Cloning, yeast expression, mutagenesis and phylogenetic analysis of a novel member of the *Fasciola hepatica* cathepsin L-like family.

Thesis presented for the Degree of

Doctor of Philosophy

by Lic. Jose F. Tort

under the supervision of John P. Dalton, Ph.D.

School of Biological Sciences

Dublin City University

August, 1997

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:________________ ID No.: __93701519__________

Date: __August 1997_____

Acknowledgments

I would like to thank Dr. John Dalton for all the support, guidance and advice during the last four years. This Ph.D. thesis would not have been possible without his help. I also like to thank to all the people in the parasitology lab, Sharon, Leda, Sandra, Bernie, Nan, Jacinta, Andrew, Ciaran, Paul and all the people that work in the lab during this period. They have been very helpful, and probably have a difficult time coping with my bad humor when things did not go just right. My thanks also to Dr. Thecla Ryan and her lab that gave me advice and support on the yeast cloning systems every time I needed it, and to Dr. Ken Wolfe whose help and guidance in the evolution issues was invaluable. I would also like to thank all the other people in the School of Biological Sciences, who have made this possible.

I want to express my gratitude to all the people at the Irish Council for Overseas Students. They made possible this study through their financial support, and they are doing a great job assisting students from all over the world. I also want to thank the people at the Facultad de Ciencias, in Uruguay, that helped me to come to Ireland in order to get a Ph. D. I hope its worth it.

Last but not least, I wish to dedicate this thesis to my family, my friends here and at home, to the Perez-Roche family, and to Olivia, for their patience, encouragement and support throughout these years.

Contents

Cloning, yeast expression, mutagenesis and phylogenetic analysis of a novel member of the *Fasciola hepatica* cathepsin L-like family.

Jose F. Tort

Abstract

Cathepsin L2, a major cysteine proteinase secreted by adult *Fasciola hepatica*, differs from other reported cathepsin L-like enzymes in its' ability to cleave peptide substrates that contain proline in the P_2 position. In the present study we have isolated a cDNA clone encoding a complete cysteine proteinase precursor from a *Fasciola hepatica* cDNA library screened with anti-cathepsin L2 serum. The deduced amino acid sequence was compared with other cysteine proteinases of *F. hepatica*. This confirmed that it belongs to a gene family composed of at least five different cathepsin L-like genes, and is different from other *F. hepatica* secreted cathepsin L-like proteinases.

The cloned gene was successfully expressed in yeast using the trafficking signals contained within its own propeptide, resulting in functionally active enzyme. Comparison of the yeast expressed enzyme and native liver fluke cathepsin L2 by immunological and biochemical analysis showed that the cloned zymogen encoded for the liver fluke cathepsin L2.

Cathepsin L2 differs in substrate specificity from *F. hepatica* cathepsin L1. To test if this difference is due to a tyrosine in the active site, site directed mutagenesis was performed to convert the leucine present in cathepsin L1 to the tyrosine present in cathepsin L2. The data obtained indicate that this substitution is not directly linked to the differences in substrate specificity observed between liver fluke cathepsin L1 and cathepsin L2. The mutated purified enzyme was not capable of cleaving substrates with proline in the P_2 position.

Phylogenetic analysis of the papain superfamily indicated that at least four different types of cysteine proteinases of the papain superfamily exist in trematodes. The liver fluke enzymes cloned so far, constitute a cysteine proteinase family equally related to the vertebrate cathepsin Ls, cathepsin Ss and cathepsin Ks. Other relationships between cysteine proteinases of diverse origin were also detected, which allowed us to group them into families and classes.

List of Abbreviations

AFC	7-amino-4-trifluoromethylcoumarin
BCA	bicinchoninic acid
BCIP	5-bromo-5-chloro-3-indolyl phosphate
Boc	t-butyloxycarbonyl
Bz	benzoyl
CBZ	benzyloxy carbonyl
cDNA	complementary DNA
E-64	L-3-carboxy-2,3*trans* epoxypropionyleucyl amido(4-guanidino) butane
EDTA	ethylenediaminetetraacetic acid
k_{cat}	first order rate constant (turnover number)
K_m	Michaelis-Menten term $(k_{-1} + k_2 / k_1)$
NHMec	7-amido-4-methyl-coumarin
PBS	phosphate buffered saline
PAGE	polyacrylamide gel electrophoresis
PAM	percentage accepted mutation
PMSF	phenylmethylsulphonylfluoride
RPMI	Roswell Park Memorial Institute
SDS	sodium dodecyl sulfate
Suc	succinyl
Tris	tris-(hydroxy-methyl)-methylamine (2-amino hydroxymethyl) propane-1,3-diol
TWEEN 20	polyoxymethylenesorbitan monolaurate
Tos	tosyl
QAE	quaternary aminoethyl
Z	benzyloxy carbonyl

1. Introduction

1.1 The parasitic way of life

Almost all existing free-living organisms are hosts to a very diverse range of parasites. A third of the known species of protozoa, more than half of the helminths and a great amount of arthropods are parasitic (Hyde, 1990). At least sixty thousand of the known species are parasites. Parasitism is then a common feature in the evolution of the eukaryotes. Although parasitism has come forth in the history of life in different phyla, the challenges encountered by a parasitic organism are in certain ways similar. Organisms that follow this way of life utilize resources produced by their hosts. Accordingly, diverse adaptations are needed to undergo the parasitic pathway.

The first challenge for a parasite (whether unicellular or multicellular) is to invade their host. Even if they are passively injected into their hosts through a vector (as happens in several unicellular parasites) they have to invade the cells or tissues where they can complete their life cycle.

Once the parasite has entered the host, it has to migrate to a certain location (a specific organ or cell type). This process usually implies traversing surrounding tissue, extracellular matrix, basement membranes and blood or lymph vessel walls. If using the circulatory systems as dissemination mechanisms, a similar journey has to be made when arriving at the tissue where the parasitic organism is going to be established. At the same time the parasite must overcome any physico-chemical and immunological barriers offered by the host.

Living in the hostile environment of another living being offers a major challenge for parasitic organisms. Although several problems confronted by free living organisms could be simplified in a restricted environment, other difficulties arise and solutions to them must be met by adapting mechanisms for the new circumstances. The mechanisms by which parasitic organisms achieve these goals are diverse, and are a fascinating example of the plasticity of the basic biological machinery. However, the range of molecular mechanisms available for coping with the new needs are limited, and so we are confronted with similar answers to the same problem in groups that are taxonomically unrelated.

In this process molecular effectors used normally for the turnover of cell components or those used in the breakdown of molecules for obtaining energy have to be redirected for the new objective. The redirection of enzymes used in cellular catabolic processes to fulfill the new requirements imposed by the parasitic way of life is observed in parasites from very diverse phylogenetic origins. Peptidases are a major component of cellular catabolic process, and is not surprising to find these enzymes as important elements in parasite invasiveness, nutrition and immune evasion. In this way, digestive enzymes can be used for the new purpose, generating in some cases a new differentiation pathway at the molecular level. Although the effectors can be similar, the origin of them can be different from an evolutionary point of view.

1.2 Proteolytic enzymes and their classification

Proteolytic enzymes or proteases are enzymes that catalyze the cleavage of peptide bonds in other proteins. They originate very early in evolution, since all organisms require them for digestion and metabolism of their own proteins. Present day digestive proteases of eukaryotes show similarities with prokaryotic ones, which allow us to estimate that they diverged from some common ancestor some billion years ago (James , 1980).

The duplication of genes and their subsequent functional divergence has been established as a fundamental process in adaptive evolution (Li and Graur, 1991, Hughes, 1994). This process leads to the formation of families of evolutionarily related, but functionally distinct genes. It is plausible to think of a set of peptide hydrolases with a broad specificity that acquired a higher degree of specialization by restricting their action to a select number of peptide bonds located at specific sites in specific protein substrates. At the same time, the expression and distribution of these new proteinases could also be restricted, generating specialized proteinases for specific purposes. By this concerted modification of substrates and genes, proteases were included in complex regulatory circuits (Neurath, 1984).

Unfortunately, we are now confronted with an assorted suite of proteinases, from diverse origins, and with their relationships obscured by millions of years of evolution. However, based on the comparison of their active sites, mechanism of action, sequence similarity and three-dimensional structure, proteolytic enzymes can be classified into consistent groups.

The first distinction that can be made is between those enzymes that remove one or more terminal amino acids from a protein substrate (exopeptidases) and those that are capable or cleaving internal peptide bonds of proteins (endopeptidases). The exopeptidases can be further subdivided into amino-peptidases or carboxy-peptidases, depending on which end of the polypeptide chain they attack. These enzymes usually take part in latter stages of protein degradation, acting on the products released by the action of endopeptidases.

Endopeptidases are the major catalytic effectors in the hydrolysis of proteins. Four mechanistic classes have been recognized, each bearing a characteristic set of functional amino acid residues arranged in a particular configuration to form the active site. The four classes are 1) Serine proteinases, 2) Aspartic proteinases, 3) Metallo proteinases, and 4) Cysteine proteinases.

The serine proteinases are probably the best understood of all endopeptidases, characterized by a catalytic triad formed by Ser, Asp and His. Two families that share the active site configuration and the catalytic mechanism, but differ in their primary sequence and their tridimensional folds have been recognized. The first family, the mammalian serine proteinases, include well known enzymes as trypsin, chymotrypsin, elastase, and the circulating enzymes of the blood coagulation cascade. The second family is represented exclusively by prokaryotic enzymes, subtilisin being the typical example of this group.

The digestive enzymes pepsin and chymosin, the lysosomal proteinase cathepsin D and the HIV protease are typical examples of aspartic proteinases. These enzymes are characterized by two Asp residues in the active site. At acidic pH, one of the active site Asp residues is ionized, allowing the peptide bond cleavage to take place by a simple acid-base mechanism (Dunn, 1989). These enzymes are usually inactive at pHs above 6.0.

Metallo proteinases are a diverse group of enzymes characterized by their optimum activity at alkaline pH (in the range 6 to 10), and by the presence of a metal ion, usually zinc, coordinating two Glu and a His residue. These enzymes are easily inactivated by chelating agents that remove the metal ion from the active site. Typical representatives of this group of enzymes are the mammalian matrix metallo proteinases (collagenases, gelatinases, stromelysins), bacterial thermolysin and several amino and carboxypeptidases.

1.3 Cysteine proteinases

The cysteine proteinases are an large group of enzymes with representatives in all eukaryotic lineages, and take part in a wide variety of cellular functions. The activity of the enzymes belonging to this group is dependent on the triad formed by Cys, His and Asp. The order of these residues is different in the primary structure of several cysteine proteinases allowing a first level of clustering of representatives in families. At least 35 different families have been identified, several of them being associated in clans. Tertiary structures of members of four families have been determined, and the differences in the folding observed are so significant that is almost impossible to think of a possible common ancestor between them. There is a strong line of evidence suggesting multiple independent origins of cysteine proteinases during evolution. Based on these structures and on the similarities at the primary sequence level, the families can be grouped into clans.

Several of the cysteine proteinases known so far, are of viral origin, and play important roles in the viral life cycle, being many times correlated with the pathogenicity of these agents (Gorbalenya and Snijder, 1996). These enzymes are responsible for the cleavage of the viral polyproteins , and are vital in generating active forms that allow the viral replication cycle to proceed. They can also hydrolyze some host proteins, being in this way associated with pathogenicity. An interesting group is the one formed by cysteine proteinases produced by picornaviruses. In the enzymes produced by this group of RNA viruses, the order of the residues in the active site in the primary sequence is His Cys. The amino acid sequence of the enzymes belonging to this group showed some relationships with the serine proteinase chymotrypsin. The crystal structure of picornain 3C from the human hepatitis A virus was determined, and it showed a general fold very similar to chymotrypsin and other serine proteinases of the same group. The active site serine residue has been changed into cysteine, while the conserved histidine residue is retained, but there is no homologue to the aspartic acid residue, the third member of the catalytic triad in serine proteinases. It is difficult to assess which was the ancestral form and which one is derived. Other RNA viruses have cysteine proteinases with the same order of residues in the active site as picornains, and furthermore, the order of genes in the viral polyproteins is the same, suggesting a common ancestor. In spite of these similarities, the primary sequence of the endopeptidases are very dissimilar, suggesting a high mutation rate in these proteins. For these reason these enzymes have been assigned to several families clustered in the same clan CB as the picornains (Barrett and Rawlings, 1996).

A second group of RNA viral cysteine proteinases are clustered in the clan CC with certain correlations with the papain superfamily, but presenting differences that suggest a different catalytic mechanism, and probably different origin. No structural data of members of this clan exists yet (Barrett and Rawlings, 1996).

The only DNA virus cysteine endopeptidase, L3/p23 from adenovirus, has a unique protein fold that categorize it as a new family, the only one in the clan CE. The viral encoded enzyme might play functions essential for infection of the host cell (Barret and Rawlings, 1996).

Cysteine proteinases of bacterial origin are scarce. So far the aminopeptidases of *Streptococcus, Lactococcus* and *Lactobacillus*, the streptopain of *Streptococcus* and *Porphyromonas,* the clostripain of the anaerobic bacterium *Clostridium histolyticum* and the gingipains of *Porphyromonas gingivalis* have been characterized (Barret and Rawlings, 1994, 1996). The gingipains are secreted enzymes with specificity towards arginyl or lysyl bonds. Clostripains on the other hand are calcium dependent enzymes that show strict specificity for arginyl bonds, and are characterized by an inhibition profile different to other cysteine proteinases. The aminopeptidases C and streptopain are more closely related to the eukaryotic cysteine proteinases, in particular with the papain superfamily. These enzymes show restricted conservation in the region around the active site residues, although the active site Asn has been replaced by an Asp residue in streptopains.

It is in eukaryotes that cysteine proteinases have flourished, being important intermediaries in very diverse biological functions. Two families of cysteine proteinases (C12 and C19) are involved in the hydrolysis of peptide bonds at the C-terminal carboxyl group of ubiquitin. Ubiquitin is a short protein that conjugates itself to other proteins through its carboxyl terminal group; it acts as a signal for protein degradation, or it may have a chaperone function in the assembly of oligomeric proteins. The removal of the ubiquitin moiety by the de-ubiquitinating peptidases allows the recycling of the signaling protein (Barrett and Rawlings, 1994, 1996).

Legumains are cysteine proteinases with strict asparaginyl endopeptidase activity, and constitute another well defined family . They were first discovered as vacuolar proteases of legumes, and have been involved in seed maturation and degradation of seed proteins during germination. Homologues to the plant enzymes have been found in yeast, invertebrates and more recently in humans. The best-known member of the family is Sm32 from *Schistosoma mansoni* (Dalton *et al*., 1995).

A single family constitutes the clan CD of cysteine proteinases. This family is represented by the interleukin 1-beta converting enzymes (ICE), and other cytosolic cysteine endopeptidases intimately involved in the process of apoptosis. The members of this family have strict specificity for the cleavage of aspartyl bonds. The crystal structure of the ICE is completely different to papain and other cysteine proteinases, indicating that they probably have an independent origin.

1.3.1 The papain superfamily

The majority of the cysteine proteinases known so far belong to the papain superfamily. This superfamily includes the enzymes related to papain, the bleomycin hydrolases and the calpains. The latter two are cytoplasmic multimeric enzymes which share restricted similarities with bacterial cysteine proteinases. The bleomycin hydrolases (cytoplasmic enzymes involved in the detoxification of the anticancer drug bleomycin) are the most similar to the bacterial aminopeptidases C, whereas the calpains (calcium activated proteinases) share some restricted homology with enzymes from the gram negative bacterium *Porphyromonas gingivalis* (Berti and Storer, 1995).

The remaining members of the papain superfamily are in general vesicular or secreted proteins that are synthesized as pre-pro-enzymes, and require the removal of the propeptide to obtain full activity. The group include endopeptidases with broad specificity (e.g. papain), others with narrow specificity (e.g. glycyl endopeptidase), aminopeptidases, dipeptidyl-pepetidases (e.g. cathepsins C) and enzymes with both endopepetidase and exopeptidase activity (e.g. cathepsins B and H). Some enzymes have substitutions in the active site cysteine residues, i.e. the soya bean oil body-associated protein (SOYBN_P34) where a Gly is in this position, and the surface protective protein of *Plasmodium* (PLAFG_SERA), and a cathepsin B like enzyme from *Schistosoma japonicum* (SCHJP_CB2) which both have the Cys residue substituted by a Ser.

