Review

Marko Novinec* and Brigita Lenarčič

Papain-like peptidases: structure, function, and evolution

Abstract: Papain-like cysteine peptidases are a diverse family of peptidases found in most known organisms. In eukaryotes, they are divided into multiple evolutionary groups, which can be clearly distinguished on the basis of the structural characteristics of the proenzymes. Most of them are endopeptidases; some, however, evolved into exopeptidases by obtaining additional structural elements that restrict the binding of substrate into the active site. In humans, papain-like peptidases, also called cysteine cathepsins, act both as non-specific hydrolases and as specific processing enzymes. They are involved in numerous physiological processes, such as antigen presentation, extracellular matrix remodeling, and hormone processing. Their activity is tightly regulated and dysregulation of one or more cysteine cathepsins can result in severe pathological conditions, such as cardiovascular diseases and cancer. Other organisms can utilize papain-like peptidases for different purposes and they are often part of host-pathogen interactions. Numerous parasites, such as Plasmodium and flukes, utilize papain-like peptidases for host invasion, whereas plants, in contrast, use these enzymes for host defense. This review presents a state-of-the-art description of the structure and phylogeny of papain-like peptidases as well as an overview of their physiological and pathological functions in humans and in other organisms.

Keywords: cysteine cathepsins; lysosomal enzymes; proteolysis.

Introduction

Peptidases are ubiquitous enzymes found in all living organisms. They are utilized for diverse biological processes ranging from digesting proteins for food to regulating highly organized processes such as the blood coagulation cascade. In humans, peptidases represent well over 10% of all known enzymes, with >800 peptidases currently listed in the MEROPS database of peptidases, their substrates, and inhibitors (available online at http://merops.sanger.ac.uk) (1). With 11 members, the papain-like cysteine peptidase family (in short, papain-like peptidases), accounts for only a small fraction of the total human peptidase count. However, owing to their broad expression patterns, potent activity, and participation in highly diverse physiological processes such as antigen presentation, extracellular matrix remodeling, and hormone processing, as well as pathological processes, such as cardiovascular diseases and cancer, these enzymes have been among the most widely studied peptidases for decades. Papain-like peptidases have been categorized as subfamily CIA in the MEROPS database and as the papain-like family in the Structural Characterization of Proteins (SCOP) database (available online at http://scop.mrc-lmb.cam.ac.uk/scop/). In recent years, animal papain-like peptidases have become collectively known as cysteine cathepsins to distinguish them from cathepsin peptidases that belong to other catalytic classes.

In the so-called postgenomic era, we have been witnessing an explosion of information in all areas of life sciences. Advances in whole-genome sequencing, proteomics, and other modern techniques have provided us with a wealth of novel information on the structure and function of papain-like peptidases as well as of their distribution in the different kingdoms of life. This review provides a state-of-the-art description of these versatile and fascinating enzymes with particular focus on their evolution and classification as well as a comprehensive overview of their biological functions in humans and other organisms.
The active center of papain-like peptidases

Mature papain-like peptidases are monomeric globular proteins with molecular masses in the 25–35 kDa range. The papain family fold is shown in Figure 1A in the standard orientation. The structure is composed of two halves of roughly equal size that have been termed the L- and R-domains according to their position in the standard orientation. The term domain in this respect is not entirely appropriate, as none of the halves was actually shown to be an independently folding and functional unit. Therefore, we will refer to the L- and R-entities as subdomains instead and the term domain will be used to describe the entire functional unit shown in Figure 1A, called peptidase unit according to MEROPS nomenclature. Accordingly, in multidomain proteins, this entity can be referred to as a peptidase or catalytic domain, in analogy to the nomenclature used for metallopeptidases.

The active site is located at the interface of both subdomains at the top of the molecule in the form of a V-shaped cleft (marked by an arrow in Figure 1A). The catalytic mechanism of papain-like peptidases is very similar to that of serine peptidases, the best-known class of peptidases in general. In comparison with serine peptidases, which contain a catalytic triad (Ser-His-Asp), papain-like enzymes are considered to contain a catalytic diad consisting of a Cys-His⁺ ion pair. There are, however, several additional residues that are necessary for catalytic competence (Figure 1B). Residues Asn175 and Gln19 are required for proper positioning of the catalytic diad. The side chain of the latter assumes the role of the oxyanion hole of serine peptidases in stabilizing the tetrahedral intermediate (2), whereas the former plays a role analogous to Asp in the catalytic triad of serine peptidases, forming a hydrogen bond to His159, but is not absolutely essential for catalysis (3). Recent evidence, however, suggests an important role for the adjacent Trp177 in the generation of the nucleophilic character of the catalytic diad (4).

The active site runs across the entire top side of the molecule and binds the substrate in an extended conformation as illustrated in Figure 1C on the model of the octapeptide Ala-Gly-Leu-Glu-Gly-Gly-Asp-Ala bound into the active site of human cathepsin K [adapted from ref. (5)]. Schechter and Berger (6) have postulated that the enzyme can interact with at least seven residues of the substrate, of which four lie N-terminal of the cleaved peptide bond (residues P4 through P1) and three lie C-terminal of the

![Figure 1](Image)
cleaved bond (residues P1′ through P3′) and have proposed seven corresponding sites on the enzyme (sites S4 through S3′). Two decades later, their definition was revised using crystallographic data available by then and five relatively well-defined sites were described (sites S3 through S2′) (7). In general, papain-like peptidases have broad specificity and the major determinant is the residue in the P2 position of the substrate. Specificities of numerous papain-like endopeptidases have been extensively studied and the results can be easily retrieved from databases such as MEROPS or BRENDA (http://www.brenda-enzymes.info). In general, most enzymes accept residues with hydrophobic side chains in the P2 position, although variations between enzymes have been observed in the preferences for individual side chains. Some enzymes have additional, less common specificities. Cathepsin B also cleaves substrates with Arg in the P2 position and other enzymes with such activity are sometimes referred to as enzymes with cathepsin B-like specificity, regardless of their evolutionary place within the papain family, which will be discussed later.

