prot data bank (hCatH, accession no. P09668) and their related X-ray structures. The SCRs are highlighted in grey, the residues forming the catalytic triad are marked bold (consider in the X-ray structure of cathepsin L the active site residue Cys has been mutated to Ser). The secondary structure elements are coloured red (E=β-sheet, H=α-helix)*

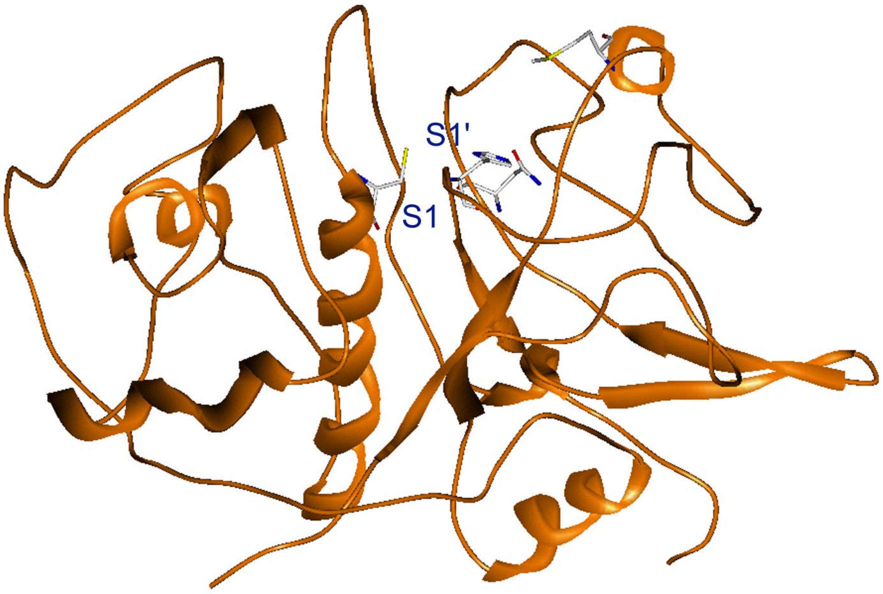

Figure 2 *Representation of the tertiary structure model of human cathepsin H. The S1' and S1 of the active site subsites are labelled. The amino acid residues of the catalytic triad and the amino acid residue Met145 are displayed. This residue is the only one which is different in the active site cleft in comparison to pCatH (Leu145).*

Moreover, we have to consider that there are also differences in the S1' subsite. The amino acid residues Trp188, Val140, Met145, His166, Asn165 form the S1' subsite of hCatH (hCatH numbering). In pCatH Met145 is substituted to Leu145 whereby the hydrophobicity of the S1' subsite will be changed. Based on the varied orientation of the side chain of Leu145 which is less exposed in comparison to the Met145 side chain attractive hydrophobic interactions to a ligand can be expected to be reduced.

Furthermore, we can observe a strong correlation between our modeled protein structure of hCatH and the X-ray structure of pCatH. The low rms deviation of all backbone atoms between hCatH and pCatH of 2.24 Å supports the principal correctness of the model of human cathepsin H determined using COMPOSER.

In a next step to evaluate the first model we used only the later available co-ordinates of the X-ray structure of pCatH as template to generate another tertiary structure model of