The plant members of the superfamily are mainly intracellular proteins. Most of the plant cysteine proteinases known so far are vacuolar enzymes present in germinating seeds, where they are involved in the hydrolysis of storage proteins. In other cases, the cysteine proteinases are expressed under stress conditions, specially as a response to high salt concentrations, dehydration, or low temperature. Some plant cysteine proteinases belonging to this superfamily have been involved in flower or leaf senescence. A few representatives are responsive to wounds, and are secreted. Papain itself is the best example of this type of secreted enzymes. In all cases, the expression of the cysteine proteinases seems to be tightly regulated, and is in most cases inducible. An important role in the induction seems to be played by phyto-hormones such as gibberellic acid or abscisic acid. It is clear from Southern blot analysis in different species that multiple cysteine proteinases exist in plants, but very little is known about their genomic organization. Two genes have been mapped to chromosome 3 in barley (Mikkonen *et al.*, 1996) , and are part of a small gene family with no more than four or five different genes. Two gene clusters with at least two genes each have been identified in *Arabidopsis thaliana* (Koizumi *et al.*, 1993). A small gene family on group 4 chromosome containing a cathepsin B like enzyme has been demonstrated by Southern blots in wheat (Cejudo *et al.*, 1992).

The cysteine proteinases of vertebrates were first described as the lysosomal cathepsins. These enzymes are responsible for more than 50% of the total cellular degradation in mammalian cells (Bond and Butler, 1987). The lysosomal proteinases are cysteine proteinases with the exception of cathepsins D and E which belong to the aspartic proteinases, and cathepsins G and A which are serine proteinases. Originally the cathepsins were considered exclusively as intracellular enzymes, but more recent data have linked these peptidases in a much broader set of extracellular functions both normal and pathological, such as tissue remodeling, arthritis, Alzheimer disease, periodontal diseases, sinusitis and tumor metastasis. In all these latter cases, the enzymes are secreted extracellularly (Muller-Ladner *et al.*, 1996, Elliot and Sloane, 1996) .

Different cathepsins can be recognized based on their substrate specificity, their sequence and tissue distribution. All the vertebrate cathepsins are glycoproteins of small size (in the range of 21 to 28 kDa), are synthesized as pre-pro-enzymes and are processed into the mature form, usually in the lysososme. The propeptide is involved in assuring a proper folding of the mature enzymes, and it is possible that the removal of the propeptide is autocatalytic. Cathepsins B, L, H and probably S and K undergo further processing by clipping the mature enzyme into two chains, usually called heavy and light chains, although the sites of processing are not conserved between all the cathepsins. Cathepsin C, a dipeptidyl-aminopeptidase, is the only multimeric enzyme having a molecular mass of 200 kDa, and consisting of four identical subunits. Each subunit comprises three polypeptide chains, two of them corresponding to the heavy and light chains of the other cathepsins and the third chain being the long propeptide (205 amino acids) (Dolenc *et al.*, 1995).

The several cathepsins found in vertebrates are single genes or low copy number clusters localized on different human chromosomes. Cathepsin B is a single copy gene mapping to chromosome 8p22 (Fong *et al.*, 1991), cathepsin L is also single copy and maps to chromosome 9q21(Chauhan *et al.*, 1993), cathepsin H maps to chromosome 15q24-q25 (Wang *et al.*,1987), cathepsin C is localized on chromosome 11q14 (Rao *et al.*, 1997) while cathepsin S and K are single genes that map to chromosome 1q21, being 150 kilobases apart (Shi *et al.*, 1994, Rood *et al*, 1997, Gelb *et al.*, 1997) . A small cathepsin L-like gene cluster containing three genes exists on chromosome 10q23 (Bryce *et al.*, 1994). The mouse genes for the cathepsins B, L , H and S map to chromosomes 14, 13, 9 and 3 respectively, lying within known regions of conserved synteny between mouse and human chromosomes (Deussing *et al.*, 1997). In the latter study, two other localizations were identified in mouse, one recognized by a cathepsin B probe in chromosome 2 and another in the X chromosome recognized by a cathepsin H probe, indicating the existence of other yet unidentified cathepsin B-like and cathepsin H-like genes.

The tissue distribution of cathepsins is variable. While cathepsin B is the most abundant lysosomal cysteine proteinase in different mammalian tissues (Barrett and Kirschke, 1981), cathepsins L and H show a more restricted expression pattern than cathepsin B, and cathepsin S is highly expressed in spleen, heart, lung and to a lesser extent brain (Shi *et al.*, 1994). Cathepsin K is highly expressed in osteoclasts, and has been associated with bone remodeling (Tezuka *et al.*, 1994). Cathepsin O is a very divergent cathepsin purified from a breast carcinoma cell line and shows a wide tissue distribution according to northern blot analysis (Velazco *et al.*, 1994). More recently, a very divergent cysteine proteinase from human lymphocytes has been cloned and termed cathepsin W. This enzyme shows restricted expression to lymphatic tissue, being higher in CD8+ cells (Linnevers *et al.*, 1997).

The restricted tissue distribution of several of the vertebrate cathepsins indicates a tight regulation of their expression. In this sense cathepsin B exhibits a TATAless GC rich promoter characteristic of a constitutive gene, although the levels of cathepsin B mRNA varies with cell type, differentiation and extracellular stimuli, indicating that the transcription is regulated over a basal level. It seems that more than one promoter exists, and furthermore, up to eight different transcripts are produced from the same locus, by alternative splicing in the 5' and 3' non coding regions. Two of these transcripts do not contain the signal peptide and the first half of the propeptide, so the resulting polypeptide should be cytoplasmic. Interestingly, cytoplasmic cathepsin B seems to be present in several tumors. The existence of several transcripts due to alternate splicing seems to be a common feature of the mammalian cathepsins since it is also documented for cathepsins L and S (Chauhan *et al.*, 1993, Shi *et al.*, 1994).

Cathepsin B is the principal cathepsin associated with malignant transformed cells from diverse tissue origins, in accordance with its wider tissue distribution than other cathepsins. Furthermore, the enzyme seems to be up-regulated in malignant cells, and is capable of degrading the basal membrane proteins laminin and type IV collagen at neutral or acidic pH, suggesting that this enzyme is partially responsible for the invasiveness of the transformed cells (Elliot and Sloane, 1996). Cathepsin L is also capable of degrading basal membrane proteins, several collagen types and proteoglycans (Muller-Ladner *et al.*, 1996). Cathepsin L is the most powerful lysosomal proteinase, with a much wider substrate specificity than cathepsin B (Barrett and Kirschke, 1981). This enzyme and cathepsin H, have also been associated with tumor invasion and metastasis, although the correlation is less strong than with cathepsin B. The tissue-specific expression of the other cathepsins, suggest that if they play a role in cancer, it would be in cancers originating in specific tissues.

Other pathological processes where cathepsins seem to be involved are Alzheimer's disease, glomerulonephritis, periodontal disease, arthritis and inflammation (Muller-Ladner *et al.*, 1996). In all these cases, the enzymes are secreted extracellularly, a feature that appears to be more common than originally believed.

The data on cysteine proteinases of invertebrates is more scarce. There are well known examples of cysteine proteinases of the papain superfamily in parasitic helminths (presented below), but the data on free living invertebrates is more restricted. However, a family of cysteine proteinases from the digestive glands of decapods have been purified. Three members of a gene family comprising probably ten genes have been cloned from the digestive gland of the American lobster (Laycock *et al.*, 1991), two cDNAs from the Norway lobster, one of them from the eye stalk, but with high homology to one of the American lobster genes (LeBoulay *et al*., 1995), and two cDNAs from the digestive gland of shrimps (LeBoulay *et al*., 1996). These genes were termed cathepsin Ls due to the sequence similarity shown with the vertebrate lysosomal enzymes.

In *Drosophila melanogaster* a digestive cysteine proteinase has also been cloned (Matsumoto *et al*., 1995). More recently, a family of cysteine proteinases has been identified in a coleopteran, the maize weevil (Matsumoto *et al*., 1997). At least four genes exist in this organism, but only one is highly expressed, predominantly in the digestive system and in oocytes and nurse cells. Further evidence for the presence of digestive cysteine proteinases in insects came from the observation that the propagation of the cereal insect pests, *Callosobruchus chinensis* and *Riptortus clavatus* is inhibited by the addition of rice cystatins to their diet (Kuroda *et al*., 1996). This latter study indicates that cysteine proteinases might play an important role in the digestion of nutrients in arthropods. A cysteine proteinase similar to cathepsin L has also been related to tissue remodeling functions during molting in the flesh fly *Sarcophaga peregrina (*Homma *et al*., 1994). The secreted enzyme has been preferentially localized to the imaginal discs and significant amounts of its mRNA were localized to unfertilized eggs. The role of cathepsin-like enzymes in the early development of invertebrates is further documented by the cloning of a cathepsin L-like proenzyme from the eggs of the silkmoth *Bombix mori* (Yamamoto *et al*., 1994), and a cathepsin B-like enzyme in the yellow fever mosquito *Aedes aegypti* (Chen *et al*., 1995). In these cases, the cysteine proteinases have been implicated in the degradation of storage proteins. Similar proteinases from parasitic invertebrates and protists have also been identified, and they are discussed in the following sections.

1.3.2 The general catalytic mechanism of cysteine proteases of the papain superfamily.

Cysteine proteinases of the papain superfamily catalyze the hydrolysis of peptide, amide, ester, thiol ester, and thiono ester bonds (Storer and Menard, 1994). By definition, proteases of this group would require the thiol group of a cysteine residue for their activity.

Cysteine proteases possess an active site cysteine residue (Cys^{25} in papain) which acts as a nucleophile that attacks the carbonyl carbon of the scissile peptide bond of the substrate. A histidine residue (His^{159} in papain) constitutes with the cysteine a thiolate-imidazole ion pair, conferring high nucleophylicity to the active site cysteine residue. An asparaginyl residue (Asn^{175} in papain) has a role in orientating the imidazolium ring of the catalytic histidine. An oxyanion hole similar to that found in serine proteases is formed between the side chains of the active site Cys^{25} residue and a glutamine (Gln^{19} in papain). Because of their importance in the catalytic functions, these four residues are absolutely conserved throughout the papain superfamily.

Several steps can be considered in the mechanism of hydrolysis. After a non-covalent union of the enzyme and the substrate (forming the Michaelis complex) the enzyme is acylated and the first product (corresponding to the carboxy portion of the peptide being cleaved) is released. The acyl-enzyme is then dissociated by a water molecule liberating the second product of the reaction, and regenerating the free enzyme (Fig. 1.1). Many intermediates are believed to exist along this pathway, and several efforts to dissect the mechanism have been made. This mechanism is well conserved for all members of the papain superfamily. But , as observed in other hydrolytic enzymes, cysteine proteases utilize a wide variety of enzyme-ligand interactions to achieve preferential hydrolysis of certain substrates over others.

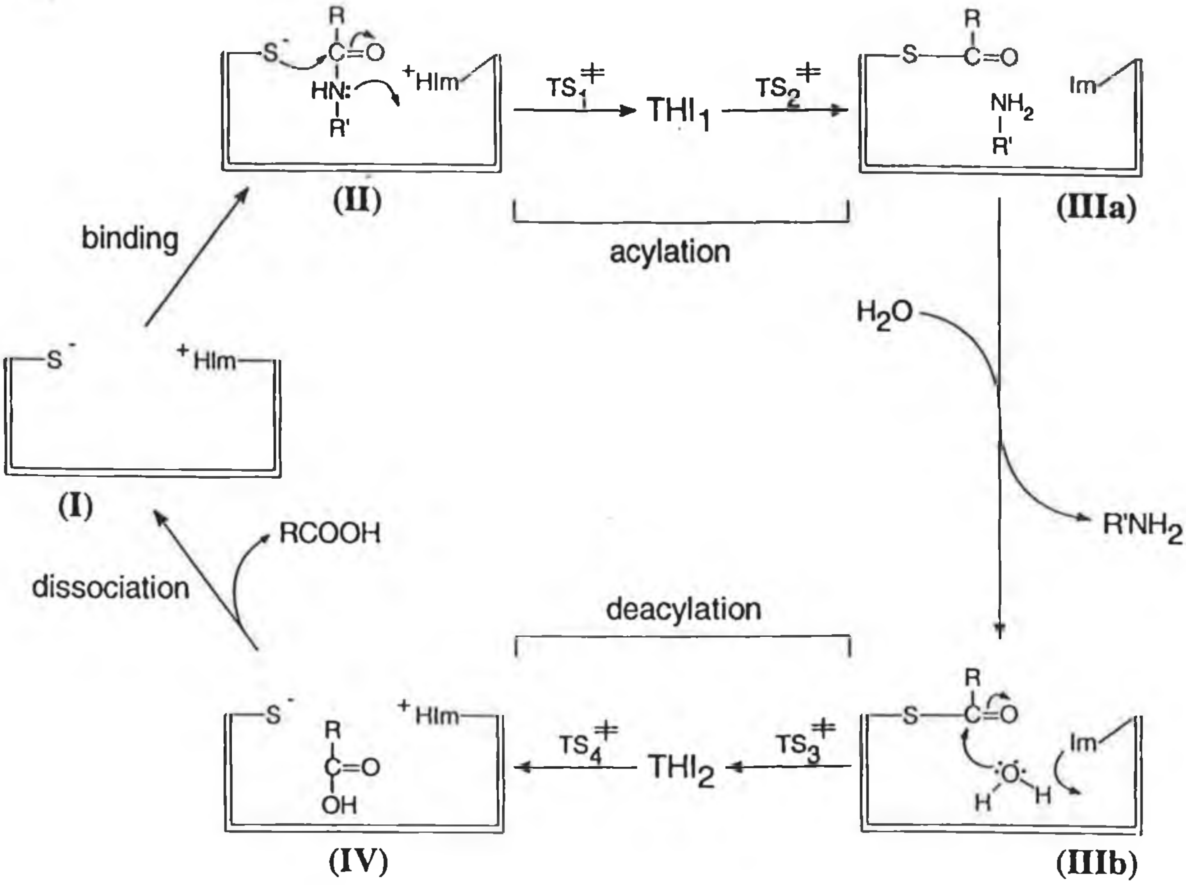


Fig. 1.1. Schematic representation of the various steps, putative intermediates and transition states involved in the reaction pathway for hydrolysis of an amide substrate by a cysteine proteinase. From Storer and Menard, 1994.

1.3.3 The active site of papain-like cysteine proteinases

The active site of papain is considered to consist of seven subsites (S_1-S_4 and S_1'-S_3'), each able to accommodate one amino acid residue of the substrate (P_1-P_4 and P_1'-P_3') located on the N-terminal and C-terminal side of the cleavage site respectively (Schechter and Berger, 1967). The differences observed in the specificity towards diverse protein or synthetic substrates should lie in the ability of the residues forming the S subsites to accommodate various different residues of the substrate. The existence of an extended binding site for the substrate, allows several interactions to be made, making it difficult to recognize the residues involved in the specificity. So far, the S_2 subsite is the only one with a clear preference that can account for the selectivity of enzymes of this superfamily. Furthermore, in the case of natural substrates, the interactions established in a particular subsite can vary according to the neighboring residues in the polypeptide chain and the interactions in which they are involved.

Several attempts have been made to identify the residues that interact with the substrate, with a special emphasis on those that could account for the specificity of the enzyme. Early studies with protein substrates demonstrate that bulky non-polar side chains like phenylanane are preferred at the S_2 position in papain and in rat cathepsin L (Koga *et al.*, 1990). Analysis of the crystal structure of chloromethane-ketone derivatives of papain allowed the identification of residues that take part in the interaction at this particular site (Drenth *et al.*, 1976). A hydrophobic cleft formed by residues Pro^{68}, Val^{133}, Val ^{157}, Asp^{158} and Ala^{160}, was postulated as the major region of interaction through van der Waals contacts with the preferred Phe.

The crystallization of human cathepsin B allowed a detailed structural comparison, and provided the first clues towards the understanding of the particular specificity of this enzyme (*Musil et al.*, 1991). Although the general fold of papain and cathepsin B are remarkably similar, an extra loop exists in the mammalian enzyme. This loop (comprising residues 108-125) lies over the top of the active site cleft, partially restricting the access to the catalytic triad, and for this reason is known as the "occluding loop". Two histidine residues in the occluding loop (His^{110}, His^{111})

positioned over the S' subsites, have been implicated in the dipeptidyl-carboxy-peptidase activity of cathepsin Bs.