**Papain-like exopeptidases**

Most papain-like cysteine peptidases are endopeptidases; some, however, are also or exclusively exopeptidases. The general evolutionary strategy for creating papain-like exopeptidases was introduction of novel structural elements into the papain fold, which restrict access of the substrate to the active site at either side of the active site cleft. Four major types of such structural modifications are known, each creating an enzyme with unique activity within the papain family (Figure 2). Size wise, these modifications are highly diverse, ranging from a three-residue insertion in cathepsin X to the addition of an entire domain in dipeptidyl-peptidase I (cathepsin C). Carboxypeptidase activity was achieved by introducing insertions into the peptidase domain. In cathepsin B, the occluding loop (shown in yellow in Figure 2), an insertion of about 20 residues in the L-subdomain, blocks the binding of substrate beyond the S2′ position and thereby enables the enzyme to act as a peptidyl-dipeptidase. The binding of the occluding loop is pH dependent and is regulated by the protonation of the side chains of His110 and His111. At low pH, these groups are protonated and the occluding loop in bound into the active site. At high pH, deprotonation of these groups causes a displacement of the occluding loop and the enzyme exhibits endoproteolytic activity (8). Functionally similar, albeit much smaller, the mini loop of cathepsin X (shown in red in Figure 2) determines the carboxypeptidase activity of the enzyme through the side chain of residue His23 (9). The evolution of aminopeptidase activity has taken a different route and was achieved by introducing additional structural elements that are not part of the peptidase domain. The aminopeptidase activity of cathepsin H is determined by its mini chain (shown in blue in Figure 2), a remnant of the propeptide that remains covalently bound to the enzyme through a disulfide bond (10). The most prominent additional structural feature found in papain-like peptidases, however, is the exclusion domain of dipeptidyl-peptidase I (shown in green in Figure 2), which remains non-covalently attached to the peptidase domain after activation of the proenzyme. The exclusion domain not only determines the enzyme’s activity but also participates in its tetramerization (11). In addition to the exclusion domain, dipeptidyl-peptidase I also requires a chloride ion for catalytic activity (shown as a green sphere in Figure 2) (12).
Evolutionary groups of papain-like peptidases

Papain like-peptidases have been identified in different forms in a wide variety of organisms. Their presence in all three domains of life indicates the existence of an ancestral papain-like peptidase in the last common ancestor of bacteria, archaea, and eukaryotes. In the past decades, there have been several papers describing the classification of papain-like peptidases into subfamilies according to sequence similarities within the mature and propeptide regions. In this section, we reevaluate the evolutionary relations between papain-like peptidases using sequence data available in the MEROPS database (2545 sequences from the CIA subfamily in the current revision). The analysis of these data in combination with available experimental data provides us with a global oversight of the major evolutionary lineages within the papain-like family, which is presented in Figure 3. First of all, comparison between bacteria, archaea, and eukaryotes reveals distinct repertoires of papain-like peptidases in each domain of life, indicating substantial structural and probably also functional divergence. According to sequence similarities, there are multiple different lineages of papain-like enzymes in bacteria and archaea; however, due to virtually non-existent experimental data, these putative enzymes cannot be further categorized at this point. Thus far, the only characterized non-eukaryotic papain-like peptidase is xylellain from the Gram-negative bacterium Xylella fastidiosa, which was found to be an endopeptidase with specificity similar to cathepsin B and a relatively low pH optimum of around pH 5 (13). The crystal structure of xylellain has also been solved (PDB accession code 3OIS) and revealed a short propeptide (56 residues), with a conformation resembling that of the cathepsin X propeptide (see Figure 3).

Eukaryotic papain-like peptidases are much more familiar to the general reader as most of these have homologs in humans, other animals, and plants. Historically, the first classification of (eukaryotic) papain-like peptidases was done by Karrer et al. (14) who have divided then known papain-like peptidases into cathepsin L-like and cathepsin B-like enzymes on the basis of an ERFNIN motif that is conserved in the propeptide region of cathepsin L-like enzymes but not in cathepsin B-like enzymes. Several years later, Wex et al. (15) proposed a third, cathepsin F-like subgroup that contains an ERFNAQ motif in place of the ERFNIN motif, while also noting that several cysteine cathepsins apparently do not fall into any of these three subgroups. Considering the structural and genetic data available to us now, eukaryotic papain-like enzymes can be divided into cathepsin B-like peptidases, cathepsin X, dipeptidyl-peptidase I, and the cathepsin L-like group, which is further subdivided into multiple subgroups (Figure 3). In this section, we will discuss only the general features of each group, focusing mostly on propeptide similarities and disulfide bond patterns, whereas other unique characteristics will be discussed in other sections. The common characteristic of all eukaryotic papain-like enzymes is having two disulfide bonds in the L-domain, whereas other disulfide bonds occur in group-specific patterns.

Cathepsin B-like peptidases, cathepsin X, and dipeptidyl-peptidase I each form a separate group and are not closely related among each other or to other groups. Cathepsin X has the shortest propeptide among all eukaryotic enzymes, which is <40 residues in length and is covalently linked to the active site cysteine residue through a disulfide bond. It contains a unique insertion in the L-subdomain (shown in red in Figure 3) and a unique pattern of five disulfide bonds in the peptidase domain. Cathepsin B-like enzymes contain a somewhat longer propeptide (62 residues in human cathepsin B) that contains two short α-helices. These enzymes contain a conserved pattern of six disulfide bonds and the characteristic occluding loop, which determines the exoproteolytic activity of animal cathepsin B (shown in red Figure 3). Plant cathepsin B-like peptidases contain a somewhat shorter occluding loop and it remains to be determined whether it is functionally equivalent to its animal counterpart. Moreover, some basal unicellular eukaryotes contain cathepsin B-like peptidases that lack the occluding loop altogether. An example of such an enzyme is the Giardia lamblia cysteine peptidase 2, which has been shown to represent the oldest diverging branch of cathepsin B-like peptidases (16). The third group of papain-like peptidases is formed by dipeptidyl-peptidase I homologs, which contain the characteristic exclusion domain that has been suggested to have evolved from a metallopeptidase inhibitor (11). The propeptide of dipeptidyl-peptidase I is similar in length to cathepsin L-like peptidases, which will be discussed in the next paragraph, but lacks the characteristic ERFNIN/ERFNAQ motifs.