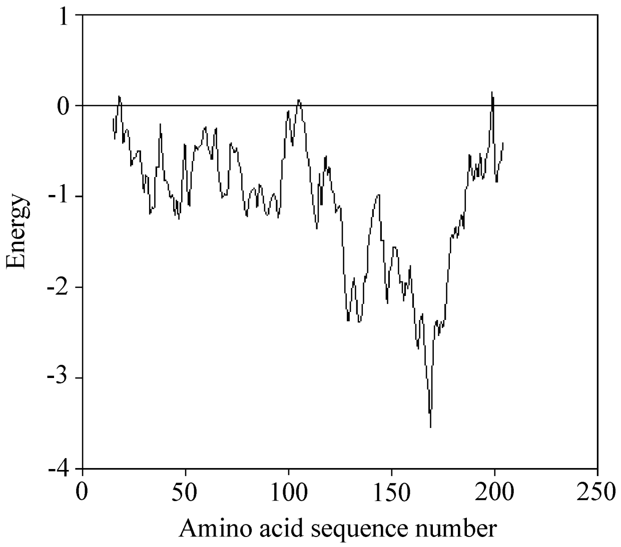

Figure 3 *Energy graph of the tertiary structure model of cathepsin H. The graph is smoothed by a window size of 30 residues. In this energy graph negative values correspond to stable parts of the molecule.*

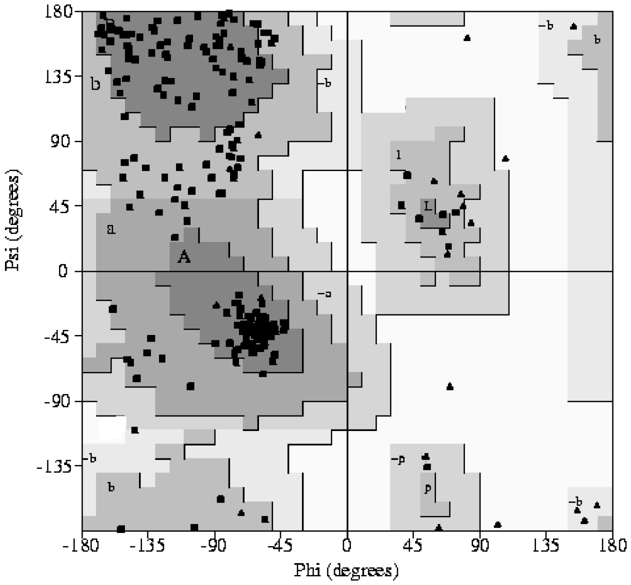

Figure 4 *Ramachandran plot of the predicted structure of cathepsin H. The good stereochemical quality is shown by the presence of 73 % of the residues in the most favoured regions.*

hCatH. The rather small deviations between the first model generated without knowledge of the structure of pCatH and the model based on the X-ray structure of pCatH (rms deviations are 2.26 Å and 1.17 Å, considering all atoms except the hydrogen atoms, and amino acid residues of the secondary structure elements only, respectively) indicates the high quality of the model discussed in this paper.

Substrate specificity of cathepsin H compared to cathepsin L

For the examination of the modeled tertiary structure of cathepsin H with the mini-chain within the active-site cleft we studied the substrate specificity in comparison with cathepsin L. For this purpose, Arg-NMec was used as a specific substrate for cathepsin H because Barrett and Kirschke found that cathepsin H hydrolysis this substrate more efficiently as cathepsin L [2]. This substrate was docked into the binding pocket of cathepsin H using FLEXIDOCK. For cathepsin L the high-resolution crystal structure was used (PDB entry: 1cjl) [37]. The propeptide of cathepsin L was removed and the mutations occurring in comparison to the mature sequence have been replaced by the correct amino acid residues. Appropriated side chain conformations where obtained by alignment with the original X-ray structure and energy optimization of these residues. Furthermore, a loop (Thr271 to Asn275, cathepsin L numbering) is missing in the PDB-structure. This loop has been formed using the LOOP-SEARCH option of SYBYL and the energetically more favored one was used for further considerations.