The residues constituting the S_2 subsite in cathepsin B were almost all different to the ones found in papain. These discrepancies correlated well with the preferences observed with synthetic substrates. Cathepsin B is capable of hydrolyzing substrates with Arg in P_2 at much better rates than papain or cathepsin L, and this property has being used as a catalytic signature for cathepsin B (Barrett and Kirschke, 1981). The bottom of the S_2 subsite is formed by a Glu moiety at position 245 in cathepsin B, while a Ser is situated in the corresponding location in papain. It has been postulated, based on the analysis of crystalline structures of cathepsin B that Glu^{245} interacts with the positively charged side chain of the Arg at P_2 (Jia *et al.*, 1995). Similar results have been advanced previously by protein engineering experiments with papain. Khouri *et al.* (1991) have constructed, by site directed mutagenesis, single and double mutants of papain bearing the amino acid residues present in the equivalent position in cathepsin B, in an attempt to comprehend the substrate specificity differences of both enzymes. The mutations Ser^{205}-Glu, and Val^{133}-Ala increased the activity towards the synthetic substrate Z-Arg-Arg-AMC. A double mutant had a k_{cat}/K_M value (a parameter that measures substrate specificity) closer to cathepsin B, and showed a similar pH profile to the mammalian enzyme. The activity was only observed at a pH where the glutamic acid is ionized and able to form a salt bridge with the Arg moiety in P_2 (Khouri *et al.*, 1991). Moreover, Hasnain *et al.* (1993) demonstrated that Glu^{245} plays an important role in the stabilization of the transition state complex of cathepsin B with substrates bearing an Arg at P_2.

Further insights into the role of the glutamic acid at the bottom of the S_2 subsite was provided by the kinetic analysis of the cysteine proteinase of the baculovirus of *Autophaga californica*. This enzyme, although showing better sequence homology to cathepsin L than to cathepsin B, has a very high specific constant (K_{cat}/K_m) for the cathepsin B specific substrate Z-Arg-Arg-AMC. Remarkably this protein also has a glutamic acid in the position equivalent to Glu^{205} in papain. The pH activity profiles of the baculovirus enzyme suggest that the Glu residue must be ionized to allow a significant rate of hydrolysis of the cathepsin B substrate, supporting the idea of an ionic interaction between the Arg in P_2 and the

Glu in the bottom of the S_2 subsite. Although the enzyme is similar to cathepsin B in its preference for Arg, these enzymes differ in the affinity for neutral hydrophobic residues at P_2. The baculovirus can hydrolyze with similar efficiencies substrates with Phe, Val, or Leu, while all vertebrate cathepsins display low activity towards Val, and high activity towards either the P_2 Phe or Leu peptides. This could indicate that the binding pocket of the viral enzyme is sterically less restricted than that of the mammalian cathepsins (Bromme and Okamoto, 1995).

The S_2 subsite of cathepsin S also shows variations that can be correlated with the preferences towards synthetic substrates. This enzyme prefers to accommodate leucine or valine residues in the P_2 pocket rather than phenylalanine. The reaction of cathepsin S with a peptide inhibitor bearing a Val in the P_2 position is 300 times faster than cathepsin L (Shaw *et al.*, 1993). From sequence alignment and by using the papain crystal structures to define the S_2 subsite, it was suspected that the presence of a Phe at position 205 (at the bottom of the S_2 subsite) was responsible for the preference for shorter hydrophobic side chains at this site. However, site-directed mutagenesis analysis of cathepsin S (towards the residues present in the closely related cathepsin L) showed that the bias towards small amino acids in the P_2 position was more dependent on the side chains at position 133 (Gly in cathepsin S, Ala in cathepsin L) than in the moiety at position 205 (Phe in cathepsin S, Ala in cathepsin L). The substitution of Phe^{205} by Glu mimicked well the cathepsin B specificity, increasing the activity towards substrates with Arg in P_2 (Bromme *et al.*, 1994).

Although the specificity of proteinases of the papain superfamily seems to be influenced mainly by the interactions at the P_2-S_2 subsite, some interesting differences also occur in the S1 subsite. The most interesting example is papaya protease IV, also known as glycil endopeptidase. This enzyme accepts almost exclusively glycine at the S1 subsite. Comparison of the sequences of this proteinase with the closely related papain, showed that two conserved glycine residues in papain (positions 23 and 65) are replaced by glutamic acid and arginine in the glycyl endopeptidase. The role of the amino acids present in this position towards the particular specificity of glycyl endopeptidase was assessed by site directed mutagenesis of these positions in cathepsin B (that have the conserved glycine residues) to Glu and Arg. Single mutants showed reduced activity against substrates with an arginine moiety in P_1 (a natural substrate for cathepsin B), but were capable of efficiently cleaving synthetic substrates with a glycine at P_1. Double mutants canceled all the enzymatic activity against different substrates, showing that the side chains of the glutamic acid and arginine can actually block completely the access to the active site cleft (Fox *et al.*, 1995). The recent crystallization of glycyl endopeptidase confirmed the suspicion that these substitutions are responsible for the restricted S_1 specificity of the enzyme. The side chains of Glu^{23} and Arg^{65} form a barrier across the binding pocket, excluding any substrate with large side chains from the S1 position of the active site cleft. Furthermore, this restriction can explain the inability of cystatins to inhibit this proteinase, as they cannot establish the required contacts for an effective inhibition (O'Hara *et al.*, 1995).

Cathepsin B also shows a peculiar S_1 subsite. The crystal structure of rat cathepsin B shows residues 191-198 folding in a different direction than in other members of the papain superfamily crystallized so far. As a consequence, a methionine at position 196 is brought closer to the active site cleft, partially blocking the entrance of large side chains in the S_1 subsite (Jia *et al.*, 1995).

The role of the S_3-P_3 interactions in the binding specificity have been analyzed for rat cathepsin B using fluorogenic synthetic substrates with different residues in P3. The kinetic constants showed a decreasing order of preference of Tyr, Val, Arg, Gly and Glu, although other residues were not tested. By enzyme modeling, a hydrophobic pocket defined by Asp^{69}, Asn^{72}, Gly^{73}, and Tyr^{75} could be recognized; the phenyl ring of Tyr^{75} and the alpha carbon Gly^{73} are the proposed major elements in the contact surface (Taralp *et al.*, 1995).

The S' subsites have not been well characterized because the synthetic substrates do not extend in this direction. However, based on the crystallographic data some properties of these binding pockets have been advanced (Storer and Menard, 1996). The analysis of the crystal structure of the complex of cathepsin B with the inhibitor CA030 (Turk *et al.*, 1995) showed that the S_2' site is not very selective, and the S_1' subsite consist of an hydrophobic pocket delimited by Val^{176}, Leu^{181}, Met^{196} and Trp^{221}, and can bind hydrophobic amino acids. Menard *et al.* (1993) investigated the preferences in the S_1' subsite for cathepsin B, L, S and papain. It was found that cathepsin B prefers to accommodate large hydrophobic side chains (Trp, Tyr, Phe, Leu), while cathepsin L has a predilection for small residues (Ala, Ser) or long non branched residues (Asp, Gln, Lys). The specificity for papain and cathepsin S seems to be wider. In general the study points out that side chain contacts in the S_1' subsite might not play a major role in specificity , with a broad range of amino acids accepted in P' positions.

The above studies illustrate that in spite of great similarities in cysteine proteinases of the papain superfamily, differences do exist. These differences can be pointed out by a combination of structure and activity analysis, and can constitute the basis for the designing of selective inhibitors. However, most of these studies take advantage of the use of small synthetic substrates, and the picture might be different if we consider natural protein substrates. In this case, requirements for interaction in a particular subsite could be overruled by the interactions of the polypeptide chain with other subsites, or by the steric restrictions imposed by the surface contacts of the enzyme and the substrate further away from the binding site cleft.

1.3.4 The propeptides in folding and inhibition

The cysteine proteinases of the papain superfamily are synthesized as pre-pro-enzymes. Early studies in vertebrate cathepsins indicated that the removal of the propeptide takes place under acid conditions when the enzyme reaches the lysosomes (Mason *et al.*, 1987, Nishimura and Kato, 1987a, Nishimura and Kato, 1987b, Nishimura and Kato, 1988, Nishimura *et al.*, 1988). Experiments with proteinase inhibitors indicate that this processing might be autocatalytic , an hypothesis supported by the accumulation of proenzymes in mutants of the active site (Salminen and Gottesman, 1990, Vernet *et al.*, 1991, Rowan *et al.*, 1992, Mach *et al.*, 1993, Mach *et al.*, 1994a, Nishimura *et al.*, 1995, Kopitar *et al.*, 1996). However, in several cases it has been found that the primary site of cleavage does not correspond to the N-terminal sequence of the mature enzyme, and semi-mature forms with short N-terminal extensions that are later trimmed (probably by exopeptidases like cathepsin C) are also detected (Rowan *et al.*, 1992, Mach *et al.*, 1993, Ishidoh and Kominami, 1994).

The propeptide is essential for the proper folding of the enzyme, since deletions of this region result in proteins that are highly trypsin sensitive, are not susceptible to mannose phosporylation, and are degraded in the endoplasmic reticulum (Tao *et al.*, 1994). A conserved motif was found in the proregion of members of the papain superfamily, consisting of Gly-Xaa-Asn-Xaa-Phe-Xaa-Asp (residues -42 to -36 in papain). Mutation of these residues indicated that some substitutions were accepted and the zymogen was processed normally, but others led to degradation of the protein probably due to misfolding. The charge of the Asp residue at position -36 seem to be important for the pH dependence of processing (Vernet *et al.*, 1995).

The folding process seems to be closely related to the sorting of the proteins to their lysosomal destination or secretion, probably via post-translational modifications. In this sense, a region in the propeptide of cathepsin L rich in basic residues (residues 33 to 41 procathepsin L numbering) that bears similarities with yeast vacuolar-sorting sequences have been identified as interacting with a membrane receptor, and this association has been implicated in the targeting of this enzyme to the lysosome through a mannose-6-phosphate independent mechanism (McIntyre and Erickson, 1991, McIntyre and Erickson, 1993, McIntyre *et al.*, 1994). Recent mutagenesis studies in this region, indicate that replacement of His^{36}, Arg^{38} or Tyr^{40} results in glycosylation at a second cryptic site in mouse procathepsin L (probably due to misfolding) and as a consequence the mutated protein is secreted (Chapman *et al.*, 1997). Furthermore, previous studies have indicated that mouse cathepsin L mutated at the normal functional glycosylation site (Asn^{204}) is properly folded and stable, but is predominantly secreted (Kane, 1993). Procathepsin B expression in yeast leads to the secretion of the precursor and mature forms (Rowan *et al.*, 1992), while procathepsin L is sorted to the yeast vacuole (Nishimura and Kato, 1992). In several cases, the secreted forms of cysteine proteinases correspond to proenzymes (Mason *et al.*, 1987, Mach *et al.*, 1993, Tagami *et al.*, 1994, McDonald and Emerick, 1995, Okamura *et al.*, van der Stappen *et al.*, 1996). It has been pointed out that the secreted forms of vertebrate cathepsins are more stable than the lysosomal enzymes (Barrett and Sloane, 1996). Mach *et al.* (1994a) demonstrated that secreted active cathepsin B forms a non-covalent complex with its propeptide, which can be dissociated under acidic conditions.

Previous studies have demonstrated that the propeptide of cathepsin B was a powerful slow-binding inhibitor of the mature enzyme. The inhibition was pH dependent being maximal at pH 6.0 (K_i 0.4 nM) but with a marked drop of the inhibition constant at lower pH values, due to an increase in the dissociation rate. In addition, the cathepsin B propeptide showed a marked specificity for cathepsin B, as the inhibition of papain activity was in the micromolar range (Fox *et al.*, 1992).

The proregions of the related papaya enzymes papain and glycyl endopeptidase were analyzed as inhibitors of their cognate enzymes and the closely related enzymes caricain, and chymopapain. Both propeptides inhibited all the papaya proteinases with inhibition constants in the nanomolar range, except for glycyl endopeptidase, which showed a lower inhibition constant. However, this enzyme was better inhibited by its own propeptide than by the papain propeptide. Furthermore, cathepsin B and L were inhibited in the micromolar range (Taylor *et al.*, 1995a, Taylor *et al.*, 1995b).

Similarly, the propeptide of cathepsin L is a very specific inhibitor of its cognate enzyme (k_i 0.08 nM), being also capable of inhibiting cathepsin S, but with an inhibition constant five hundred times higher. This propeptide showed no inhibition of cathepsin B or papain. A marked pH dependence of the process was also detected (Carmona *et al.*, 1996). In a similar study, the proregion of a cysteine proteinase of the ciliate *Paramecium tetraurelia* was capable of inhibition of its cognate enzyme and a second *P.tetraurelia* cysteine proteinase, but exhibited no inhibition of papain, cathepsin B or cathepsin H (Volkel *et al.*, 1996). Taken together, these results indicate that cysteine proteinase propeptides are potent and highly selective inhibitors of their cognate enzymes, and can also inhibit closely related enzymes but with much lower efficiency.

The crystallization of procathepsin B, procathepsin L and procaricain provided a framework to a better understanding of the inhibition process. The proregion of procathepsin B folds over cathepsin B, and interacts with the substrate binding cleft. The N-terminal portion of the proregion folds into a hairpin formed by a three turn alpha helix, a tight bend and a short beta strand, maintained together by hydrophobic interactions; a loop connected this hairpin to a second short alpha helix, that sits on the S' side of the active site, and helps fasten the polypeptide chain to be accommodated in the substrate binding cleft ; from there the prosegment follows an unordered and highly mobile path on the surface of the mature enzyme towards the first amino acid of the mature enzyme. The hairpin super-structure interacts with an aromatic cluster in a loop (now named the prosegment binding loop PBL) at the top of the R domain of cathepsin B. A short beta sheet is formed between the propeptide beta strand and residues of the PBL, and aromatic interactions are also established in this region, and together function as an anchor for the propeptide over the surface of the mature enzyme. The substrate binding cleft is shielded from exposure to the solvent by the extended polypeptide chain that enters the cleft, but in the opposite direction to substrates, i.e., it binds in a reverse mode (Cygler *et al.,* 1996, Turk *et al.,* 1996).

Although cathepsin Bs have shorter propeptides than the remaining members of the papain superfamily, the crystal structure of procathepsin L and procaricain showed that the general fold of the propeptides was similar. Consequently the inhibition mechanisms were also alike. The N-terminal portion of these propeptides constitute a globular domain formed by three alpha helixes packed together and a short beta strand that makes contacts with the mature enzyme; the propeptide continues as an extended polypeptide chain that after traversing the active site cleft, adopts an unordered conformation. The core of the globular domain is a long second alpha helix, with the C-terminal turn of the first helix packed against one side, and the N-terminal of the third helix packed against the other. The central helix corresponds with a previously described conserved motif present in non cathepsin B cysteine proteinases, the ERFNIN motif (Karrer *et al.*, 1993). The C-terminal end of the ERFNIN helix is topologically homologous to the first helix of procathepsin B, and from there on the structures of procathepsin L, procaricain and procathepsin B are topologically similar. In cathepsin L and also in procaricain the most important interactions are established in the prosegment binding loop, and in the active site cleft . The binding to the active site cleft is also in reverse orientation, and shields the active site from exposure to the solvent (Coulombe *et al.*, 1996, Groves *et al.*, 1996).

Further evidence of the importance of the regions of contacts is given by the analysis of the inhibition of cathepsin L by truncated propeptides. It was shown that the inhibition is much weaker with a propeptide lacking portions of the globular domain (Carmona *et al.*, 1996). Using two series of peptides corresponding to successive deletions every five residues both from the amino and carboxy- terminus of the proregion of cathepsin B, the most important regions for the inhibition were delimited. Two regions were identified, one corresponding to the hairpin structure that interacts with the propeptide binding loop (PBL) and the other to the residues that make contacts with the active site cleft (Chen *et al.*, 1996). In an independent study, using short overlapping peptides (15mers) spanning the proregion of cathepsin B, a minimal region with good inhibitory activity (in the micromolar range) was identified, and corresponded to the region that was shown by crystallography to interact with the active site cleft (Chagas *et al.*, 1996).

These results, indicate that the proregions have evolved to play very different roles in their interaction with the mature enzyme. The high specificity shown by the propeptides as inhibitors, opens the possibility of designing highly specific inhibitors of cysteine proteinases based on the interactions established by the propeptide and the corresponding mature enzyme.

1.4 Proteinases of Parasites

Serine proteinases have been identified in several parasitic organisms, including unicellular parasites like Trypanosomatids (Cazzulo,1991, Londsdale-Eccles, 1991), and *Plasmodium* (Rosenthal, 1991). In helminths, serine proteinases are present as minor components of excretion secretion products. At least three serine proteinases are present in the secretions of the acetabular glands of *Schistosoma* cercariae. The major component seems to be a 28/30 kDa proteinase, and the existing evidence suggests that this enzyme (or a combination of enzymes produced by the acetabular gland) is involved in skin penetration and shedding of the glycocalyx of the cercaria. A membrane bound serine proteinase has also been identified in the tegument of *Schistosoma*, and is the prime suspect for the cleaving of host complement and Fc receptor-bound antibodies (Dalton and Brindley, 1997).