The largest group of papain-like peptidases can be collectively termed the cathepsin L-like group. It is characterized by a conserved pattern of three disulfide bonds, although individual members may contain additional disulfide bonds, and the largest propeptide among all papain-like peptidases, which is >100 residues long. The propeptide forms a folded, predominantly α-helical domain, with a central 25-30 residues long α-helix that...
contains the aforementioned ERFNIN/ERFNAQ motif (illustrated on the structure of human procathepsin K in Figure 3). The cathepsin L-like group is divided into several subgroups: cathepsin L-like peptidases, cathepsin F-like peptidases, cathepsin H-like peptidases, and cathepsin O, which is least similar to the remaining members. Cathepsin L-like peptidases contain the ERFNIN motif; they are mostly endopeptidases and have the characteristic papain...
structure, usually with only minor structural alterations within the peptidase domain. This group was evolutionarily very successful and represents the largest group of enzymes in most eukaryotic species. In humans, there are four members of this subgroup (cathepsins L, V, K, and S) and multiple cathepsin L-like peptidases were also found in most other animal species. Furthermore, a recent phylogenetic study revealed a large expansion of cathepsin L-like peptidases in the plant kingdom and subdivided plant cathepsin L-like peptidases into six distinct subgroups. In plants, cathepsin L-like enzymes occur not only as stand-alone mature proteins, but also in multidomain organizations, in which a granulin domain is located C-terminal of the peptidase domain (17). A multidomain cathepsin L-like enzyme, called the 26/29 kDa peptidase, is also found in insects and many other animal species, including zebrafish and chicken, but was lost in mammals. The 26/29 kDa peptidase architecture is similar to dipeptidyl-peptidase I, i.e., the propeptide is located between the 26- and 29-kDa chains that constitute the mature protein (18). Cathepsin F-like peptidases have been proposed to constitute a separate subgroup upon the characterization of human cathepsins F and W at the genomic level (15). Despite their similarity in the genomic organization and the conserved ERFNAQ motif, these two enzymes are quite different at the protein level as will be discussed later. Cathepsin F-like peptidases have also been identified in plants (17). Cathepsin H-like peptidases can be categorized as a third subgroup within the cathepsin L-like group. They contain the ERFININ motif characteristic for cathepsin L-like enzymes; what sets them apart, however, is the mini chain that remains attached to the peptidase domain after activation. As discussed above, this propeptide remnant determines the aminopeptidase activity of cathepsin H-like enzymes (see Figure 2). Plant homologs of cathepsin H are called aleurains (19). The last subgroup within the cathepsin L-like group is formed by cathepsin O. This enzyme can be considered the odd-one-out because the peptidase domain closely resembles that of cathepsin L-like peptidases and contains the characteristic disulfide bond pattern; its propeptide, however, is shorter and the characteristic ERFININ/ERFNAQ motifs are not conserved. A recent computational study has also predicted that the interactions between the propeptide and the peptidase domain of cathepsin O are different from those in cathepsin L (20). Thus far, cathepsin O has been identified only in animals, indicating that it may have evolved from a cathepsin L-like ancestral peptidase.

It should be noted that while the described groups are widespread among eukaryotes, they are by no means ubiquitous. Most notably, papain-like peptidases are rarely found in fungi. In contrast, fungi appear to be a rich source of inhibitors of cysteine peptidases, including macromolecular inhibitors as well as low molecular mass inhibitors such as E-64 from Aspergillus japonicus, probably the most widely used inhibitor of cysteine peptidases for experimental purposes. Furthermore, plants contain a modified set of papain-like peptidases, as both cathepsin X and dipeptidyl-peptidase I are absent, while the cathepsin L-like peptidase repertoire is expanded (17). Some organisms also contain lineage-specific innovations. An example is the Ser5 antigen from Plasmodium, which contains a papain-like domain within a larger architecture. The crystal structure of this domain has been solved (PDB access code 2WBF) and even though this specific domain does not appear to be catalytically active due to the mutation of the catalytic Cys residue to Ser, homologs with conserved catalytic residues have been identified. Apart from these, numerous other lineage-specific patterns can be observed; however, owing to lack of their functional and structural characterization, they will not be further discussed at this point.

Functional characteristics of papain-like peptidases

Animal papain-like peptidases, now commonly known as cysteine cathepsins, were long referred to as lysosomal peptidases. The advances of recent decades, however, have shifted our perception of cysteine cathepsins from non-specific lysosomal scavengers to specialized processing enzymes with a broad subcellular distribution that are involved in numerous physiological as well as pathological processes. Because of the latter, several cysteine cathepsins are among today’s top priority drug targets, e.g., cathepsin K for the treatment of osteoporosis (21). A lot of effort has been dedicated toward experimental characterization of the biochemical and functional properties of these enzymes, using essentially all available experimental techniques. In recent years, a number of excellent reviews have been published that contain detailed descriptions of their roles in different physiological processes and pathological conditions (22–27).

In the following sections, we will discuss the functional properties of each group of cysteine cathepsins with major focus on human (mammalian) enzymes and a brief description of their functions in other organisms. To provide an integrated perspective of the functional diversity of cysteine cathepsins in humans, Figure 4 shows a schematic overview of their major functions.
(physiological and at the molecular level) in different cells of the human body and their relatedness to pathological conditions. Human cysteine cathepsins are highly versatile and can act both as non-specific hydrolases as well as specific processing enzymes. Their major functions (highlighted in the shaded boxes in Figure 4) can be divided into endolysosomal digestion of protein substrates, which includes the processing of antigens in antigen-presenting cells, degradation of the extracellular matrix, and a diverse group of processes involving limited proteolysis of specific substrates, such as processing of hormone precursors and growth factors or activation of serine peptidases.
in the immune system. All processes outlined in Figure 4 are discussed in detail in the following sections.