After minimization of all obtained enzyme ligand complexes the resulting structures were checked in their interaction behavior considering non-bonded interaction energies and binding energies (see Table 5). In comparison with cathepsin L the interaction of the substrate Arg-NMec with cathepsin H is significantly energetically preferred. The lost of the affinity of the ligand to cathepsin L is expressed by the reduced interaction energy of this complex compared to cathepsin H (about 10 kcal·mol^{-1}) (Table 5). Furthermore, the position and conformation of the substrate docking in both cathepsins show differences. The arginine in the P1 position

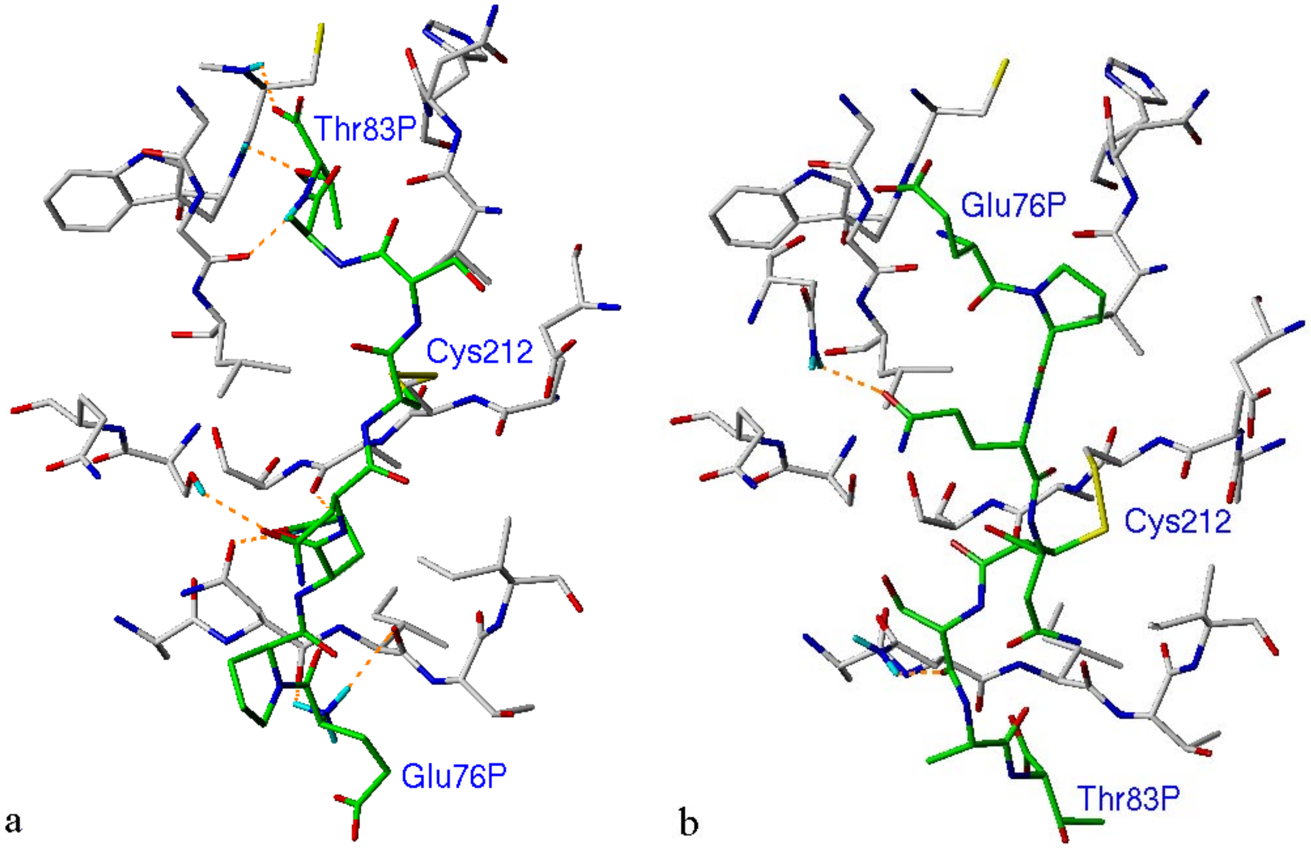

Figure 5 *Illustration of the model of cathepsin H including the mini-chain: (a) The mini-chain binds within the active-site cleft in the direction of a bound substrate* ***(1)****. The negatively charged carboxylic group of its C-terminal Thr83P binds into the S_2 binding site of cathepsin H. (b) The mini-chain is rotated 180° in the active site of cathepsin H, whereas the Glu76P occupies the S_2 subsite* ***(2)****. Only the hydrogen atoms involved in hydrogen bonds are displayed. The carbon atoms of the mini-chain are highlighted bold.*