A dipeptidyl-peptidase secreted by all the vertebrate stages of *Fasciola hepatica* was isolated. This enzyme was characterized as a serine proteinase of more than 200 kDa, and has optimal activity at neutral pH, and differs from mammalian dipeptidyl-peptidases in substrate specificity and susceptibility to inhibitors. It was postulated that this enzyme may take part in the latter stages of protein digestion of host macromolecules, providing dipeptides that can be adsorbed as nutrients (Carmona *et al.*, 1994).

Two serine proteinases have been identified in secretions of the frog lungs' trematode *Haplometra cylindracaea,* as well as several cysteine proteinases (Harthorne *et al.*, 1994). In *Diplostomium pseudopathaceum,* a trematode parasite of aquatic birds, a serine proteinase has been identified in the cercarial cecum. This enzyme is most active at pH 8.0 and is dependent on calcium and magnesium ions for stability (Moczon, 1994b). Similarly, in the cestode *Schistocephalus solidus* a chymotrypsin-like proteinase with collagenolytic activity was identified in procercoids (Poltzer and Conradt, 1994). A trypsin-like serine proteinase and a serine proteinase inhibitor were purified from the nematode *Anisakis simplex* by Morris and Sakanaki (1994).

Aspartic proteinases have also been identified in malarial parasites, as minor components of the hemoglobin degradation complex (Rosenthal, 1991). In helminths a single copy gene with homology to cathepsin D has been isolated in *Schistosoma japonicum,* and a similar gene seems to be present in *Schistosoma mansoni*. A cathepsin D-like enzyme has been previously localized to the gastrodermis, cecal lumen, dorsal tegument and tubercules of male schistosomes, indicating that this gene, is widely expressed. The secreted enzyme (possibly originated in the digestive tract) is capable of digesting haemoglobin in vitro (Becker *et al.*, 1995).

Metalloproteinases have been detected in trichomonads, in *Trypanosoma cruzi*, in *Leishmania donovani* and in the cestode *Protocephalus ambliplitis*. Several proteinases inhibited by the chelating agent EDTA have been identified in *Trichomonas tenax, Trichomonas vaginalis,* and *Tritichomonas mobilensis* (North, 1991). An alkaline membrane-associated proteinase of 60 kDa has been identified in *Trypanosoma cruzi*. The major surface protein of the promastigotes of *Leishmania* *donovani* is a membrane bound metalloproteinase more usually known as PSP (for Promastigote Surface Proteinase) or gp63 (for its molecular wight of 63kDa) Surface metalloproteinase activity has been detected in several other *Leishmania* species. This proteinase has been widely used as a vaccine target due to its abundance and ease of purification, with conflicting results (Etges and Bouvier, 1991).

In plerocercoids (cysts) of the cestode *Protocephalus ambliplitis,* a proteinase of 30 kDa with optimal activity at pH 9.0 has been identified. This proteinase has collagenolytic, hemoglobinlytic and partial elastinolytic activity, and is inhibited by chelating agents, but the inhibition can be reversed by the incubation with Zn ions, confirming that this enzyme activity belongs to the metalloproteinase group (Poltzer *et al.*, 1994).

1.5 Cysteine Proteinases of Parasites

A growing body of evidence has indicated that parasite cysteine proteinases are the major protein hydrolases involved in the interaction with their hosts and in pathogenesis. Several cysteine proteinases have been detected and purified from secretions of parasitic organisms. More recently, several cysteine proteinase genes have been cloned by PCR using primers homologous to the cysteine proteinases found in other organisms. However, as cysteine proteinases are widespread in eukaryotes, it is not surprising to find them in parasitic organisms, and several of them could be playing housekeeping roles different to the ones involved in the interaction with their hosts and pathogenesis.

1.5.1 Cysteine Proteinases of unicellular parasites

Cysteine proteinases are secreted by the most primitive eukaryotes, like the amitochondriate diplomonads and triplomonads. In the diplomonad *Giardia*, three genes similar to the vertebrate cathepsins B have been isolated. One of them has been identified as the proteinase released during the excystation of trophozoites in the digestive tract of their host. Further evidence of its role in the parasitic life cycle is provided by the fact that cysteine proteinase inhibitors block the excystation process. The expressed genes localize to vacuolar bodies. The excystation cathepsin B gene and a second expressed cathepsin gene map to chromosome IV, while the third gene is also single copy and maps to chromosome II. There is no evidence of expression of this third gene (Ward *et al.*, 1997).

A family of cysteine proteases more related to the vertebrate cathepsin Ls and than to cathepsin B has been identified in the trichomonads *Trichomonas vaginalis* and *Tritrichomonas foetus,* that parasitize the reproductive tract of humans and cattle, respectively. Cysteine proteinases have been identified as a main component in the culture media of axenically grown parasites, indicating that at least some of these are secreted. Up to four different genes have been identified in *T. vaginalis*, three of them being single copy and the fourth a member of a low copy multigenic family. All the genes are expressed, as detected in Northern blots (Mallison *et al.*, 1994). Similarly, up to nine different sequences have been reported from *T. foetus*. The sequences show a greater diversity, some of them being more distant than human cathepsin L is to cathepsin H. Only one of the *T. foetus* cysteine proteinases is a member of a low copy multigenic family, while the remaining are single copy genes, and they are expressed to different levels (Mallison *et al.*, 1995). There is no clear evidence to indicate which of the multiple genes from trichomonads encode the secreted enzymes that may be important for its role in pathogenesis, and which encode lysosomal enzymes.

There is a correlation between secretion of cysteine proteinases and virulence in *Entamoeba hystolitica*. Several genes encoding cysteine proteinases that are more related to vertebrate cathepsins H and L have been identified, and three highly expressed proteinases have been related to virulence and hystolysis. However, the non-pathogenic *E. dispar* contains four genes highly homologous to the ones present *in E. hystolitica*, including one of the highly expressed genes, indicating that not all the cysteine proteinases are involved in pathogenesis (Bruchhaus *et al.*, 1996). Cysteine proteinases are also secreted by the facultative parasitic amoebas *Acanthamoeba* and *Naegleria,* and have been related to the hystolysis produced by these organisms (Mitro *et al.*, 1994, Aldape *et al.*, 1994).

In trypanosomatids several cysteine proteinases have been identified, and the *Trypanosoma cruzi* representatives have been called cruzipains (Cazzulo *et al.*, 1997). Cruzipains are encoded by multicopy genes arranged in tandems on different chromosomes. As many as 130 copies may be present in certain strains of *T.cruzi* (Campetella *et al.*, 1992). A similar multicopy gene family arranged in tandem is found in other species of *Trypanosoma* (Lonsdale-Eccles, 1991, Tanaka *et al.*, 1994, Martinez *et al.*, 1995), and in species of the genus *Leishmania*, although the copy number within the tandems seem to be restricted to around 20 in this genus (Souza *et al.*, 1992, Traub-Cseko *et al.*, 1993, Sakanaki *et al.*, 1997). Studies with general cysteine proteinases inhibitors demonstrate that these can block the transformation of trypomastigotes to epimastigotes and amastigote replication, indicating that these proteinases are vital to the life cycle of the parasite (McKerrow *et al.*, 1995). Several of these genes are transcribed, and are developmentally controlled, exhibiting higher expression levels in the replicating forms, especially the epimastigotes. There is evidence that different substrate preferences could exist within the gene family, and that different genes of the cluster are differentially expressed during the cycle of the parasite (Mottran *et al.*, 1977). The coded enzymes contain a long C-terminal extension (over 100 amino acids) joined to the catalytic domain by sequence motifs comprising threonine, proline and serine residues. The function of this terminal extension is still unknown, although it has proven to be highly immunogenic. It is not essential for proteinase activity, as recombinant cruzipain lacking the C-terminal extension exhibits the same activity as the native counterpart, and is not involved in the targeting to the lysosomes. Furthermore, some of the fully functional variants expressed on *L.mexicana* have a truncated C-terminal domain (Mottran *et al.*, 1977). Although these enzymes have been localized to lysosomes or lysosomal-like organelles, they also have been detected on plasma membranes. Cruzipain is identical to the glycoprotein GP57/51, the major antigen in Chagas disease patients. It is still to be determined which of the members of the multigene family are effectively lysosomal, and which are on the surface.

A second cysteine proteinase of the papain superfamily has been identified in species of the genus *Leishmania*. These proteinases are encoded by a single copy gene, and show stage specific expression (Robertson *et al.*, 1996). A cathepsin B-like proteinase encoded by a single gene is also present in *L. mexicana* and *L. major*

(Robertson *et al.*, 1996, Sakanaki *et al.*, 1997). Gene knockout experiments with *L. mexicana* have demonstrated that none of the three types of cysteine proteinases present in this species is individually essential for the life cycle of the parasite, although complementation between the different genes has not been ruled out (Robertson *et al.*, 1996).

A cysteine proteinase termed falcipain has been identified in the food vacuole of the causative agents of human malaria, the apicomplexans *Plasmodium falciparum* and *P.vivax* (Rosenthal *et al.*, 1993). Similar enzymes have been detected in several species of the genus *Plasmodium* (Rosenthal, 1996). Although originally this activity was related to the hemoglobinase activity of *Plasmodium*, new evidence points at a secondary role for this enzyme in relation to plasmepins, aspartic proteinases also present in the food vacuole (Francis *et al.*, 1996). As many as five different cysteine proteinases seem to be present in *P.falciparum*. Three other genes have been cloned, and are characterized by a serine stretch (Knapp *et al.*, 1989). Falcipain is expressed mainly in the intra-erythrocite ring stage, and for that reason is the only one that can be implicated in hemoglobin degradation. The remaining genes are expressed during the schizont stage of the parasite. All the *Plasmodium* genes are characterized by a very long propeptide, unique in the cysteine proteinases of the papain superfamily. However, the C-terminal region of the propeptide bears some similarity to the remaining propeptides.

Cysteine proteinases have also been found in another apicomplexan, *Theileria*, a parasite of erythrocytes of ruminants. The cysteine proteinases of two species of *Theileria* also have an unusual long propeptide, but not as long as the one present in *Plasmodium* (Nene *et al.*, 1990, Nene *et al.*, 1992, Baylis *et al.*, 1992).

A family of cysteine proteinases has been identified in the free living slime mold *Dictyostelium discoideum* (North *et al.*, 1988). The cysteine proteinases present in this organism are developmentally regulated, with a dramatic increase in their expression at the end of the vegetative growth and start of the differentiation process. Six different genes have been isolated, and it is probable that more cysteine proteinases are present (Williams *et al.*, 1985, Pears *et al.*, 1985, Souza *et al.*, 1995, Ord *et al.*, 1997). The sequence of four of the *D.discoideum* cysteine proteinases show insertions not present in other cysteine proteinases, rich in serine residues. Interestingly, it has been found that *D.discoideum* cysteine proteinases carry phosphoglycosil moieties (N-acetyl-glucosamine-1-P) linked to serines residues. Deletion of the serine rich region yields a product that no longer carries GlcNAc-1-P residues, indicating that this particular region is the target of the modification (Ord *et al.*, 1997).

1.5.2 Cysteine Proteinases of helminths

1.5.2.A Nematodes

Based on the use of specific cysteine proteinase inhibitors, cysteine proteinase activity has been detected in the secretions of the nematodes *Dirofilaria immitis* (Richer *et al.*, 1992), *Angiostrongylus cantonensis, Ascaris suum* (Maki and Yanagisawa, 1986), *Haemonchus contortus* (Karanu *et al.*, 1993), *Ancylostoma caninum* (Dowd *et al.*, 1994), *Strongylus vulgaris* (Caffrey and Ryan, 1994), *Heterodera glycines, Globodera pallida* (Lilley *et al.*, 1996), and *Trichuris suis* (Hill and Sakanaki, 1997) and in somatic extracts of *Dictyocaulus viviparus* (Rege *et al.*, 1989), *Trichuris muris* (Drake *et al.*, 1994) and *Nippostrongylus brasiliensis* (Kamata *et al.*, 1995). Most of these enzymes are secreted from the gut of adult worms, or in other cases from third or fourth larval stages, and are, in most of the cases, important immunogens.

The enzymes in the nematode secretion products have usually maximum activity at acidic pH, and are capable of Z-Phe-Arg-NHMec cleaving activity, but in several cases have no Z-Arg-Arg -NHMec cleaving activity (Dowd *et al.*, 1994, Caffrey and Ryan, 1994, Drake *et al.*, 1994, Lilley *et al.*, 1996, Rhoads and Fetterer, 1995, Hill and Sakanaki, 1997). This activity is consistent with the presence of cysteine proteinases similar to cathepsin L and papain, and not to cathepsin B. However, with the exception of a cysteine proteinase from *Toxocara canis* and an EST from the free living nematode *Caenorhabditis elegans,* all the cysteine proteinase genes cloned so far from nematodes are similar to the cathepsin B of vertebrates.

A developmentally regulated family of cathepsin B-like enzymes has been identified in *H. contortus* (Cox *et al.*, 1990, Pratt *et al.*, 1990, Pratt *et al*, 1991). At least two of the five genes cloned are linked in tandem. The enzymes seem to be highly abundant in the adult worm and expressed to lower levels in the larval stages. A very similar small gene family has been detected in *Ostertagia ostertagi* (Pratt *et al.*, 1992). Two cathepsin B-like genes were detected in *Ancylostoma caninum*, and localized to the gut (Harrop *et al.* , 1995). PCR with primers designed on conserved regions of cysteine proteinases of nematodes, allowed the detection of cysteine proteinase genes from *Strongyloides ratti, S. stercoralis, Ancylostoma caninum* and *C. elegans* (Harrop *et al.*,1995). All the genes isolated were cathepsin B-like genes.

In the free-living nematode *Caenorhabditis elegans* a family of developmentally regulated cathepsin B-like genes have been detected (Ray and McKerrow, 1992, Larminie and Johnstone, 1996). These genes although similar to the ones in *H.contortus* and *O.ostertagi*, are more divergent. Four of the genes map to different localizations on chromosome V, and another gene map to chromosome X. Different patterns of temporal expression have been detected, some of them being more abundant in the adult stage, others in the different larval stages. One gene has been localized to the intestine of adult and larval worms (Ray and McKerrow, 1992).

The contradiction of finding cathepsin B-like genes but cathepsin L-like activities was analyzed by modelling the secreted cysteine proteinase of the hookworm *Ancylosotoma caninum*. A substitution of a Tyr residue to a Trp in the S_2 subsite position 75 (69 according to papain numbering), and substitutions in the occluding loop were considered responsible for the difference. The Tyr residue in cathepsin Bs have aromatic hydrophobic interactions with a Phe side chain occupying the S_2 subsite. Also, the guanidino group of an Arg occupying the P_2 position can interact with the negatively charged phenolic ring of the Tyr^{75}. While the aromatic interactions with a Phe would be enhanced by a Trp in position 75, the positive charge of the indole ring of Trp would prevent this interaction with Arg in the P_2 position. Furthermore, a negatively charged Glu^{122} in the occluding loop of vertebrate cathepsins B which allows an electrostatic interaction with the guanidino group of the Arg, has been substituted by an hydrophobic stretch, in the hookworm enzyme (Brinkworth *et al.*, 1996). This would explain the preference for cathepsin L-like substrates by a cathepsin B-like enzyme. However, a non-cathepsin B-like enzyme has been cloned in *Toxocara canis* (Loukas *et al.*, 1997), and a closely related cysteine proteinase has been detected in a contig from chromosome III of *C.elegans* (EMBL locus CER07E3, accession Z49207). Furthermore, a partial sequence of the plant cyst nematode *Heterodera glycines* shows homology to other invertebrate cathepsin L-like cysteine proteinases, indicating that cathepsin L-like enzymes exist in nematodes.

Although nutritional roles have been ascribed to the cysteine proteinases secreted by the blood feeding nematode *Haemonchus conturtus* and *Ascaris suum,* and to the plant parasitic nematode *H.glycines,* this might not be the only function of the secreted enzymes. It has been demonstrated that the secretions of *H. contortus* are capable of degrading the proteins of the extracellular matrix (Rhoads and Fetterer, 1996). Furthermore, cysteine proteinases have been involved in the molting process in several nematodes. Cysteine protease inhibitors can arrest the L3 to L4 larva molting in *Dirofilaria immitis* (Richer *et al.*, 1993) and *Onchocerca volvulus* (Lustigman *et al.*, 1996).