**Cathepsin B-like peptidases**

The proteolytic activity of cathepsin B depends on the conformation of the occluding loop, as described in one of the previous sections. One of the earliest known examples of cathepsin B activity was its processing of muscle fructose 1,6-bisphosphate aldolase following stress-induced lysosomal leakage in muscle cells. Under these conditions, cathepsin B acts by sequentially removing up to nine dipeptides from the C-terminus of aldolase. This alters the activity of aldolase by reducing its affinity for fructose 1,6-bisphosphate without affecting its activity on fructose 1-phosphate and thereby affects the metabolism of the cell (28). In addition to its peptidyl-dipeptidase activity, cathepsin B can also function as an endopeptidase or as a carboxypeptidase (29). It is ubiquitously expressed and found in lysosomes in extremely high concentrations (up to 1 mM) (30). Apart from non-specific protein turnover, it is involved in antigen processing (31), processing of thyroglobulin in the thyroid gland (32), and maturation of lysosomal β-galactosidase (33), and its release from lysosomes into the cytosol has been shown to trigger apoptosis (34).

The main reason for the great scientific interest in cathepsin B should not be sought in its physiological roles but in its involvement in numerous pathological conditions. The most widely investigated pathologies in regard to cathepsin B activity are rheumatic diseases (35–37) and different types of cancer, such as breast and colon carcinomas, where it is usually associated with the stages of tumor progression and malignancy (23). The common denominator in both pathologies is the degradation of extracellular matrix by cathepsin B, which occurs in an environment that favors the endoproteolytic activity of cathepsin B. In numerous tumors, cathepsin B expression is upregulated (23) and mature cathepsin B is secreted from the cells where it is associated with the plasma membrane in caveolae (38). Localization of cathepsin B on the cell surface was recently confirmed by live imaging of human umbilical vein endothelial cells, which have been proposed to utilize cathepsin B for extracellular matrix degradation in angiogenesis (39). Similar experiments with transformed fibroblasts have shown that cathepsin B is secreted from the cells through secreted lysosomes, which are concentrated in the podosomes of these cells (40). The enzyme has been found to be active and relatively stable under such conditions that are usually considered unfavorable for lysosomal enzymes (41). It has, in fact, been shown that the activity of cathepsin B on small synthetic endoproteolytic substrates steadily increases with increasing pH, reaching maximal values between pH 7 and 8 (42). In vitro, cathepsin B has been found to degrade various extracellular matrix components, including the basal lamina components laminin and type IV collagen, fibronectin, and tenascin C (43–45), which is probably the molecular mechanism underlying the contribution of cathepsin B to cell invasion (46).

In recent years, there has been growing evidence connecting cathepsin B to Alzheimer’s disease, albeit its exact role(s) remain unclear. Cathepsin B has been shown to function as a major β-secretase in secretory vesicles of neurons, cleaving amyloid precursor protein to produce amyloid β, which accumulates extracellularly in patients with Alzheimer’s disease (47) and genetic deficiency in cathepsin B has been shown to reduce amyloid β deposition (48). However, cathepsin B also appears to degrade amyloid β and thereby reduce its deposition (49) and the overall effect has been proposed to be controlled by a balance between cathepsin B and its endogenous inhibitor cystatin C (50).

Apart from mammals, cathepsin B-like enzymes have been isolated and characterized from numerous other organisms. Plant homologs have been shown to be involved in the hypersensitive response, a form of programmed cell death in response to pathogen invasion, basal defense mechanisms, and plant senescence (51, 52). Similarly, the cathepsin B homolog CPC from Leishmania species is involved in programmed cell death (53). In this unicellular parasite, cathepsin B has also been found to be involved in the parasite stage of its life cycle within macrophages (54). A cathepsin B-like peptidase is also critical for host cell invasion by the parasite Toxoplasma gondii (55). Similarly, the G. lamblia cysteine peptidase 2 mentioned above is the major peptidase involved in encystation of this primitive parasite (56). Cathepsin B-like peptidases have also been described in several species of parasitic worms in relation to their putative roles in host-parasite interactions. In the liver fluke Fasciola hepatica, for example, three cathepsin B-like peptidases are among the major proteolytic enzymes secreted by juvenile parasites (57).

**Cathepsin X**

Cathepsin X is a carboxypeptidase. Initial reports have indicated that it can also act as a peptidyl-dipeptidase
Cathepsins L-like peptidases

Cathepsin L-like peptidases are the largest group of the papain family. The repertoire of cathepsin L-like enzymes in vertebrates is relatively diverse and can sometimes cause ambiguities in the comparison of their functional properties between species. In humans, there are four cathepsin L-like peptidases – cathepsins L, V, S, and K; in the mouse, the most frequently used animal model, however, there are only three counterparts – cathepsins L, S, and K, as well as an additional eight placenta-specific enzymes that arose by tandem duplications of the cathepsin L gene (81). Cathepsin L-like peptidases are involved in numerous physiological and pathological conditions, and depending on the specific situation they can act alone or in tandem with other peptidases. The roles of cathepsin L-like and other cysteine cathepsins in cells of the immune system have been intensively investigated [for a review, see ref. (26)]. Cathepsin S is the most important peptidase in antigen presentation (82–84) and cathepsin L has also been found to be critical for proper major histocompatibility complex (MHC) class II-associated antigen presentation in murine thymus (85). In humans, the role of cathepsin L in thymal antigen presentation is taken over by cathepsin V (86), a cathepsin L-like peptidase with an expression pattern limited to a few types of immune cells and testis. Cathepsin V is found only in higher pri-mates and is, in fact, more similar to mouse cathepsin L than human cathepsin L is, indicating that cathepsin V is the human ortholog of mouse cathepsin L (87, 88). The p41 splice variant of the invariant chain, a chaperone for MHC class II molecules and regulator of MHC class II antigen presentation [reviewed in ref. (89)], has been characterized as a potent inhibitor of cathepsin L-like peptidases, with the exception of cathepsin S, which is only weakly inhibited (90). Moreover, the p41 fragment of the invariant chain has been proposed to function as a chaperone for extracellular cathepsin L, which is relatively unstable in the extracellular milieu (91). Secretion of several cysteine cathepsins (K, L, S, and B) from activated macrophages in inflammation has been demonstrated, and these enzymes have been proposed to be the major factor contributing to tissue damage in chronic inflammation (92, 93). Indeed,
all cathepsin L-like peptidases were found to be potent elastinolytic enzymes that act through a mechanism different from that of neutrophil elastase (87, 94). In comparison with cathepsin L, other enzymes from this group are significantly more stable in the extracellular environment. Cathepsin S is optimally active at pH values around 7 (95), whereas cathepsin K is relatively unstable at neutral pH in vitro; however, its stability is markedly increased in the presence of heparin or the extracellular chaperone clusterin (5, 96). Not surprisingly, extracellular elastinolytic activity of cathepsin L-like peptidases has been linked to the development of pathological conditions that are accompanied by excessive degradation of extracellular matrix components, such as cardiovascular and pulmonary diseases [reviewed in refs. (24, 25)]. The elastinolytic activity of macrophages occurs in distinct compartments, with cathepsins K and S being the major extracellular elastases and cathepsin V the major intracellular elastase (87). Elevated levels of cathepsin L-like peptidases have also been identified in other inflammatory conditions, such as pancreatitis (97), and there have been indications that their contribution to these conditions may be due to their kinin-processing activity (25). Apart from inflammation, cathepsin L-like peptidases were identified in numerous types of cancer (23). In recent years, they have also been implicated in the regulation of adipogenesis and their elevated activity has been linked to obesity and, by implication, diabetes (98–101).