has a varied docking behavior in both cathepsins. For cathepsin H a number of interactions has been detected to the S1 subsite (Gly68, Gly69) as well as to Gly24 (cathepsin H numbering) (see Figure 6). Hydrogen bonds were formed between the Arg side chain to the carbonyl oxygen atom of Cys66 and from the carbonyl oxygen atom of the backbone of Arg to the Gln20 side chain of cathepsin H which forms the oxyanion hole of hCatH. The formation of the hydrogen bond between the positively charged N-terminus of this substrate and the negatively charged C-terminus of the mini-chain is very important for further stabilization of the enzyme-ligand complex.

In the case of cathepsin L, the attractive hydrophobic interactions of the substrate residue Arg in the P1 position are diminished (Gly67) (cathepsin L numbering). For the tendency to form hydrogen bonds the same statement is valid.

In both cathepsins attractive hydrophobic interactions occur between the phenyl ring of the leaving group NMec with the side chain of Trp188 (hCatH numbering). In cathepsin H an additional interaction can be formed between this group and the side chain of Met145. In cathepsin L Leu144 is located in this position. However, the side chain of this amino acid residue is rotated to 90° compared to the Met145 side chain of hCatH. Therefore, the hydrophobic interaction between the aromatic ring of the substrate leaving group to Leu144 is reduced. These results explain why the substrate Arg-NMec has a stronger affinity to cathepsin H. This effect is also reflected by the calculated non-bonded interaction energies and binding energies (Table 5).

Moreover, Tchoupé et al. have studied a high specific substrate for cathepsin L (Z-Phe-Arg-NMec) [30]. With the investigations of the docking and interaction behavior of this ligand in comparison with ArgNMec in cathepsin L we tried to explain the decrease of the affinity of the Arg-NMec substrate into the active site of cathepsin L. We calculated several complexes of cathepsin L with the substrate Z-Phe-Arg-NMec (see Methods section). Based on the resulting attractive interactions of cathepsin L the different affinities of both substrates (Arg-NMec and Z-Phe-Arg-NMec) to cathepsin L (see also Table 5) can be explained. Stable hydrophobic interactions can be detected between the aromatic ring of the Z group and of the side chain of phenylalanine of the substrate to the side chains of Leu69, Ala214, Ala135, and Met70 that form the S2 subsite (see Figure 7). The side chain of arginine in the P1 position is oriented to the conserved amino acid residues of the S1 subsite. Between the leaving group NMec