1.5.2.B Cestodes

In cestodes cysteine proteinase activities with particular properties have been detected. Cysteine proteinases have been detected in *in vitro* secretions of the oncosphere of *Taenia saginata* as well as other proteinase activities related to the serine proteinase type (White *et al.*, 1996). In cysts of other taenids several proteinase activities have been detected, including metalloproteases, aspartic proteases and cysteine proteinases (White *et al.*, 1992, White *et al.*, 1997). The cysteine proteinase activities of taenid cestodes have been purified, and correspond to proteins of 18 kDa in *T.saginata,* 32 kDa in *T. solium* and 43 kDa in *T. crassiceps.* All show activity towards the synthetic substrate Z-Phe-Arg-NHMec, with maximum activities at acidic pH.

Taenia species have an active uptake of host immunoglobulins into the cysts, probably via specific receptors, and host immunoglobulins are found in the cystic fluids. Furthermore, immunoglobulins are degraded, in a process that starts at the cyst wall, and it has been proposed that the parasites actually use IgG as a major source of nutrient (Ambrosio *et al.*,1994). The cysteine proteinase purified from *T. crassiceps* cysts seem to be the major effector in the IgG cleaving activity, having at the same time an immuno-evasive function as well as participating in a nutrition related process.

In several species of *Spirometra* an acid cysteine proteinase has been detected in plerocercoids (Song *et al.*, 1992,Song and Chapell, 1993). The cysteine proteinase activity was purified from secretions and from tissue extracts of the plerocercoids of *S. mansoni* and *S. erinacei*, and corresponds to a protein of 28 kDa. The purified enzyme exhibits an acid pH optimum, and is capable of cleaving the substrate Z-Phe-Arg-NHMec, and also collagen, and to a lesser extent haemoglobin. Immunoglobulin cleaving activity of plerocercoids of *S. mansoni* were analyzed *in vitro*, and the enzyme responsible was purified. The resulting enzyme was also a cysteine proteinase of 28 kDa, with a pH profile almost identical to the previously purified enzymes. N-terminal sequencing revealed an enzyme similar to vertebrate cathepsin S (Kong *et al.*, 1994). A cysteine proteinase was cloned from *S. erinacei,* with homologies to the vertebrates cathepsin L. A molecular mass of 34 kDa was determined for *in vitro* translated products, that could correspond well with the proform of the previously characterized enzyme (Liu *et al.*, 1996). Two other cysteine proteinases have also been identified in these organisms, a small molecule of 21 kDa similar to vertebrate cathepsin B, and a proteinase of 57 kDa In all the cases, the enzymes were strongly recognized by sera from infected patients, indicating that they are important immunogens.

The plerocercoids of *Spirometra* have the ability to stimulate body growth of their hosts. The parasite releases a mammalian hormone-like substance, that interacts with growth hormone receptors. The purification of the plerocercoid growth factor produced a protein of 27.5 kDa with proteinase activity. The purified protein was a cysteine proteinase with maximum activity at slightly acid pH, and showed activity towards Z-Phe-Arg-NHMec. Furthermore, the addition of E64, while abolishing the proteinase activity, also greatly diminished the ability of the purified plerocercoid growth factor to compete with human growth factor for it's receptor (Phares and Kubik, 1996). A cDNA clone for the plerocercoid growth factor was obtained, and corresponds to a cysteine proteinase almost identical to the one cloned in *S. erinacei* (Phares, 1996).

1.5.2.C Trematodes

As in other parasitic organisms, the trematodes have a complex set of proteinases involved in the interaction with their hosts. Although this group comprises several species, very little is known on the biochemistry of the interaction with their hosts for most of them, although an interesting body of data exists for the human parasite *Schistosoma mansoni* , and the liver fluke *Fasciola hepatica.* The only other trematodes where the proteinases have been analyzed are the lung fluke *Paragonimus westermani*, the gull fluke *Diplostomum pseudopathaceum,* the chinese liver fluke *Clonorchis sinensis* and the lung fluke from frogs *Haplometra cylindrea* .

The cercariae of *Diplosotomum pseudophataceum*, a parasite of aquatic birds, actively penetrate the skin of fish, their intermediate host. While a serine proteinase has been found associated with the cercarial cecum, a neutral cysteine proteinase probably associated with the penetration glands could be responsible for the host invasion (Moczon, 1994a, 1994b).

Three cysteine proteinases with preferential activity towards Z-Arg-Arg-NHMec, a typical cathepsin B substrate, and a fourth with preferential activity towards the cathepsin L substrate Z-Phe-Arg-NHMec were found in the excretion/secretion products of *Haplometra cylindracea,* the lung fluke of the frog. The activities were correlated with proteins of 48, 23 , 14 and 55 kDa respectively. Previous hystochemical studies have indicated the presence of hydrolytic enzymes in the ceacal epithelium of the parasite, released upon a blood meal, and capable of degrading hemoglobin. It has been proposed that the identified cysteine proteinases, together with two serine proteinases also released, would be involved in hemoglobin degradation (Hawthorne *et al.*, 1994).

Evidence for several cysteine proteinases developmentally regulated were found in the chinese liver fluke, *Clonorchis sinensis.* Four different proteins seem to be present in the adult. An immunogenic protein of 18 kDa with cathepsin B-like activity has been purified (Song *et al.*, 1990), as has a secreted neutral cysteine proteinase of 24kDa, that seems to be related to the cytotoxity associated with

C.sinensis infections (Park *et al.*, 1995). Cysteine proteinase activity also exists in immature worms, but is higher in metacercariae, where a cysteine proteinase of 32 kDa has been purified (Song and Rege, 1991).

In the lung fluke *Paragonimus* several cysteine proteinases are evident at different stages of the life cycle of the parasite. In adults, an acidic activity that corresponds with a protein of 27-29 kDa, and a neutral activity corresponding to an immunogenic protein of 20 kDa, both capable of hydrolyzing hemoglobin, were detected independently in *P.ohirai* (Yamakami,1986) and *P.westermani* (Song and Kim, 1994), and probably are both present in each species. In extracts of *P.westermani* eggs a cysteine proteinase of 35kDa with maximal activity at pH 6.0 was purified (Kang *et al.*, 1995).

In excysting metacercariae several cysteine proteinases are present and at least some of them are secreted. Yamakami and Hamajima (1987) identified a neutral thiol protease (NTP) of 22 kDa. This protease is also found in culture supernatants of excysting metacercariae indicating that is secreted (Yamakami and Hamajima, 1990). Furthermore, the enzyme was localized to the digestive tract of the metacercariae. The purified NTP enzyme showed low specificity towards hemoglobin, has collagenolytic activity, and cleaved preferentially the carboxylic side of the basic residue in N-substituted peptides. A gene was isolated from a cDNA library using anti-NTP serum (Yamamoto *et al.*, 1994). The cloned enzyme showed weak homology to papain and vertebrate cathepsins L and H. Southern blot analysis showed the presence of a cysteine proteinase gene family in *P.westermani* (Hamajima *et al.*, 1994). Consistent with this, a 20 kDa, acidic proteinase was reported from metacercariae (Song and Dresden, 1990), and two other slightly acidic cysteine proteinases of 27 and 28 kDa, both capable of hydrolyzing collagen, fibronectin and myosin were also identified (Chung *et al.*, 1995). The 27 kDa acidic proteinase was purified , and characterized, showing a pH optimum of 4, and was capable of degrading albumin, immunoglobulins and complement components (Yamakami *et al.*, 1995).

Infection with *P.westermani* appears to induce immuno-tolerance, since repeated infections are not infrequent. Intraperitoneal injection of NTP into mice resulted in reduced expression of the major histocompatiblity complex and interleukin-II receptor on lymphocytes, induction of spleen supressor cells and also suppression of rejections of skin grafts. Collectively these data suggest that the neutral cysteine proteinase can induce immune supression and tolerance to parasite antigens, and is a key player in the immune evasion mechanism (Hamajima *et al*., 1994).

While only members of the papain superfamily have been detected in other trematodes, in *Schistosoma spp* cysteine proteinases of different families have been identified, including calcium activated neutral proteases (calpains), and legumains, together with vertebrate cathepsin like enzymes.

Calpains are heterodimeric cytoplasmic proteins formed by a large subunit of 80kDa and a small subunit of 30 kDa. Two different but structurally related types of calpains exist (CANP-I and CANP-II), and they have been implicated in the regulation of cytoskeletal proteins, receptors and protein kinases, and also in membrane biogenesis and cytoplasmic protein regulation. Both CANP-I and CANP-II activities exist in *Schistosoma* adult worms and have also been detected by northern blots in sporocysts. The large unit of a CANP from *S. mansoni* and *S. japonicum* have been cloned, and they show high homologies with their vertebrate counterparts, although the distribution of the conserved EF motifs found in calcium-modulated proteins is different to the vertebrate enzymes (Andersen *et al*, 1991). The enzymes have been immunolocalized to the tegumental syncytium and *in vitro* assays of incubation of worms with calpain inhibitors demonstrated an inhibition of the calcium-mediated uptake of methione and phosphatidylcholine, indicating a probable role in the biogenesis of the surface membrane (Siddiqui *et al*., 1993) .

The legumain from *Schistosoma mansoni* was originally identified as an hemoglobinase, corresponding with a highly immunogenic protein (Sm32) present in the gut of schistosomula and adult worms. The cloned gene show high similarity with legumains, cysteine proteinases from legumes involved in the post-translational modification of proteins and had a strict preference for cleavage after asparaginyl bonds (Davis *et al*., 1987, Merckelbach *et al*., 1994). Due to the low specific activity shown by the schistosome legumain, which is inconsistent with the large uptake of erythrocytes by the blood flukes, it has been proposed that these enzymes might modify post-translationally other proteases rather than being directly involved in the haemoglobin degradation. Asparaginyl residues are present in the C-terminal end of the propeptides of cathepsin-like proteinases from this organism, suggesting that they could be the target for the action of the asparaginyl endopeptidase (Dalton and Brindley, 1996).

Several cysteine proteinases of the papain superfamily are found in *Schistosoma*. A cathepsin B-like enzyme was first recognized as a potent antigen (Sm31) that was secreted from the parasite gut (Ruppel *et al.*, 1985). A cDNA encoding Sm31 was obtained from *S. mansoni*, and the homologous gene from *S. japonicum*, and a closely related gene bearing a substitution of the active site cysteine have also been cloned (Klinkert *et al.*, 1988, Merckelbach *et al.*, 1994). The cloned enzyme from *S. mansoni* expressed in insect cells was capable of cleaving the cathepsin B substrate Z-Arg-Arg-NHMec (Gotz and Klinkert, 1993). Analysis of the cathepsin B-like enzyme with synthetic and protein substrates indicated an activity profile similar to vertebrate cathepsins B, and also that this enzyme would not be the major component in the globin degradation (Ghoneim and Klinkert, 1995). Further evidence for cathepsin B not being the major hemoglobin degrading enzyme came from studies of recombinant *S.mansoni* cathepsin B expressed in *Saccharomyces cerevisiae*. The zymogen expressed in yeast was not capable of autoprocessing, but pepsin treatment generated the active mature enzyme. The activity towards synthetic substrates was similar to the one previously characterized, and the enzyme was capable of cleaving haemoglobin but did not show a marked substrate preference for this protein substrate (Lipps *et al.*, 1996). Molecular modeling of the fluke enzyme based on the coordinates of the human cathepsin B, showed a wider and more hydrophilic S_2 subsite than the human counterpart. The superimposition of inhibitors on the modeled structures revealed that, as a consequence of the differences in the binding cleft, a better binding of the inhibitor Z-Trp-MetCHN$_2$ is expected in the schistosome enzyme, consistent with a 15 fold stronger inhibitory activity for the schistosome enzyme than for the human cathepsin B (Klinkert *et al.*, 1994).

A predominant Z-Phe-Arg-NHMec cleaving activity at acidic pH was detected in soluble extracts of adults of *S. mansoni* and *S. japonicum*, and ascribed to a cathepsin L-like enzyme. By PCR with degenerate primers of the conserved regions of cysteine proteinases and futher screening of an adult cDNA library, an enzyme more similar to vertebrate cathepsin L than to cathepsin B was cloned. The cloned enzyme (SMCL1) showed a relatively low homology to vertebrate cathepsins L (46%), and it appeared to be a single copy gene according to Southern blot data (Smith *et al.*, 1994). A second cathepsin L-like enzyme from *S. mansoni* (SMCL2) was cloned from an adult cDNA library, corresponding to a proenzyme with higher homology to vertebrate cathepsins L . The expression of this enzyme is five times higher in female worms, and is encoded by a single copy gene. Immunolocalization studies indicate that this latter enzyme is expressed in the epithelium of the vitelloduct, the ovo-vitelloduct, the ootype and the uterus of female worms, while it is detected in the gynaecophoric canal in male worms, and its involvment as a lubricant agent lowering the viscosity of the fluid in the ovo-vitelloduct was proposed. (Michel *et al.*, 1995).

Two cysteine proteinases were cloned from an adult *S. japonicum* cDNA, one closely resembling SMCL1 (92% identical) and a second one with high homology to SMCL2 (78% identical) (Day *et al.*, 1995). An analysis of the activities of cathepsin L-like and cathepsin B-like enzymes of *S. mansoni* indicated that both enzymes were secreted by the adult worms. The cathepsin L-like specific activity was twice the one found for the cathepsin B-like activity, and sex related differences were also detected, with much higher activities in females. It was also observed that Z-Phe-Ala-CHN2 is a more potent inhibitor than Z-Phe-Phe-CHN2, in contrast to what is found in vertebrate cathepsin L, opening the possibility of using these enzymes as targets for specific antiparasitic drug design (Dalton *et al.*, 1996).

Cysteine proteinase activities have been detected in *S.mansoni* eggs, corresponding to acidic proteinases of aproximately 25 and 27 kDa (Asch and Dresden, 1979). Three cysteine proteinases were later purified from eggs, and showed a pH optimum at pH 5.5 for CBZ-Arg-Arg-AFC , the two most prominent being of 25kDa and 30 kDa (Sung and Dresden, 1986). In zymograms of soluble extracts of *S.japonicum* eggs activity towards Z-Phe-Arg-NHMec was detected (Day *et al.*, 1995). It was suggested that these activities were associated with the pathogenesis induced by the schistosome eggs, either directly or through immunological responses to the proteases. In support of this hypothesis, an IgG1 anti-cysteine proteinase monoclonal antibody was shown to inhibit the activity from eggs towards CBZ-Arg-Arg-AFC (Dresden *et al.*, 1983).

Proteinase activity corresponding to cysteine proteases was also detected in miracidial culture media and in extracts of miracidia and primary sporocysts. The miracidial activity was higher at early stages of transformation, and it was suggested that it could be involved in miracidial snail penetration. When sporocyst extracts where incubated with snail cell-free hemolymph, degradation of hemolymph proteins was detected, including the snail hemoglobin. This suggested a role of the parasite enzymes in establishment or maintenance of the infection in snail hosts. Two proteinases were purified from transforming miracidia medium, corresponding to a 19 kDa and a 36 kDa protein (Yoshino *et al.*, 1993).

Based on substrate preferences, cathepsin L-like activities and cathepsin B-like activities were found in cercariae and schistosomula of *S. mansoni*. These cysteine proteinases are present in the acetabular gland of cercariae, and they might collaborate with the cercarial serine proteinases in the process of skin penetration (Dalton *et al.*, 1997). Furthermore, using specific cysteine proteinase floromethyl-ketone inhibitors it was demonstrated that *in vitro* haemoglobin degradation was blocked in schistosomules. Moreover mice experimentally infected show reduction in worm burden and egg production when treated with the inhibitors (Wasilewski *et al.*, 1996).

A different cysteine proteinase, cathepsin C, a dipeptidyl-peptidase, has also been cloned from cDNA libraries of adult *S. mansoni* and *S. japonicum* (Butler *et al.*, 1995, Hola-Jamirska *et al.*, 1997). It has been hypothesized that this enzyme might play a role downstream in the degradation of host-derived hemoglobin to absorbable peptides, acting on fragments released after the action of the secreted endopeptidases (Dalton and Brindley, 1996).

Since the sixties several reports have indicated the presence of proteolytic enzymes in the excretion/secretion products of the common liver fluke *Fasciola hepatica*. Histochemical studies indicated that the predominant proteinase activity was associated with cells lining the digestive system of the parasite, and were then implicated in nutrition and tissue degradation (Howell, 1966, Halton, 1967, Howell, 1973, Simpkin *et al.*, 1980). A role in immune evasion of the secretions of *F. hepatica* was proposed based on the observation that excretion /secretion products of newly excised juveniles were capable of reducing the attachment of cells from *F. hepatica* resistant rats (Goose, 1978). Further studies demonstrated that acidic and neutral cysteine proteinases in fluke culture media were capable of cleaving immunoglobulins at the hinge region (Chapman and Mitchell, 1982). Furthermore, two closely migrating bands of 27 kDa corresponding to the cysteine proteinases were detected with sera of infected rats, indicating their possible role as antigens (Coles and Rubano, 1988). A cysteine proteinase of 14.5 kDa was purified from extracts of adult *F.hepatica* worms, capable of digesting hemoglobin, collagen and IgG, and active at pH between 4.5 and 7.5 with a maximum at pH6.0 (Rege *et al.*, 1989).