In addition to (patho)physiological processes involving utilization of multiple cysteine cathepsins acting in synergy, specific roles have been described for individual cathepsin L-like peptidases. The most thoroughly studied member is cathepsin K, the principal peptidase involved in bone remodeling. Its importance has been revealed by the discovery that loss-of-function mutations in the CTSK gene cause the rare autosomal recessive disease called pycnodysostosis, which is characterized by severe bone abnormalities (102). Cathepsin K is a highly potent peptidase that can cleave most extracellular proteins, including the triple helix of type I and type II collagens. In contrast to other endogenous collagenases, such as neutrophil elastase and matrix metallopeptidase-1, -8, -13, -14, and -18, which cleave the collagen triple helix at a single position (103, 104), cathepsin K can cleave at multiple positions along the molecule (105, 106). Excessive activity of cathepsin K has been associated with osteoporosis and several cathepsin K inhibitors are currently under investigation as potential therapeutics (107), as will be discussed in more detail later. Cathepsin K is also associated with age- and osteoarthritis-related type II collagen digestion in cartilage (108) and its inhibition has been shown to reduce cartilage degradation in animal models (109). It is also involved in osteosarcoma where its expression is correlated with metastasis (110) and has been shown to process SPARC, a known cancer-related protein, in bone metastasis of prostate cancer (111). In addition, cathepsin K is associated with schizophrenia through the conversion of β-endorphin to enkephalin (112) and is involved in Toll-like receptor signaling (113) and cardiovascular homeostasis (114).

Recent research has also identified a plethora of extracellular substrates for cathepsin S, including the basement membrane component nidogen-1 (115), the epithelial sodium channel (116), leptin (117), and plasminogen (118). Similar to cathepsins B and X, cathepsin S has also been found associated with the cell surface (119). Cathepsin L is indispensable for skin homeostasis and mice deficient in cathepsin L experience periodic hair loss (120), abnormal skin development (121), and bone anomalies (122). At the cellular level, the origin of this defect has been localized to keratinocytes, which exhibit abnormal growth factor processing in the absence of cathepsin L (123). Cathepsin L knockouts also developed late-onset cardiomyopathy (124) and revealed an important role for cathepsin L in the processing of neuropeptides (125). In humans, cathepsin V may substitute cathepsin L in some instances. It has, for example, been shown that human cathepsin V can compensate for murine cathepsin L in mouse skin functions (126). Recently, cathepsin V has also been shown to participate in the production of neuropeptides (127) and to release angiostatin-like fragments from plasminogen (128). Cathepsin L has also been found in the cell nucleus where it has been shown to be involved in the regulation of cell cycle progression by proteolytic processing of the transcription factor CDP/Cux (129, 130). Moreover, nuclear cathepsin L has been implicated in mechanisms of epigenetic regulation in mouse embryonic stem cells through proteolytic processing of histone H3 (131, 132).

Cathepsin L-like peptidases have also been investigated in numerous non-mammalian organisms. Much work has been done on cathepsin L-like enzymes of Plasmodium species as potential antimalarial targets, as these enzymes are the major peptidases involved in the degradation of hemoglobin in host erythrocytes. In Plasmodium falciparum, four homologs have been identified and termed falcipains [for a recent review, see ref. (133)]. The falcipains contain two unique structural features that are necessary for their specific functions (shown in Figure 5), an N-terminal extension of approximately 20 residues that is indispensable for correct folding of the enzyme and has been termed the refolding ‘domain’ (134) and a unique insertion in the R-subdomain that is involved in specific recognition of hemoglobin (135).
Cathepsin L-like peptidases also play crucial roles in the life cycles of many other unicellular organisms, including Trypanosoma (136), Tetrahymena (137), and Paramecium (138), as well as in plants and invertebrate animals. A particularly interesting example is cruzipain (also called cruzain), the major cysteine peptidase of Trypanosoma cruzi used by the parasite for feeding and invasion. It is a member of the cathepsin L-like group and it shows higher sequence similarity to human cathepsin F than to human cathepsin L, but it has cathepsin B-like activity and a unique C-terminal domain [for a recent review, see ref. (139)]. In flatworms, cysteine cathepsins represent about 10 – 30% of all active transcripts. In the liver fluke F. hepatica, cathepsin L-like peptidases are the major secreted enzymes, representing about 80% of all proteolytic activity of the adult parasite, which has been found to rely solely on cathepsin L-like peptidases for its feeding upon hemoglobin and serum proteins. The Fasciola cathepsin L-like subgroup is the largest subgroup of cysteine cathepsins described to date, with >20 paralogs produced by gene duplications [reviewed in ref. (140)].