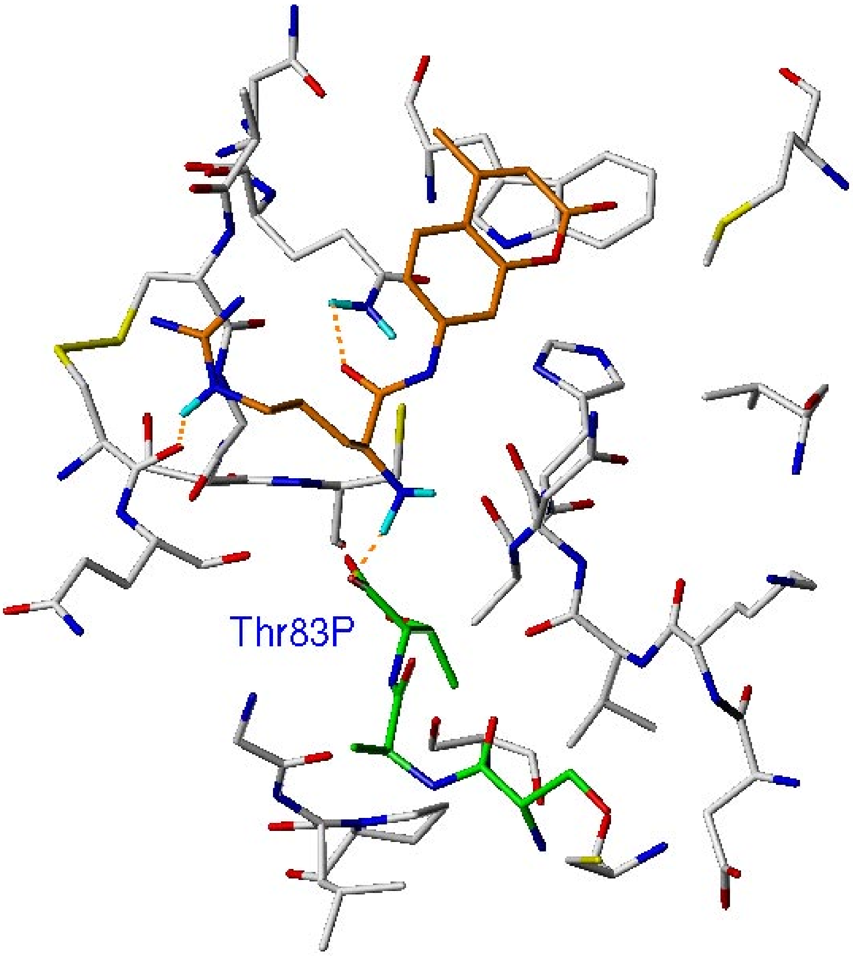

Figure 6 *Presentation of the binding pocket model of cathepsin H with the specific substrate Arg-NMec. The carbon atoms of the substrate are coloured orange. The amino acid residues of the min-chain Thr83P, Ala82P, and Ser81P are displayed green.*

additional attractive interactions can be observed to the side chains of Gln19 and Trp188. The number of hydrogen bonds from the substrate to cathepsin L emphasizes the energetically favorable position of this ligand into the active-site cleft of this enzyme. Since in cathepsin H the mini-chain occupies the S2 and S3 subsites the amino acid residue Phe and the Z group of the substrate cannot form a stable enzyme-ligand complex. In this case the position of the ligand is energetically unfavorable.

Besides, the values describe also the different kinetic data. The K_m value of Arg-NMec with cathepsin H is significantly higher compared to cathepsin L complexed with Z-Phe-Arg-NMec (Table 5). These findings are in agreement with the calculated non-bonded interaction energies and behavior of the complexes of the cathepsins H and L.

Discussion

All papain-like cysteine proteases have the same basic mechanism of action, but their substrate specificity differs from one member of the family to another. The active-site cleft of cysteine proteases like papain [38], actinidin [39], cathepsins L [40], and K [41] is unoccupied and is able to bind substrates along its full length, whereas the active-site cleft of cathepsin H is partially filled, thereby limiting the free substrate-binding sites [18].

The model of human cathepsin H that we constructed is in good agreement with the known high-resolution crystal structures of cathepsins B [34], L [37], and K [41]. The validity of the model is supported by a mostly negative energy graph of cathepsin H obtained by PROSA. We could establish that the structure corresponds to stable parts of this molecule (Figure 3).

By the determination of the position of the so called mini-chain within the active site-cleft of cathepsin H it could be shown in agreement to the described X-ray structure of porcine cathepsin H that the mini-chain runs in an extended conformation in the substrate binding direction along the front side of the active-site cleft (see Figure 5a). The disulfide bridge between Cys80P of this octapeptide EPQNCSAT and Cys212 of cathepsin H could be detected as an anchor. The model of human cathepsin H shows that the S2 and S3 subsites are occupied with the amino acid residues of the mini-chain and form a large number of attractive hydrophobic interactions