Multiple cysteine proteinase activities were detected in gelatin-substrate SDS-PAGE (GS-PAGE) of culture media of adult flukes and also of immature flukes, with activities in the acid (pH 3.0 to 4.5) or slightly acid to neutral range (pH 4.5 to 8.0) (Dalton and Heffernan, 1989). A single 27 kDa cysteine proteinase was purified in our laboratory by gel filtration and ion exchange chromatography. However, the purified enzyme showed several activity bands in non reducing GS-PAGE in the range of 60 to 90 kDa, indicating that at least some of the multiple bands detected previously were due to a single enzyme. N-terminal sequencing and substrate specificity analysis with synthetic substrates indicated that the enzyme has a cathepsin L-like activity, with a preference for hydrophobic residues in the P_2 position. Immuno-localization studies indicated the presence of the enzyme in secretory granules of the digestive epithelium of the parasite (Smith *et al.*, 1993). Identical localization was obtained independently for a 27 kDa cysteine proteinase of a japanese strain of *Fasciola sp.* (Yamasaki *et al.*, 1992). It was demonstrated that the secreted enzyme was capable of cleaving immunoglobulins at the hinge region (Smith *et al.*,1993). The purified enzyme was responsible for the prevention of the antibody mediated attachment of eosinophils to newly excysted juveniles of *F. hepatica* previously described by Goose (1978) (Carmona *et al.*, 1993). This enzyme was termed cathepsin L1.

A second cysteine proteinase of 29.5 kDa was also purified from excretory/secretory products of adult *F. hepatica* (cathepsin L2). The enzyme produced a multiple band pattern on GS-PAGE complementary to the cathepsin L1 pattern. The purified enzyme showed an unusual substrate specificity when analyzed with synthetic substrates. First, although both enzymes were capable of cleaving the cathepsin L substrate Z-Phe-Arg-NHMec, and had negligible activity towards the cathepsin B substrate Z-Arg-Arg-NHMec, and cathepsin H substrate Z-Arg-NHMec, they also showed higher specific activity towards the substrate Boc-Val-Leu-Lys-NHMec, an unusual feature for vertebrate cathepsins. Furthermore, cathepsin L2 can cleave substrates with proline in the P_2 position such as Tos-Gly-Pro-Arg-NHMec and Tos-Val-Pro-Arg-NHMec, with high efficiency whereas these are poor substrates for cathepsin L1, and for mammalian cathepsin L, indicating that *F. hepatica* cathepsin L2 is an enzyme with novel activity. In addition, both enzymes showed structural differences indicated by treatment with tetranitromethane (which nitrates tyrosine residues); only the cathepsin L2 was completely inactivated by the modification (Dowd *et al.*, 1994). Cathepsin L2 is capable of cleaving fibrinogen, generating a clot that differs from the one produced by thrombin. As *F. hepatica* feeds on blood, by puncturing the bile duct wall, it was proposed that cathepsin L2 -induced coagulation may prevent excessive bleeding from the lesions produced (Dowd *et al.*,1995). Both cathepsin L1 and L2 could be involved in tissue invasion, as shown by their capacity of degrading collagen, fibronectin and laminin (Berasain *et al.*, 1997).

With a similar approach to the one applied to *F. hepatica*, several proteolytic enzymes with immunoglobulin and globin cleaving activities were detected in extracts of adult *F. gigantica* worms, (Fagbemi and Hyller, 1991). A protein in the range of 26-28 kDa was further purified from *F. gigantica*, corresponding to a dominant cysteine proteinase (Fagbemi and Hyller, 1992). Since both *F. hepatica* and *F. gigantica,* and intermediate forms of these species exist in Japan, and considering the immuno-localization data of Yamasaki *et al.* (1992), it is very likely that this represents the homologue to one of the proteins isolated in our laboratory.

However, these are not the only cysteine proteinases present in *Fasciola*. Using degenerate primers based on the conserved regions of cysteine proteinases , seven different clones were obtained by PCR from adult *F.hepatica* cDNA (Heussler and Dobbelaere, 1994). Five of these genes showed varying levels of homology to vertebrate cathepsin L, while the remaining two were more similar to vertebrate cathepsin B. Northern blot analysis indicated several fold differences in the level of expression. Three of the cathepsin L-like enzymes were highly expressed, and one of the cathepsin L -like and both cathepsin B-like enzymes were expressed at low levels.

Southern blot hybridization indicated that while the poorly expressed and more divergent cathepsin L-like (Fcp4 E) and the cathepsin B genes appear to be single copy, the remaining cathepsin L-like genes could be in multiple copies according to a complex pattern of hybridization, although cross-hybridization was not ruled out. Antibodies raised against a full length cDNA clone (Fcp1C) expressed in *E.coli*, detected a single band of 30 kDa and a precursor band of 38 kDa in worm extracts and also in excretion/secretion products, indicating that this protein is secreted (Heussler and Dobbelaere, 1994). By a similar PCR approach a cDNA clone was reported from japanese flukes *Fasciola sp.* with similarities to vertebrate cathepsin L, H and S. Although other clones slightly divergent from the reported one were detected in this study they were not further analyzed (Yamasaki and Aoki, 1993).

Further evidence for the existence of multiple cysteine proteinases similar to cathepsins came from gel filtration purification and two dimensional SDS-PAGE analysis of excretion/secretion products of *F. hepatica*. A major activity was detected at 28 kDa, that was separated in seven bands in two dimensional SDS-PAGE, indicating the presence of isozymes or different enzymes of the same size (Wijffels *et al.*, 1994). In addition, different N-terminal sequences were obtained from the purified cysteine proteinases. It was also established that proline residues in these proteinases were modified to 3-hydroxy-prolines derivatives, a modification that occurs in several *F. hepatica* proteins (Bozas and Spithill, 1996).

A gene was purified from an adult cDNA library by screening with ovine anti-cysteine proteinase serum, and showed similarities to cathepsins L and H, however it was not established if this gene correspond to one of the secreted enzymes (Wijffels *et al.*, 1994). A different gene was obtained from a cDNA library, and localized to the gut and the Mehlis gland (Panaccio *et al.*, 1994). Using anti-cathepsin L1 serum an adult cDNA library was screened in our laboratory and a clone was purified, corresponding to a cysteine proteinase with homology to one partial clone obtained by Heussler and Dobbelaere (1994), and the clones obtained by Yamasaki and Aoki (1993) and Wijffels *et al.*(1994). This clone was functionally expressed in *Saccharomyces cerevisiae,* and the resulting enzyme showed physicochemical properties indistinguishable from the cathepsin L1 purified from the excretion/secretion of flukes by Smith *et al.*(1993). Immunoblot analysis confirmed that the cloned gene encoded the secreted cathepsin L1 (Roche *et al.*, 1997).

Cysteine proteinases with similarities to vertebrate cathepsins where detected in several stages of the life cycle of *F. hepatica*. Dalton and Heffernan (1989) detected cysteine proteinase activities in metacercariae and newly excysted juveniles (NEJ) by gelatin substrate gels, and this results were confirmed by Carmona *et al.* (1993). Heussler and Dobbelaere (1994) demonstrated by immunoblotting the presence of different cysteine proteinases in redia, cercaria, metacercariae, NEJ, immature and adult flukes. By N-terminal sequencing of the major proteins produced by the NEJ, a band of 31 kDa with sequence homology to cathepsin B, and a 34 kDa band with homology to cathepsin B precursor were detected along with a band with homology to cathepsin L, but distinct from the sequences obtained from adult cDNA (Tkalcevic *et al.*, 1995). More recently a neutral cathepsin B activity was detected in the excretion/secretion products of newly excysted juveniles, corresponding to a 29 kDa protein. A cDNA clone was obtained from this stage, corresponding to a cathepsin B-like enzyme. Using anti-bovine cathepsin B serum, the NEJ cathepsin B was localized to the gut (Wilson *et al.*, 1997).

Because the cathepsin L-like proteinases are involved in vital functions of tissue penetration, nutrition and immune evasion, and at the same time constitute a major component of the secretion products, they are considered good targets of immunoprophylaxis. The cysteine proteinases of adult fluke were used in vaccination assays with different results. Wijffels *et al.* (1994) reported that no protection was obtained in sheep vaccinated with purified cathepsin L-like enzymes, but a significant reduction in fecal egg count was achieved. A mixture of cathepsin L1 or cathepsin L2 and *F. hepatica* hemoglobin induced high levels of protection in cattle, and induced a very significant reduction in the viability of the eggs produced by the surviving flukes (Dalton *et al.*, 1996). The consistent effect on egg count or egg viability could be related to the presence of cathepsin L-like enzymes in the reproductive tract of the fluke, a localization also detected in other organisms (see above).

In this study we have isolated a cDNA clone by screening an adult cDNA library with anti-cathepsin L2 serum. The resulting clone was analyzed, and it relationships with the gene sequences of other cysteine proteinases from *F. hepatica* established. Functional expression of the cathepsin L2 is described, resulting in a proteinase with similar physicochemical properties to the native cathepsin L2 secreted by adult liver fluke. In addition, in an attempt to understand the particularities of the differences in substrate specificity between the two secreted cathepsin L-like enzymes from *F. hepatica,* mutagenesis of cathepsin L1 is described. Finally, an evolutionary analysis of the cysteine proteinases of the papain superfamily is presented, showing the relationships of the different gene families detected in several parasitic species and also the relationships between the diverse cysteine proteinases present in vertebrate hosts.

2. Materials And Methods

2.1 Materials

Bachem

Synthetic peptide substrates

Cambridge Research Biochemicals (Chesire, UK)

Synthetic peptide substrates , Peptide synthesis

Department of Biological Sciences, University of Durham, U.K.

DNA sequencing (Applied Biosystems)

Department of Genetics, Trinity College Dublin

DNA sequencing (Applied Biosystems)

Gibco, Life technologies Ltd.

RPMI-1640 (10X) w/o L-glutamine

Kodak

Polaroid film

Pharmacia LKB Biotechnology

Sephacryl S200 HR, DEAE Sepharose

Promega

Agarose, Anti-rabbit IgG (Fc) alkaline phosphatase conjugate.

Deoxytrinucleotides, DNA molecular weight markers,

isopropylthio-ß-D-galactoside (IPTG),

1,5-bromo-4-chloro-3-indolyl-ß-D-galactosidase (X-Gal)

pGem®T vector system, Restriction enzymes, T4 DNA ligase, Taq DNA Polymerase,

Wizard™ DNA clean-up system, Wizard™ lambda preps.

Schleicher and Schuell

Nitrocellulose

Sigma Chemical Company

5-bromo-5-chloro-3-indolyl phosphate (BCIP), Adenine diphosphate (ADP),

Anti-bovine IgG conjugated to alkaline phosphatase (rabbit),

Anti-bovine IgG conjugated to alkaline phosphatase (rabbit),

Anti-rat IgG conjugated to alkaline phosphatase (rabbit),

bovine serum albumin (BSA), coomassie brilliant blue R, dithiothreitol (DTT),

diethylpyrocarbonate (DEPC), Freund's complete and incomplete adjuvants, gelatin, gentamycin, glutamine, iodoacetamide, leupeptin, lysozyme,

L-trans-epoxysuccinyl-leucylamido-[4-guanidino]-butane (E-64),

nitro-blue tetrazolium (NBT), prestained molecular weight markers, proteinase K,

p-nitrophenyl phosphate, phenylmethylsulphonyl fluoride (PMSF),

salmon sperm DNA, sodium dodecyl sulphate (SDS).,Tween 20.

The Oswel DNA Service (Edinburgh University, UK)

Synthetic oligonucleotides.

2.1 Methods

2.2.1 Preparation of Excretion/Secretion products (E/S) from mature liver fluke

Mature *F. hepatica* parasites were removed from the bile ducts of infected bovine livers obtained at a local abattoir. After several washes in phosphate buffered saline pH 7.3 (PBS), the parasites were incubated at 37°C for 18 h in RPMI-1640, pH 7.3, containing 2% glucose, 30 mM HEPES and 25 mg/ml gentamycin. The culture medium was removed, centrifuged at 12,000 g for 30 min and the supernatant collected and stored at -20°C (Dalton and Heffernan, 1989). The preparation was termed excretory /secretory (E/S) products.

2.2.2 cDNA library construction

A *F. hepatica* λgt11 cDNA library was prepared in our laboratory (Smith , 1994). In brief, total RNA was isolated from mature adult flukes according to Chomczynski and Sacchi (1987) protocol . After mRNA isolation by oligo dT column binding, double stranded cDNA was generated using the Promega Riboclone cDNA synthesis kit. *EcoR* I linkers were added to the cDNA, which was then ligated to λgt11 arms and packaged into lambda heads using the Packagene system. The packaged phage was titred on LE392 *E.coli* cells and then amplified on Y1090 *E.coli* cells.

2.2.3 Immunological screening of an adult *Fasciola hepatica* cDNA library

The adult fluke cDNA library was plated on Y1090 *E.coli* cells, and duplicate nitrocellulose filters were lifted from the culturing petri dishes. After blocking with 1% bovine serum albumin in PBS 0.1% Tween, the filters were incubated first with anti-cathepsin L1 or anti-cathepsin L2, and then with alkaline phosphatase - conjugated anti-rabbit rat IgG. Antibody binding was visualised using NBT and BCIP as substrates for alkaline phosphatase in AP buffer (100 mM Tris, pH 9.5, containing 100 mM NaCl, 5 mM $MgCl_2$) and the positive phage lysis plaques were identified. Those phage lysis plaques that showed reaction with one or both of the sera, were collected as agar plugs in SM buffer and the phages were recovered with 5% chloroform in SM buffer. The screening procedure, which was repeated twice, yielded 22 purified clones, 13 of which were recognized by the anti-cathepsin L2 serum, and the remaining recognized by both anti-cathepsin L1 and anti-cathepsin L2 sera.

2.2.4 Phage DNA isolation

DNA of the selected phage clones were purified according to Sambrook *et al.* (1989) or with the Promega Wizard Lambda prep DNA purification kit. Briefly, phages were adsorbed to *E.coli* Y1090 cells and cultured in liquid media until bacterial cell lysis was achieved. The phages were recovered with SM buffer 0.5 % chloroform, centrifuged, and the phage lysate was treated with a nuclease mix (RNAse A and DNAse I to a final concentration between 0.25 to 1 mg/ml). The phages were precipitated with 0.4 volumes of 33% Polyethylene glycol (PEG) 3.3 M NaCl for 30 min to 3 h, and centrifuged at 10000g. The pellets were resuspended in phage buffer (40 mM Tris-HCl, pH 7.4, containing 150 mM NaCl , 10 mM $MgSO_4$) and cleared by centrifugation. The purified phages were mixed with Promega Wizard Purification Resin, passed through a column, washed and eluted according to the manufacturers manual. Alternatively, the purified phages were treated with 50 mg/ml Proteinase K, 0.5% SDS, and extracted once with phenol, twice with chloroform. The DNA was precipitated in ethanol and then resuspended in sterile ultrapure water according to the procedure described by Sambrook *et al.* (1989).

2.2.5 Isolation of clones by Polymerase Chain Reaction (PCR)

DNA of the immunoreactive phage clones was analysed by Polymerase Chain Reaction (PCR). PCR amplification of anti-cathepsin L2 positive clones was performed with universal λgt11(Promega Lambda gt 11 Forward and Reverse) primers and with specific primers designed using the consensus sequences around the active site cysteine (JDF), and asparagine (MCB) of papain-like cysteine proteinases. The primers contained internal restriction enzymes sites as underlined.

JDF ACA GAA TTC GGY TAT GTG ACT GGY GTG AAG G
Eco R I G Y V T G V K

MCB TTA AAG CTT CCA $^{IGA}/_{RCT}$ RTT YTT IAC RAT CCA RTA
Hin d III W S N K V I W Y

Cycling conditions were : 4 min at 94°C initial denaturing step, 35 cycles of denaturation at 94°C for 1 min, primer annealing at 55°C for 1 min and extension at 74°C for 1 min, and a final extension step of 3 min at 74°C.