Similarly, cathepsin L-like peptidases represent the majority of the papain-like peptidase repertoire in plants. These include some of the most commonly known enzymes, such as papain from papaya, bromelain from pineapple, and ficin from figs, which are produced by the plants in large quantities and are widely used for medicinal purposes, in the food industry and for numerous other applications [reviewed in ref. (141)]. On the basis of their characteristics, plant cathepsin L-like peptidases have been further subdivided into six groups (17). Functionally, these enzymes have been shown to be involved in plant innate immunity and in programmed cell death associated with different physiological or pathological circumstances. Papain and bromelain have been shown to protect plants from parasites such as insects or fungi (142, 143). Roles in plant immunity have also been described for the Arabidopsis RD21 protease and its tomato ortholog C16 (144). The xylem-specific peptidases XCP1 and XCP2 have been found to be involved in programmed cell death associated with xylogenesis (145), whereas the SAG12 protease has been associated with general plant senescence (146).

**Cathepsin H**

Cathepsin H is one of the earliest identified lysosomal cysteine cathepsins. Despite this, it has remained rather an enigma with respect to its biological functions. As an aminopeptidase, it has been shown to process several neuropeptides (147, 148) as well as granzyme B (149). It is also involved in surfactant protein B maturation in the lung, but not indispensable for this process (150). Interestingly, cathepsin H lacking the mini chain is capable of acting as an endopeptidase (151). The putative biological relevance of this, however, remains to be investigated.

**Cathepsin F-like peptidases**

Cathepsin F is distinguished from other papain-like peptidases by an unusually long prodomain, which in addition to the common propeptide contains a sequence of approximately 100 residues that shows similarity to cystatins (152). Human cathepsin F is an endopeptidase with specificity and activity on small synthetic substrates similar to cathepsin L-like peptidases, and its crystal structure shows no substantial structural differences from the cathepsin L-like subgroup (153). The parallels between the two subgroups appear to extend to the functional level, where cathepsin F has been found to participate in the same (patho)physiological processes as cathepsin L-like peptidases, even though it has been much less investigated than its cathepsin L-like orthologs. Thus far, it has been shown to be involved in antigen presentation (154) and has been found in atherosclerotic plaques where it has been implicated in the degradation of low-density lipoprotein particles (155).
Cathepsin F may not have received much spotlight among its human counterparts, it has been, however, somewhat more extensively investigated in some invertebrates. There have been several reports relating cathepsin F to the molecular mechanisms of host invasion by parasitic worms such as *Clonorchis sinensis* (156), *T. dorsagia circumcincta* (157), *Opisthorchis viverrini* (158), and *Manduca secta* (159). Indeed, in Asian flukes from the genera *Clonorchis*, *Paragonimus*, and *Opisthorchis*, cathepsin F-like peptidases are the predominantly expressed cysteine peptidases (140).

**Cathepsins O and W**

Although cathepsins O and W are not members of the same subgroup, several parallels can be drawn between them. They are both animal specific and they are the only two human papain-like peptidases that remain to be characterized in respect to their catalytic activities and three-dimensional structures. Given that all important catalytic residues in their primary sequences are conserved and that they have no special features surrounding their active sites, they can be best described as putative endopeptidases. Interestingly, cathepsin W contains an insertion in the R-subdomain, which is analogous to the hemoglobin-binding loop of falcipains; its function, however, remains unknown. Regarding their biological functions, cathepsin W has been shown to be expressed only in cytotoxic T cells and natural killer cells but is not essential for their cytotoxicity (160), whereas cathepsin O, despite exhibiting a broad expression pattern, has remained largely uninvestigated after its initial characterization (161).

**Regulatory strategies in papain-like peptidases**

Given their high destructive potential, papain-like peptidases must be tightly regulated to assure their proper functioning. Not surprisingly, in higher animals such as humans and other mammals, dysregulation of cysteine cathepsins, or their loss by genetic mutations, has been shown to directly cause severe pathological conditions or contribute to their progression. Aside from regulation at the level of protein expression and cellular trafficking of the newly synthesized products [reviewed by Brix et al. (22)], the most important level of regulation is the inhibition of active peptidases by macromolecular inhibitors, which will be discussed in more detail in the next section. Other well-investigated mechanisms of regulation are the processing of zymogens to yield mature enzymes, which usually occurs in the lysosomal compartment, and environmental parameters such as the pH and the redox potential. There are also a few other regulatory mechanisms that have emerged in recent years and seem to have significant impact on the biological activity of papain-like peptidases. One of these mechanisms is the regulation of active enzymes by glycosaminoglycans, which can play diverse regulatory roles and are further discussed in one of the following sections. The other mechanism is alternative splicing of mRNA transcripts, which affects the protein expression rate or yields proteins with alternative trafficking patterns. This phenomenon has been thoroughly investigated for cathepsin B in association with rheumatoid arthritis (162), and alternative splicing has been shown to either facilitate extracellular secretion or promote mitochondrial targeting of the enzyme, resulting in cell death (163, 164). Similarly, alternative splicing of cathepsin L has been associated with cancer (165). As described in one of the previous sections, cathepsin L can also localize to the nucleus where it is involved in the processing of several nuclear substrates. Apart from it, cathepsins B and F have also been detected in the nuclei of different cells (129, 166, 167). Their proteolytic activities in this environment, however, remain to be thoroughly investigated.

**Regulation by macromolecular inhibitors**

Arguably the most important mechanism in enzyme regulation in nature, competitive inhibition is also widely considered the most important regulatory mechanism for papain-like peptidases in human biology. Similar to other peptidase families, balance between cysteine cathepsins and their endogenous inhibitors is crucial for the well-being of the organism and there have been numerous studies highlighting its importance for the physiological and pathological processes described in previous sections (50, 168–170). The most studied inhibitors are proteinaceous and belong to several different evolutionary families (1, 171). The cystatins and the thyropins are the largest families in animals [comprehensively reviewed in ref. (172)], while a third family of inhibitors, called the chagasin family after its archetypal representative from *T. cruzi*, is found in bacteria, archaea, and some unicellular eukaryotes but is absent in animals and plants.
Representing a fascinating example of convergent evolution, inhibitors from all three families act in a similar manner, by inserting three loops into the active site of the target peptidase, as shown in Figure 6 on the crystal structures of the stefin A/papain, p41 fragment/cathepsin L, and chagasin/cathepsin L complexes.