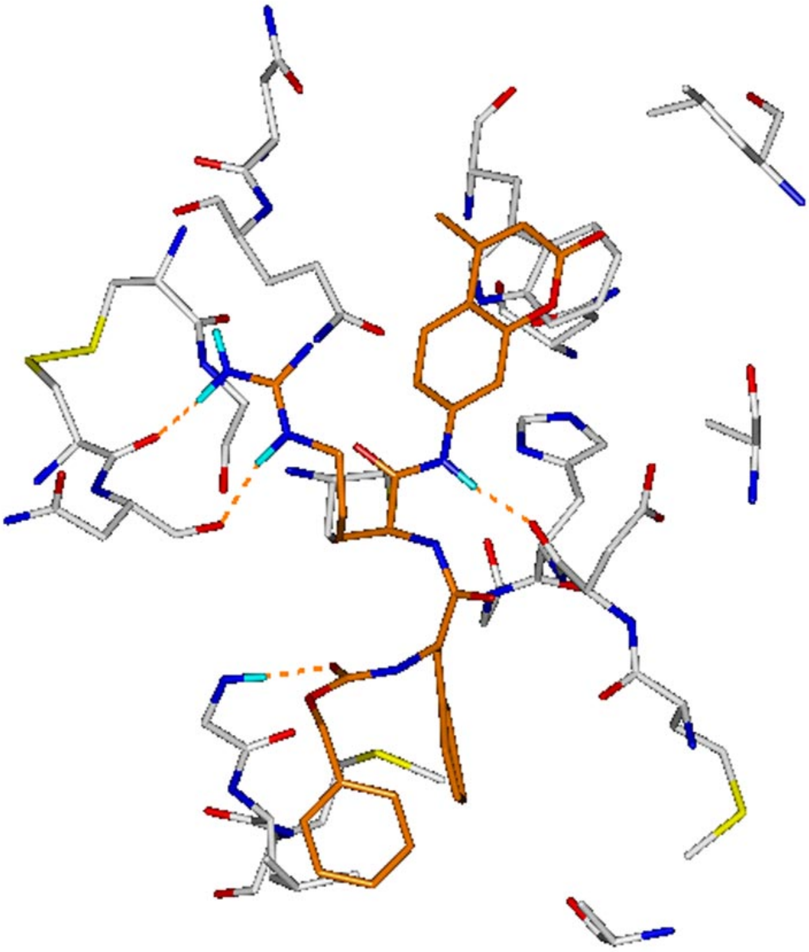

Figure 7 *Presentation of the active-site cleft of cathepsin L with the specific substrate Z-Phe-Arg-NMec (orange)*

Table 5 *Characteristics of cathepsins H and L interactions with their ligands [2, 30]*

Substrate	Cathepsin H			Cathepsin L		
	K_m (mM)	ΔE [a] (kcal·mol^{-1})	ΔE [b] (kcal·mol^{-1})	K_m (mM)	ΔE [a] (kcal·mol^{-1})	ΔE [b] (kcal·mol^{-1})
Arg-NMec	0.150	-52.03	-129.34	–	-39.19	-101.96
Z-Phe-Arg-NMec	–	–	–	0.006	-54.09	-138.59

[a] non-bonded interaction energies
[b] binding energies calculated using LEAPFROG

and hydrogen bonds. The amino acid residues Cys80P, Ala82P, and Thr83P fill mainly the S2 binding site, Gln78P the S3 subsite of cathepsin H. With these docking studies we confirm that the mini-chain as a part of the cathepsin H propeptide binds in the mature enzyme along the active-site cleft in the substrate-binding direction.

The results of our investigations show the model of human cathepsin H correlates with the X-ray structure of porcine cathepsin H. Although the sequence identity of hCatH to pCatH is very high (91%) some differences between both enzymes were found. In contrast to the X-ray structure of pCatH the thiolat-imidazol ion pair between His159 and Cys26 can be formed in the model due to altered orientations of the side chains of the active-site residues in accordance to other related cysteine protease.