2.2.6 Subcloning of PCR gene fragments

The PCR amplified products were electrophoretically separated in 1 or 1.5 % agarose-Tris-acetate gels. The gels were stained with ethidium bromide, and the bands of the expected sizes (approx. 1 kb) were cut from the gel and purified by spinning through siliconized glassbeads. The DNA was precipitated with ethanol, and resuspended in sterile distilled water. The purified DNA fragments were ligated to pGEMT vector DNA (Promega) according to the protocols provided by the manufacturer.

2.2.7 Transformation and culturing of *E. coli*

Escherichia coli strains JM109 and DH5α were used as the host for pGEM-T plasmid propagation and manipulations. Strains LE392 and Y1090 were used for phage propagation and manipulations. Fresh competent cells were prepared using the calcium chloride method and transformation was performed according to standard protocols (Sambrook *et al.* 1989). *E. coli* cells were normally cultured on Luria Bertani (LB) or SOC media. One hundred mg/ml ampicillin was added to plates and liquid media of cultures of cells harbouring plasmids. LB plates containing 100 mg/ml ampicillin, 0.5 mM IPTG and 40 mg/ml X-Gal were used for blue/white selection of recombinant plasmids derived from pGEM-T.

2.2.8 Screening of recombinant colonies.

Positive or recombinant colonies were picked from the plates and replicated in fresh media. Plasmid DNA was prepared using the alkaline lysis miniprep method (Sambrook *et al.* 1989) or using the Wizard miniprep DNA purification system (Promega) according to the suppliers protocols. DNA was analyzed by restriction enzyme digestions and visualized by electrophoresis on agarose-TAE gels.

2.2.9 DNA Sequencing

The complete DNA sequence of selected clones (pFheTCL2, pFheHCL2, pTYH3) was determined by an automated method (Applied Biosystems) at the Department of Biological Sciences, University of Durham, U.K., or at the Department of Genetics, Trinity College, Dublin.

2.2.10 Subcloning into yeast expression plasmid

Primers based on the sequences crossing the start and stop codons of cathepsins L1 and L2 were designed (URUF and IREB respectively). These primers include a *Hin* d III site overhang (underlined). PCR of the anti-cathepsin L2 positive clones with URUF and IREB under the cycling conditions previously described (section 2.2.5) generated, therefore, the complete coding sequence of cathepsin L2.

URUF: AAC AAT AAG CTT ATG CGR TTM TTC RTA TTA GCC GTC
Hin d III M R L F $^{V}/_{I}$ L A V

IREB: TGAC AGA AGC TTA TCA CGG AAA TCG TGC CAC
Hin d III Z P F R A V

The amplified fragments were gel purified, cloned into pGEMT (Promega) as described previously and sequenced. The insert was excised with *Hin* d III and ligated to *Hin* d III linearized pAAH5 yeast expression vector (Ammerer, 1983). The orientation of the insert in pAAH5 vector was assessed by restriction mapping and PCR with the primers used for cloning and a primer (termed PBR) synthesized using the pBR322 sequences of pAAH5 (50 bp from the *Bam H* I site near the ADC1 terminator).

PBR: GTG ATG TCG GCG ATA TAG

E. coli strains MC1061 and JM109 were used for propagating the pAAH5 plasmids. A diagram of the cloning strategy is depicted in figure 2.1.

2.2.11 Construction of *Fasciola hepatica* cathepsin L1 Leu-70-Tyr mutant

A comparison of the sequences of the recombinant liver fluke cathepsin L1 (Roche *et al.* 1997) and L2 show several differences. Two residues in the S_2 subsite of the active site differed between the enzymes. In cathepsin L1 the residues at position 70 and 161 (67 and 157 in papain numbering) are leucine and valine respectively, while tyrosine and leucine are present in the corresponding positions in cathepsin L2. A mutation of one of those residues was performed, changing the aliphatic non-polar Leu^{70} present in cathepsin L1 to the aromatic polar Tyr^{70} in the cathepsin L2.

Two overlapping oligonucleotide primers with opposite orientations covering the residue of interest were designed (substituted residues according to *Fasciola hepatica* cathepsin L1 are shown in bold) :

FOTYR GGT GGA T**AT** ATG GAA AAT GC**A** TAC CAA T
G G Y M E N A Y Q
Nsi I

BATYR ATT TTC CAT **ATA** TCC ACC ACC GCA
N E M Y G G G C

A *Nsi* I restriction site was introduced in the forward primer (underlined) by means of a synonymous substitution (Ala GCT to Ala GCA) in the cathepsin L1 sequence. This restriction site was used to assess the introduction of the mutation. The primers overlap in 18 nucleotides. DNA of the pGEMT derivative plasmid pFheT1.9 that contains the entire coding region of *Fasciola hepatica* cathepsin L1 (Roche *et al.*, 1997) was used as template in PCR amplifications. Two amplification reactions were performed using as primers FOTYR and the SP6 promoter primer (Promega), and BATYR and T7 promoter primer (Promega). The cycling conditions were: 4 min at 94°C initial denaturing step, 20 cycles of denaturation at 94°C for 45 secs, primer annealing at 52°C for 45 secs and extension at 72°C for 1 min, and a final extension step of 3 min at 72°C. The products were separated on 2% agarose-TAE electrophoresis, the bands of the expected size (approx. 1 kb) were cut from the gel and purified by spinning through siliconized glassbeads, precipitated and resuspended in sterile distilled water. A second PCR reaction was performed using both products as templates, and SP6 and T7 promoter primers under the same conditions as above. Aliquots of the secondary PCR reaction were digested with *Nsi* I to verify the incorporation of the mutation. The rest of the reaction, was electrophoresed and the fragments of the expected size were purified and subcloned in pGEMT as previously described. The sequence of the resulting clone (pTYH3) was then determined (see section 2.2.9). Digestion of pTYH3 DNA with *Hin* d III produced a whole length mutant insert that was purified and ligated to *Hin* d III linearized pAAH5 yeast expression vector as described before (see section 2.2.10).

2.2.12 Transformation and culturing of *Saccharomyces cerevisiae*

Saccharomyces cerevisiae strain DBY746 (Mat a his3-D1-leu2-3 leu2-112 ura3-52 trp1-289a) (Yeast Genetic Stock Center, Department of Biophysics and Medical Physics, University of California, Berkeley, CA, USA) was transformed with the plasmid pFheCL42 by the lithium acetate method (Carter *et al.*, 1987). Yeast transformants were cultured in selective minimal media (Bacto Yeast Nitrogen Base 6.7 g/l,, D-glucose 20 g/l, uracil 20 mg/ml in 0.1M phosphate buffer pH 6.5) at 30°C for 3 or 4 days. DBY746 strain was routinely maintained in buffered complex media (YEPD): Yeast extract 10 g/l, peptone 20 g/l, D-glucose 20 g/l in 0.1M phosphate buffer pH 6.5.

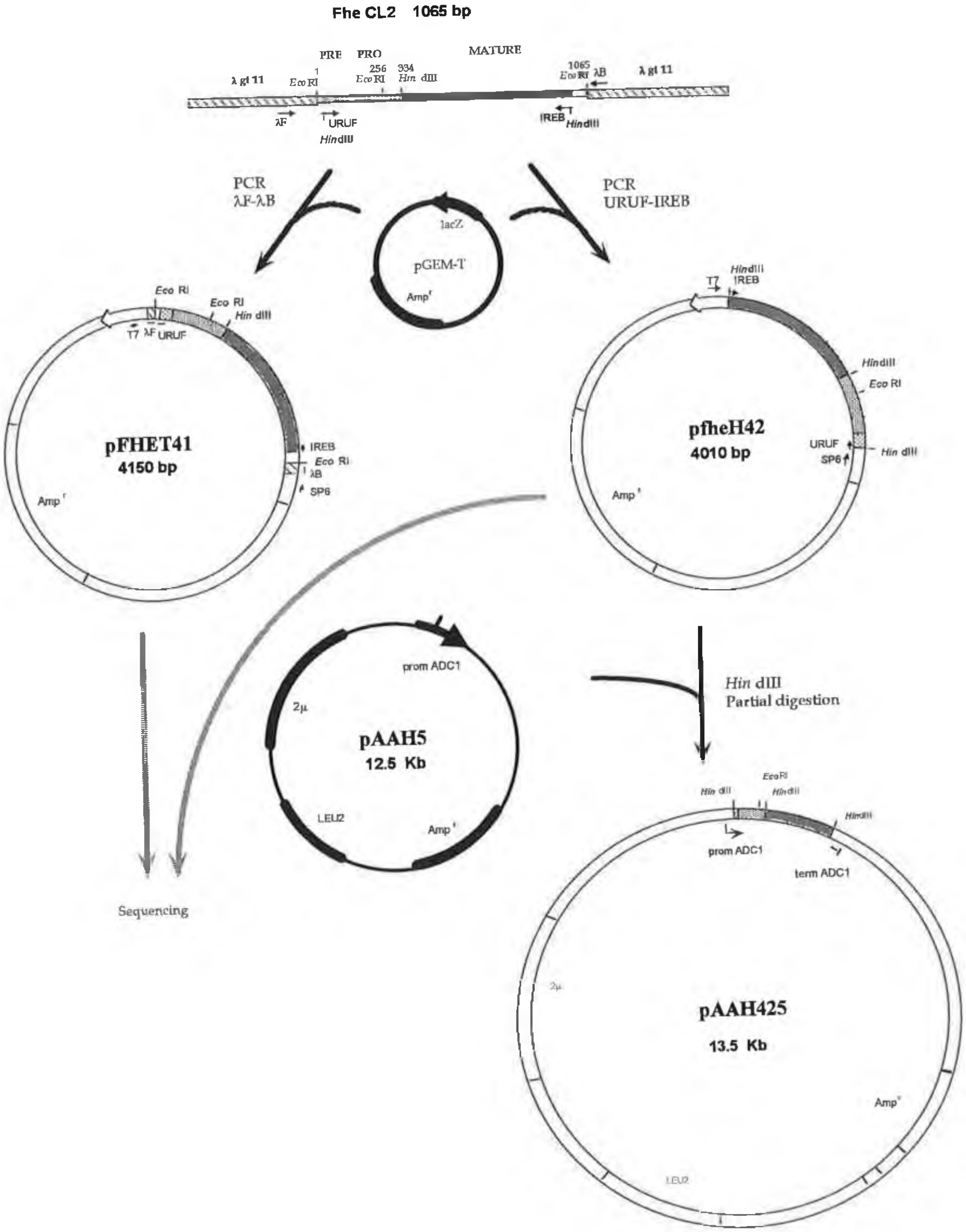


Fig. 2.1

Diagrammatic representation of the cloning strategy.

2.2.13 Preparation of yeast cell extracts and culture supernatants

Yeast cells were grown in selective minimum media at 30°C in flasks with vigorous agitation and the OD_{600} was monitored. Cultures were harvested when they reached the early stationary phase. The cells were collected by centrifugation, and the supernatant stored. The culture supernatants (CS) were concentrated tenfold in an Amicon 8400 concentrator using an Amicon YM3 membrane (3000 Da molecular mass cut-off, Amicon, WI) and stored at -20°C. The cells pellets were then washed in distilled water and resuspended in one-tenth of the culture volume in 50 mM phosphate buffer pH 6.5, 10 mM β-mercaptoethanol, 100 mg/ml Lyticase. The cells were incubated at 30°C with agitation for 1 h, and then one-tenth volume of lysis buffer (50 mM phosphate buffer pH 6.5, 10 mM b-mercaptoethanol, 10 mM EDTA, 0.1% Triton X-100, 0.1% Sarcosyl) was added. Cellular debris were removed by centrifugation at 13,000g for 10 min and the supernatant representing the cell extract (CE) were collected and stored. The integrity of yeast cells in cultures was assessed microscopically after mixing samples with 0.1% crystal violet.

2.2.14 Purification of yeast-expressed enzyme

Thirty litres of yeast culture media supernatant was concentrated to 200 ml in an Amicon 2000A concentrator using an Amicon YM3 membrane (3000 Da molecular mass cut-off, Amicon, WI, USA) and subsequently concentrated to 20 ml in an Amicon 8400 concentrator using an Amicon YM3 membrane. The concentrate was applied to a Sephacryl S200HR gel filtration column (2.5 cm x114 cm) equilibrated in 0.1M Tris-HCl, pH 7.0, at 4°C (Buffer A). The column was eluted with Buffer A, and 5 ml fractions were collected. Each fraction was assayed for cathepsin L activity using the fluorogenic substrate Z-Phe-Arg-NHMec at a final concentration of 10 mM in Buffer A (see section 2.2.19). The Sephacryl S200HR fractions containing Z-Phe-Arg-NHMec cleaving activity were pooled and concentrated to 10 ml on an Amicon 8400 concentrator with a YM3 membrane.

2.2.15 SDS-Polyacrylamide gel electrophoresis (SDS-PAGE)

Protein samples were analysed by one dimensional 10% or 12% SDS-PAGE gels according to the method of Laemmli (1970). Samples were prepared in non-reducing sample buffer (0.12 M Tris-HCl, pH 6.8, 5% (w/v) SDS, 10% (w/v) glycerol and 0.01% (w/v) Bromophenol Blue) or reducing sample buffer (5% 2-mercaptoethanol in non-reducing sample buffer and boiled for 5 min). Gels were stained for protein with Coomassie brilliant blue R or silver salts according to Sambrook *et al.* (1989).

2.2.16 Immunoblotting

Adult *F. hepatica* E/S products, yeast CS, yeast CE, and the purified native cathepsin L enzymes were separated by reducing SDS-PAGE and electrophoretically transferred to nitrocellulose paper using a semi-dry electroblotting system. Following blocking of non-specific binding in 1% bovine serum albumin in TBST (10 mM Tris, pH 8, containing 150 mM NaCl, 0.1% Tween), the nitrocellulose membrane was incubated in anti-cathepsin L1, anti-cathepsin L2 or non-immune rabbit serum. Alkaline phosphatase -conjugated anti-rabbit rat IgG was used to detect the bound immunoglobulin using Nitro-blue tetrazolium (NBT) and 5-bromo-4-chloro-indolyl phosphate (BCIP) were used as substrates for the alkaline phosphatase.

2.2.17 Protein concentration estimation.

Protein concentration was measured using a BCA protein assay kit in microtitre plates according to the method of Redinbaugh and Turley (1986). Bovine serum albumin was used as a protein standard.

2.2.18 Characterisation of enzyme activity

Proteinase activity was measured fluorometrically using peptide-NHMec as substrates. Each substrate was stored as a 1 mg /100 ml stock solution in dimethyl-formamide. The routine assays were carried out using a final concentration of 10 mM substrate in 0.1 M sodium phosphate buffer, pH 6.5, containing 0.5 mM dithiothreitol, in a volume of 1 ml. The mixtures were incubated at 37°C for 30 min or 1 h before stopping the reaction by the addition of 0.2 ml 1.7 M acetic acid. The amount of 7-amino-4-methylcoumarin (NHMec) released was measured using a Perkin-Elmer fluorescence spectrophotometer with excitation set at 370 nm and emission at 440 nm. One unit enzyme activity was defined as the amount which catalysed the release of 1 mmol NHMec/min at 37°C. A standard curve of different concentrations of NHMec was made in parallel with each experiment.

The substrate specificity of cathepsins L1 and L2 purified from *F. hepatica* E/S products and yeast recombinant cathepsin L2 were determined with the fluorogenic peptide substrates Z-Arg-Arg-NHMec, Z-Phe-Arg-NHMec, Bz-Phe-Val-Arg-NHMec Tos-Gly-Pro-Arg-NHMec, Boc-Val-Pro-Arg-NHMec, and Bz-Val-Leu-Lys-NHMec. Relative activity against these substrates was determined in parallel activity assays using the same sample of enzyme.

The kinetic constants of the purified yeast expressed and native *F. hepatica* cathepsin L2 proteinases were determined with the substrates: Z-Arg-Arg-NHMec, Z-Phe-Arg-NHMec and Tos-Gly-Pro-Arg-NHMec. The kinetic constants, k_{cat} and K_m, were obtained by non-linear regression analysis by using the program Enzfitter (Leatherbarrow, 1987).

2.2.19 Active site titration of cathepsins L1 and L2

Active site titration using the cysteine proteinase inhibitor L-*trans*-epoxysuccinyl-leucylamido-(4-guanido)-butane (E-64) was performed according to the method of Barrett *et al.* (1982) using the fluorogenic substrate Z-Phe-Arg-NHMec.