Cystatins (inhibitor family I25 according to MEROPS nomenclature) emerged early in evolution and are found in most organisms, including bacteria (173). In vertebrates, they are divided into three groups, of which the stefins (group A) and the cystatins (group B) are standalone proteins, whereas the kininogens (group C) are composed of multiple cystatin repeats (172). They are usually general inhibitors of papain-like peptidases and inhibit both endopeptidases and exopeptidases, albeit some of them, e.g., cystatin F, have a more limited specificity (174). Stefin B is the most abundant intracellular inhibitor, whereas cystatin C is the major extracellular inhibitor of cysteine cathepsins (172). An even more diverse repertoire of cystatins is found in plants, where they are usually referred to as phytocystatins and have been implicated in defensive functions and regulation of protein turnover. The can occur as stand-alone proteins or in multidomain organizations called multicystatins. In addition, some members of the family have a unique C-terminal extension that enables these inhibitors to inhibit a second family of cysteine peptidases, the legumains (175).

Thyropins (MEROPS family I31) are more specialized than cystatins both in terms of activity and physiological action. They are found only in animals and the inhibitory unit, a thyroglobulin type-1 domain, is usually part of a multidomain architecture. The best-studied thyropin is the p41 fragment of the MHC class II-associated invariant chain, which has been discussed briefly in one of the previous sections and inhibits cathepsin L-like peptidases and cathepsins F and H (90). Additional contacts with the R-subdomain of the peptidase outside of the active site have been proposed to contribute to its selectivity (90). Apart from the p41 fragment, four other thyropins have been characterized up to date. These include one human member, testican-1, which is a selective inhibitor of cathepsin L (176) and three non-human proteins that occur in lineage-specific patterns, saxiphilin from North American bullfrog, equistatin from sea anemone, and the cysteine peptidase inhibitor from chum salmon eggs (172, 177). Interestingly, equistatin has been found to inhibit the aspartic peptidase cathepsin D in addition to papain-like peptidases (178).

Several members have also been characterized in the chagasin family (MEROPS family I42), including chagasin, cryptostatin from Cryptosporidium parvum, EhlCP1 and EhlCP2 from Entamoeba histolytica, and several inhibitors from different Plasmodium species. All were found to inhibit endogenous as well as host peptidases (179–182). Apart from these three major groups, several other inhibitors of papain-like peptidases have been described. Fungal mycocypins have been shown to inhibit papain-like peptidases as well as trypsin and asparaginyl endopeptidase (183). Moreover, three serpins, which are usually serine peptidase inhibitors, have been shown to also inhibit papain-like peptidases. These include the squamous cell carcinoma antigen 1 and endopin 2C, which show
inhibitory specificity for cathepsin L-like peptidases (184, 185), and hurpin that has been found to selectively inhibit cathepsin L (186).

Regulation by glycosaminoglycans

Glycosaminoglycans are abundant components of the extracellular matrix, the cell surface, and other cellular compartments that perform structural and regulatory roles. Interactions of lysosomal enzymes with intralysosomal membrane-bound glycosaminoglycans have been known for decades. In the strongly acidic environment of the resting lysosome, mature cysteine cathepsins and other lysosomal enzymes are kept in a latent form through inhibition by these negatively charged polymers (187). Later, glycosaminoglycans have also been shown to facilitate the autocatalytic activation of several cysteine procathepsins (188–190). The link between specific regulation of cysteine cathepsin activity and glycosaminoglycans has been established with the discovery that the collagenolytic activity of cathepsin K is regulated by chondroitin sulfate (191). The crystal structure of the cathepsin K/chondroitin-4-sulfate complex has shown that the glycosaminoglycan binds at a site distant from the active site (192), which led to the characterization of glycosaminoglycans as allosteric regulators of cathepsin K, as the first example of allosteric regulation in papain-like endopeptidases (5). Further experiments have shown contrasting effects of different glycosaminoglycans on cathepsin K, indicating that glycosaminoglycans can affect cathepsin K through multiple mechanisms, and a second glycosaminoglycan-binding site on cathepsin K has been predicted (5, 193). Apart from cathepsin K, regulatory effects of glycosaminoglycans have been described for cathepsin B (194), papain (195), and trypansosomal cathepsin L (196), indicating that this kind of regulation may be a widespread mechanism for the regulation of papain-like peptidases.

Pharmacological targeting of papain-like peptidases

Because of their involvement in pathological conditions, several papain-like peptidases are among the high-priority targets of the pharmaceutical industry. In recent years, human cathepsin K has been receiving a lot of attention as a target for the treatment of osteoporosis, a disease characterized by progressive loss of bone mass that affects mostly elderly individuals, primarily postmenopausal women. Several cathepsin K inhibitors have progressed into clinical trials. Thus far, the drug odanacatib (Merck & Co. Inc., Whitehouse Station, NJ, USA) has been the most successful in reducing bone turnover and increasing bone mineral density and is about to successfully conclude phase 3 clinical trials. Two further inhibitors are currently in development, ONO-5443 (Ono Pharmaceutical Company, Osaka, Japan) has progressed into phase 3 clinical trials as well, whereas MIV-711 (Medivir, Huddinge, Sweden) is currently undergoing phase 1 trials (197). Unfortunately, the development of two promising inhibitors, balicatib (Novartis, Basel, Switzerland) and relaxatib (GlaxoSmithKline, Collegeville, PA, USA), was stopped due to side effects associated with off-target reactivity [see ref. (107) and references therein for more details].