Based on this model of the tertiary structure of human cathepsin H with the mini-chain the substrate specificity of cathepsin H compared to cathepsin L could be investigated and explained. Cathepsin H shows a preference for residues with large hydrophobic (Phe, Trp, Leu, Tyr) or basic (Arg, Lys) side chains at the P1 position.

Considering the substrate specificity, cathepsin H hydrolyses the substrate Arg-NMec, whereas for cathepsin L the enzymatic activity to this ligand is decreased. In comparison to cathepsin L we can observe more attractive hydrophobic interactions and formation of hydrogen bonds between the substrate and cathepsin H particularly for the arginine residue in the P1 position. Moreover, essential hydrogen bonds can be detected between the negatively charged C-terminus Thr83P of the mini-chain to the side chain of Arg of the substrate. The reduced interaction energy of this substrate to cathepsin L explains the observed lower substrate affinity compared to cathepsin H.

The estimated non-bonded interaction energies and the binding energies of the investigated complexes of cathepsin H and cathepsin L correlate with their K_m values (see Table 5). By using of the LEAPFROG program to calculate the binding energies of the complexes the essential influence of the solvation of the ligand, the desolvation of the enzyme as well as the contribution of the formation of hydrogen bonds could be considered. The energy values are listed in Table 5. The differences of binding energies obtained with both methods are in the same range, however, the absolute values obtained by LEAPFROG seem to be too high in comparison to the experimental results (K_m). Probably, contributions of hydrogen bonds are overestimated in this method [33].

Based on these results of our theoretical studies presented in this paper it can be concluded that the modeling of the tertiary structure of cathepsin H including the mini-chain and the docking studies of specific ligands is an effective way to determine the specificity of the binding pocket of cathepsin H.

This model of human cathepsin H together with our recently developed model of cathepsin S [19] and known X-ray structures of cathepsins (B, K, L) will be subject for further investigations to develop more specific substrates and inhibitors for these enzymes.

Acknowledgements We are grateful for research support from the Deutsche Forschungsgemeinschaft (SFB387/TP A8).

Supplementary material available statement Additional 3D information for structures shown in Figures 5-7 is available in pdb format (fig5a.pdb, fig5b.pdb, fig6.pdb, fig7.pdb).

References

1. Berti, P. J.; Storer, A. C. *J. Mol. Biol.* **1995**, *246***,** 273.
2. Barrett, A. J.; Kirschke, H. *Methods Enzymol.* **1981**, *80*, 535.
3. Katunuma, N.; Kominami, E. *Rev. Physiol. Biochem. Pharmacol.* **1987**, *108*, 1.
4. VanNoorden, C.; Smith, R. E.; Rasnick, D. *J. Rheumatol.* **1988**, *15*, 1525.
5. North, M. J.; Mottram, J. C.; Coombs, G. H. *Parasitol. Today*, **1990**, *6*, 270.
6. Sloane, B. F.; Moin, K.; Krepela, E.; Rozhin, J. *Cancer Metastasis Rev.* **1990**, *9*, 333.
7. Delaisse, J. M.; Ledent, P.; Vaes, G. *Biochem. J.* **1991**, *279*, 167.
8. Mort, J. S.; Recklies, A. D.; Poole, A. R. *Arthritis Rheumat.* **1984**, *27*, 509.
9. Takahashi, T.; Dehdarani, A. H.; Yonezawa, S.; Tang, J. *J. Biol. Chem.* **1986**, *261*, 9375.
10. Fruton, J. S.; Irving, G. W.; Bergmann, M. *J. Biol. Chem.* **1941**, *138*, 249.