Perhaps not as widely known, but just as promising as the targeting of cathepsin K, inhibitors of cruzipain, the major peptidase of T. cruzi, are in development for the treatment of Chagas’ disease. The causative agent is transmitted through blood-feeding insects and the disease, which is common to Latin America, causes severe damage to the nervous and digestive systems and the heart. Thus far, the most promising compound is K777, which has been showing good results in animal models, such as dogs and mice (198, 199), and is about to enter phase I clinical trials (139). Apart from these two enzymes, several other papain-like peptidases are being considered as promising drug targets but are still several steps away from clinical trials, including dipeptidyl-peptidase I in autoimmune and inflammatory diseases (79), cathepsin B in Alzheimer’s disease (200), and plasmodial falcipains for the treatment of malaria (201).

Expert opinion

Papain-like peptidases are a structurally and functionally diverse family of enzymes that have been extensively studied for the past several decades. Nevertheless, they still remain a mystery in many aspects of their physiological roles in humans, let alone in other organisms. It is remarkable how evolution used a group of relatively simple and straightforward enzymes to perform a great variety of different biological functions. On the basis of currently available data, papain-like peptidases in eukaryotes can be divided into several evolutionary groups, of which the cathepsin L-like group is the largest. The papain-like peptidase repertoire of bacteria and archaea is different from that in eukaryotes and, judging from sequence data, can also be divided into several groups, with several
unique and interesting multidomain organizations predicted for the putative proteins. However, owing to lack of experimental data, a classification of these enzymes would not be feasible at this point. As microbial enzymes are frequently used for industrial applications, these enzymes will likely be further investigated in the following years in order to assess their usability for biotechnological applications.

In humans, papain-like peptidases are involved both in non-specific proteolytic processes, such as endolysosomal protein digestion, antigen processing, and extracellular matrix degradation, and in highly specific proteolytic pathways, such as zymogen activation and peptide hormone processing. Their importance is reinforced by the fact that their excessive activity has been associated with the development of numerous diseases. Multiple cysteine cathepsins, most frequently cathepsin L-like enzymes and cathepsin B, have been associated with pathological conditions characterized by excessive extracellular matrix degradation, such as atherosclerosis, arthritis, and pulmonary emphysema. Most cysteine cathepsins have also been associated with at least one type of cancer and some of them, e.g., cathepsin X in gastric cancer and cathepsin B in colorectal cancer, are currently under investigation for their usefulness as predictive markers for these pathologies. At the moment, cathepsin K, which is the principal peptidase in bone resorption, is probably the most intensively investigated cysteine cathepsin, as it is considered one of the most promising targets for the treatment of osteoporosis in postmenopausal women. Ongoing clinical trials of several cathepsin K inhibitors, most notably odanacatib, are showing promising results, which is a good outlook not only for the targeting of cathepsin K but also for targeting other cysteine cathepsins. Numerous studies have highlighted the importance of the peptidase/inhibitor balance for homeostasis and interactions between cysteine cathepsins and their macromolecular inhibitors have been exhaustively studied. In recent years, allosteric regulation has surfaced as a novel, previously overlooked mode of cysteine cathepsin regulation, indicating that our knowledge of the mechanisms underlying the regulation cysteine cathepsins is still far from complete and at the same time providing novel opportunities for drug design.

Diverse roles have also been described for cysteine cathepsins in other, non-vertebrate, organisms. Because of their potential roles as drug targets, these enzymes have been investigated in common human parasites, such as the malarial agent *Plasmodium* and parasitic worms, which utilize cysteine cathepsins in processes of host invasion and for feeding on host proteins. The roles of cysteine cathepsins have also been investigated in plant physiology, where these enzymes have been implicated in processes that are similar to those observed in humans, such as plant immunity and programmed cell death. Interestingly, two widely distributed subgroups of eukaryote cysteine cathepsins, cathepsin X and dipeptidyl-peptidase I, are absent in plants, indicating that their roles have been taken over by other enzymes.

**Outlook**

The work done in the past years has provided a solid foundation for future research, which will be focused on further improving our current understanding and implementing it for practical applications in medicine and technology. In one or two decades, we are likely to see a holistic view of the functional roles of cysteine cathepsins in human biology. At the moment, there are several focus areas that have remained largely uninvestigated, including the molecular characterization of human cathepsins O and W, endogenous mechanisms of allosteric regulation, characterization of the non-proteolytic functions of cysteine cathepsins, and their interactions with non-inhibitory binding partners. Judging by the current trends, we can expect to see a series of novel and improved inhibitors of cathepsin K for the treatment of osteoporosis and further attempts at targeting other cysteine cathepsins. Apart from studies on human enzymes, substantial progress can be expected in the studies of papain-like peptidases in other organisms, such as plants and prokaryotes, which will not only greatly contribute to our scientific knowledge on the structure and function of these fascinating enzymes but also will provide novel opportunities for their practical utilization.

**Highlights**

- Papain-like peptidases are found in all domains of life.
- Papain-like enzymes are mostly endopeptidases. A few of them are also or exclusively exopeptidases.
- Eukaryote papain-like peptidases are divided into several evolutionary groups according to disulfide bond patterns and propeptide homology.
- The largest group of papain-like peptidases is the cathepsin L-like group.
- Human papain-like cysteine peptidases are involved in numerous physiological and pathological processes. Their major functions include endolysosomal protein digestion, extracellular matrix degradation,
and processing of specific substrates by limited proteolysis.

- Papain-like peptidases are utilized by parasites for host invasion and for feeding.
- In plants, papain-like peptidases function in innate immunity and in programmed cell death.
- Balance between papain-like peptidases and their endogenous inhibitors is vital for homeostasis.

Acknowledgements: The authors thank Prof. Antonio Baici (Zürich, Switzerland) for critical reading of the manuscript and the Slovenian Research Agency for financial support (grants P1-0140, Z1-4077 and J1-4247). The authors declare no conflict of interest.

Received December 7, 2012; accepted February 11, 2013

References


156. Kang JM, Bahk YY, Cho PY, Hong SJ, Kim TS, Sohn WM, Na BK. A family of cathepsin F cysteine proteases of Clonorchis sinensis is the major secreted proteins that are expressed in the intestine of the parasite. Mol Biochem Parasitol 2010; 170: 7–16.