11. Kirschke, H.; Langner, J.; Wiederanders, B.; Ansorge, S.; Bohley, P.; Hanson, H. *Acta. Biol. Med. Ger.* **1977**, *36*, 185.
12. Takahashi, T.; Dehdarani, A. H.; Tang, J. *J. Biol. Chem.* **1988**, *263*, 10952.
13. Machleidt, W.; Müller-Esterl, W. In Turk, V., (ed.) C*ysteine Proteinases and Their Inhibitors*, **1986**, pp 3-18, Walter de Gruyter and Co., Berlin-New York.
14. Takio, K.; Towatari, T.; Katuna, N.; Teller, D. C.; Titani, K. *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 3666.
15. Ritonja, A.; Popovic, T.; Kotnik, M.; Machleidt, W.; Turk, V. *FEBS Lett.* **1988**, *228*, 34.
16. Lafuse, W. P.; Brown, D.; Castel, L.; Zwilling, B. S. *J. Leukocyte Biol.* **1995**, *57*, 663.
17. Baudyš, M.; Meloun, B.; Gan-Erdene, T.; Fusek, M.; Mares, M.; Kostka, V.; Pohl, J and Blake, C. C. *Biomed. Biochim. Acta* **1991**, *50*, 569.
18. Guncar, G.; Podobnik, M.; Pungercar, J.; Štrukelj, B.; Turk, V.; Turk, D. *Structure* **1998**, *6*, 51.
19. Fengler, A.; Brandt, W. *Prot. Eng.* **1998**, *11*, 1007.
20. Blundell, T. L.; Carney, D.; Gardner, S.; Hayes, F.; Howlin, B.; Hubbard, T.; Overington, J.; Singh, D. A.; Sibanda, B. L.; Sutcliffe, A. J. *Eur. J. Biochem.* **1988**, *172*, 513.
21. http://www.rcsb.org/pdb/
22. TRIPOS Associates Inc., 1699 S. Hanley Road, Suite 303, St. Louis, MO 63144.
23. Needlemann, S. B.; Wunsch, C. D. *J. Mol. Biol.* **1970**, *48*, 443.
24. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer Jr., E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanaouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535.
25. Clark, M.; Cramer III, R. D.; VanOpdenbosch, N. *J. Comput. Chem.* **1989**, *10*, 982.
26. Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
27. Sippl, M. J. *Proteins* **1993**, *17*, 355.
28. Sippl, M. J. *J. Comp. Aided-Mol. Design* **1993**, *7*, 473.
29. Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Cryst.* **1993**, *26*, 283.
30. Tchoupé, J. R.; Moreau, T.; Gauthier, F.; Bieth, J. G. *Biochim. Biophys. Acta* **1991**, *1076*, 149.
31. Judson, R. In *Reviews in Computer Chemistry*, Lipkowitz, K. B.; Boyd, D. B., Eds.; VCH: New York, 1997; Vol. 10, pp 1-73.
32. Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849.
33. Bohacek, R. S.; McMartin, C. *J. Med. Chem.* **1992**, *35*, 1671.
34. Podobnik, M.; Kuhelj, R.; Turk, V.; Turk, D. *J. Mol. Biol.* **1997**, *271*, 774.
35. Cygler, M.; Sivaraman, J.; Grochulski, P.; Coulombe, R.; Storer, A. C.; Mort, J. S. *Structure* **1996**, *4*, 405.
36. Turk, D.; Podobnik, M.; Kuhelj, R.; Dolinar, M.; Turk, V. *FEBS Lett.* **1996**, *384*, 201.
37. Coulombe, R.; Grochulski, P.; Sivaraman, J.; Menard, R.; Mort, J. S.; Cygler, M. *EMBO J.* **1996**, *15*, 5492.
38. Drenth, J.; Jansonius, J. N.; Koekoek, R.; Swen, H. M.; Wolthers, B. G. *Nature* **1968**, *218*, 929.
39. Baker, E. N. *J. Mol. Biol.* **1980**, *141*, 441.
40. Fujishima, A.; Imai, Y.; Nomura, T.; Fujisawa, Y.; Yamamoto, Y.; Sugawara, T. *FEBS Lett.* **1997**, *407*, 47.
41. McGrath, M. E.; Klaus, J. L.; Barnes, M. G.; Brömme, D. *Nat. Struct. Biol.* **1997**, *4*, 105.