2.2.20 Database search and retrieval of sequences

A database search of cysteine proteases of the papain superfamily was performed using World Wide Web browsing and retrieving facilities. The databases searched were translated GenBank, PIR, TREMBL, SWISSPROT and DDBJ, and the browsers used were Entrez, Biologists Search Palette, and Sequence Retrieval

System. The entries obtained from different databases were cross-chequed and all redundant entries were eliminated. Sequences were named following the SWISSPROT criteria, but swapping the five characters of the species names to the beginning of the entry name.

2.2.21 Alignment of sequences

Sequences were aligned by progressive alignment using the ClustalW program. The alignments were performed on full length coding sequences (pro-enzymes), mature regions, proregions, or conserved regions around the active site residues. Mature and pro-regions of the proteinases were separated according to the features tables of each entry or by comparison after primary alignment. A primary alignment of papain (papa_carpa), caricain (pap3_carpa), glycil-endopeptidase (pap4_carpa), actinidin (actn_actch) human cathepsin L (catl_human), human cathepsin B (catb_human) and rat cathepsin B (catb_rat) was performed under the program defaults values for gap penalties. This alignment was manually corrected based on the crystallographic data of these entries (papain pdb entry 1PE6, caricain pdb entry 1PPO, glycil-endopeptidase pdb entry 1GEP, actinidin pdb entry 1AEC, human cathepsin B, pdb entry 1HUC, rat cathepsin B pdb entry 1CTE, and human cathepsin L data from Coulombe *et al.*, 1996), the corresponding entries in the Homology derived Secondary Structure of Proteins database (HSSP), and the structure optimised alignment of rat cathepsin B and papain presented in Musil *et al* (1991). A secondary structure mask was constructed based on this optimized alignment in order to restrain the insertions and deletions to non-structured regions. For the pro-regions the sequences used were rat procathepsin B (pdb entry 1MIR), and the secondary structures tables of human procathepsin L (Coulombe *et al.* 1996) and procaricain (Groves *et al.*, 1996).

All the alignments were performed with and without using the corresponding secondary structure masks (i.e. proenzyme, mature enzyme or proregion) using the following parameters: protein weight matrix : Blosum series, Helix gap penalty : 4, strand gap penalty : 4, loop gap penalty : 1, secondary structure terminal penalty : 2. Sequence alignments were visualized with GENEDOC (Nicholas *et al.*, 1997).

2.2.22 Phylogenetic inferences

Primary phylogenetic inferences were made using the CLUSTALW program. CLUSTALW calculates Phylogentic trees using the Neighbor joining method of Saitou and Nei (1987) based on a matrix of distances between sequences. The distances were normalized using the Kimura formula for correcting the distance data for multiple substitutions (Kimura, 1983). The unrooted neighbor joining trees generated by ClustalW were visualized using TreeView (Page,. 1996) or DRAWTREE of the PHYLIP package (J.Felsenstein, 1989). The effects of different sequences in the topology of the resulting trees were analyzed by adding sequences or profiles of aligned sequences to previously aligned sets. Subsets of sequences were also analyzed using the programs SEQBOOT, PROTDIST and PROTPARS of the PHYLIP package (J.Felsenstein, 1989).

2.2.23 Protein modeling

Tridimensional models of cathepsin L1 and cathepsin L2 were made through the Swiss-Model server (http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html) . The sequences were first aligned to the available structures in the Protein Data Bank. The highest scores were with caricain (1PPO), and actinidin (2ACT, 1AEC), so these structures were used as frames for the model building.

2.2.24 Peptide design

The possible region of the propeptide of the *F. hepatica* cathepsins in contact with the active site was predicted, based on the comparison with rat procathepsin B (Cygler *et al*, 1996; Chen *et al.*, 1996; Chagas *et al.*, 1996), human procathepsin B (Turk *et al.*, 1996), human procathepsin L (Coulombe *et al.*, 1996), and procaricain (Groves *et al.*, 1996). For that purpose, these sequences and pro-regions of other mammalian cathepsins were aligned. Based on this optimal alignment, two overlapping peptides were designed, covering the region of possible interaction of the propeptide with the enzymes's active site, for analysis of their inhibitory properties. The peptides were synthesized by Zeneca (Cambridge Research Chemicals, Cheshire, UK).

3 Cloning of *Fasciola hepatica* cathepsin L2

3.1. Isolation of a cDNA clone encoding a cathepsin L proteinase

A λgt11 cDNA library of the adult liver fluke *Fasciola hepatica* (5 x 10^4 lysis plaques) was immunologically screened in order to isolate clones that express cathepsin-L like proteases. The rabbit antibodies employed were raised against two different excretion-secretion products of the parasite that show cathepsin L proteinase activity (cathepsin L1 and cathepsin L2), that were previously isolated in our laboratory (Smith *et al.*, 1993b, Dowd *et al.* 1994). After three rounds of immunological screening twenty-two clones were purified and assigned to two categories according to their reactivity with the sera. Although cathepsin L1 serum recognized only nine of these clones, these and thirteen other clones were detected with anti-cathepsin L2. Six of these clones were selected for further analysis, three from each reactivity group.

As a first approach to assess the size of the inserts, PCR with DNA of the selected clones using universal lambda phage primers was performed, and the resulting DNA fragments separated on agarose gel electrophoresis. Five of the clones contained inserts of approximately 1 kb and the remaining clone contained an insert of 1400 bp. PCR reactions using primers designed against conserved sequences around the active site residues in cysteine proteinases (JDF, MCB), showed amplification bands of the expected size in all the clones selected. These results support the hypothesis that the clones were the liver fluke cathepsins (Figure 3.1). On this basis one of the 1 kb clones recognized only by anti-cathepsin L2 serum was chosen for further investigation and named FheCL2 (the characterization and expression of this clone is the subject of this study). Another clone that was recognized by both anti-cathepsin L1 and anti-cathepsin L2 sera was named FheCL1, and was characterized in collaboration with others in the laboratory (Roche *et al.* 1997).

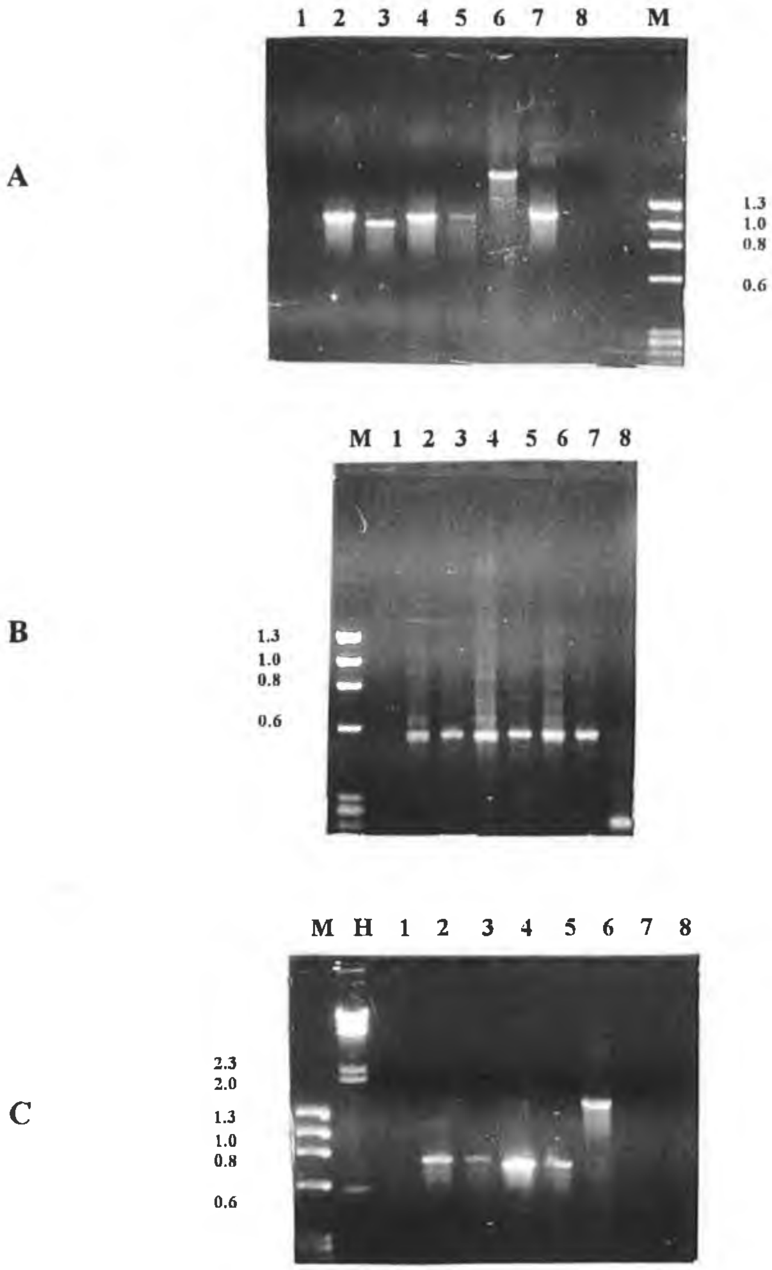


Fig.3.1

PCR amplifiction of six selected lambda clones.

(a) Amplification with lambda forward and lambda reverse primers (b) Amplification with JDF and MCB cathepsin primers (c) Amplification with JDF and lambda reverse primers. Lane 1, control with no DNA; lanes 2-7, clones 1 to 6; lane 8 λgt11 DNA control. The DNA marker (M) is Φ X 174 DNA digested with *Hae* III. The clones 1 , 4 , 5, and 6 were further analyzed corresponding to pFheCL1, pFheCL2, pFheCL1-5 and pFheCL2-6. Lane H DNA marker (λ *Hind* III)

3.2. Determination of the nucleotide sequence of FheTCL2

The presence of an internal *Eco* RI restriction site in the selected clone, impaired the purification of the whole insert by direct restriction enzyme digest, as the amount of the correct product from partial digestions was almost neglectable. A different approach had to be taken. A new PCR amplification of FheCL2 clone with λ primers was performed and the product was gel purified and subcloned into pGEMT vector. The complete nucleotide sequence of both strands of the derived plasmid pFheTCL2 was determined (Genebank, Accession Number U62289). Two other of the six clones that were originally selected were also subcloned as described, and partial sequences for them were obtained. These clones were termed pFheCL1-5, and pFheCL2-6, because they were related to the FheCL1 and FheCL2 clones respectively.

The nucleotide and deduced amino acid sequence of the cathepsin L2 is shown in Fig. 3.2. The sequence obtained for the FheTCL2 clone was 1065 nucleotides long and corresponds to the complete coding region of the pre-procathepsin L2. It comprises 21 nucleotides of untranslated 5' end, an open reading frame of 978 nucleotides coding for the 326 amino acids of the pre-procathepsin L2 and 66 nucleotides of the untranslated 3'end. A typical AATAA polyadenylation signal is present at position 1034-1039 in the 3' untranslated region.

A stretch of 15 hydrophobic amino acids can be found at the N-terminus of the predicted polypeptide. According to the physical characteristics of eukaryotic signal sequences (von Heijne, 1985), this region could be considered as the predicted signal peptide. N-terminal sequencing of *Fasciola hepatica* cathepsins L1 and L2 purified from excretion/secretion products revealed that these enzymes differed from other cathepsins L in having an additional alanine on the N-terminus (Smith *et al.*, 1993, Dowd *et al.*, 1994). This alanine was present in the deduced amino acid sequence of FheTCL2 at position 106, and allowed the exact positioning of the cleavage point between the pro-peptide and mature protein. Based on this data the deduced protein sequence could be divided into a 15 amino acid signal sequence, a 91 amino acid propeptide and a 220 amino acid mature protein.

```
                                                                                        ↓
   1 GAATTCCGTAACAATCAAACG ATG CGG TGC TTC GTA TTA GCC GTC CTC ACG GTC GGA GTG TAC GCC TCG AAT GAC      75
   1                        M   R   C   F   V   L   A   V   L   T   V   G   V   Y   A   S   N   D       18

  76 GAT TTG TGG CAT CAA TGG AAA CGA ATA TAC AAT AAA GAA TAT AAT GGG GCT GAC GAT GAG CAC AGA CGA AAT   147
  19  D   L   W   H   Q   W   K   R   I   Y   N   K   E   Y   N   G   A   D   D   E   H   R   R   N     42

 148 ATT TGG GGG AAA AAT GTG AAA CAT ATC CAA GAA CAC AAC CTA CGT CAC GAT CTC GGC CTC GTC ACC TAC AAG   219
  43  I   W   G   K   N   V   K   H   I   Q   E   H   N   L   R   H   D   L   G   L   V   T   Y   K     66

 220 TTG GGA TTG AAC CAA TTC ACT GAT TTG ACA TTC GAG GAA TTC AAG GCC AAA TAT CTA ATA GAA ATC CCA CGC   291
  67  L   G   L   N   Q   F   T   D   L   T   F   E   E   F   K   A   K   Y   L   I   E   I   P   R     90
                                                                          ▽
 292 TCG TCT GAG TTA CTC TCA CGC GGT ATC CCG TTT AAG GCG AAC AAG CTT GCC GTA CCC GAG AGC ATT GAC TGG   363
  91  S   S   E   L   L   S   R   G   I   P   F   K   A   N   K   L   A   V   P   E   S   I   D   W    114

 364 CGT GAC TAT TAT TAT GTG ACT GAG GTG AAA AAT CAG GGA CAA TGT GGT TCC TGT TGG GCT TTC TCA ACA ACC   435
 115  R   D   Y   Y   Y   V   T   E   V   K   N   Q   G   Q   C   G   S   C   W   A   F   S   T   T    138

 436 GGT GCT GTG GAG GGA CAG TTT AGG AAG AAC GAA AGA GCT AGT GCT TCA TTC TCT GAG CAA CAA CTG GTC GAT   507
 139  G   A   V   E   G   Q   F   R   K   N   E   R   A   S   A   S   F   S   E   Q   Q   L   V   D    162

 508 TGT CCC CGT GAT TTG GGC AAT TAT GGT TGC GGT GGA GGA TAT ATG GAA AAC GCT TAT GAA TAT TTG AAA CAC   579
 163  C   P   R   D   L   G   N   Y   G   C   G   G   G   Y   M   E   N   A   Y   E   Y   L   K   H    186

 580 AAC GGA TTG GAA ACT GAG TCC TAT TAT CCA TAC CAG GCT GTG GAA GGT CCG TGT CAA TAC GAT GGG CGG TTG   651
 187  N   G   L   E   T   E   S   Y   Y   P   Y   Q   A   V   E   G   P   C   Q   Y   D   G   R   L    210

 652 GCA TAT GCC AAA GTG ACT GGC TAC TAT ACT GTG CAT TCT GGC GAT GAG ATA GAA TTA AAG AAT TTG GTC GGT   723
 211  A   Y   A   K   V   T   G   Y   Y   T   V   H   S   G   D   E   I   E   L   K   N   L   V   G    234

 724 ACC GAA GGA CCT GCG GCG GTC GCT TTG GAT GCG GAT TCT GAC TTC ATG ATG TAC CAG AGT GGT ATT TAT CAG   795
 235  T   E   G   P   A   A   V   A   L   D   A   D   S   D   F   M   M   Y   Q   S   G   I   Y   Q    258

 796 AGC CAA ACT TGT TTA CCG GAT CGC TTG ACT CAT GCA GTC TTG GCT GTC GGT TAT GGA TCA CAA GAT GGT ACT   867
 259  S   Q   T   C   L   P   D   R   L   T   H   A   V   L   A   V   G   Y   G   S   Q   D   G   T    282

 868 GAC TAT TGG ATT GTG AAA AAT AGT TGG GGA ACG TGG TGG GGT GAG GAC GGT TAC ATT CGG TTT GCC AGG AAC   939
 283  D   Y   W   I   V   K   N   S   W   G   T   W   W   G   E   D   G   Y   I   R   F   A   R   N    306

 940 CGA GGT AAT ATG TGT GGA ATT GCT TCT CTG GCC AGT GTC CCG ATG GTG GCA CGA TTT CCG TGA TAA TTTGCT  1011
 307  R   G   N   M   C   G   I   A   S   L   A   S   V   P   M   V   A   R   F   P   *   *            326

1012 GTCATTATGGAGACGCAATGAACAATAAATCTCACTCGGCCTTGCACGGAATTC                                           1065
```

Fig. 3.2

Nucleotide and deduced amino acid sequence of *Fasciola hepatica* cathepsin L2. The predicted start of the propeptide (arrow) and mature enzyme (triangle) are indicated. The residues of the catalytic triad are encircled. Conserved cysteine residues (dotted squares) involved in disulphide bonds (dotted lines) in other cysteine proteinases are indicated. The poly-adenylation signal is underlined.

The predicted molecular mass for the precursor enzyme is 37,009 Da and for the mature protein is 24,459 Da. No N-linked glycosylation signals (NXT/S) were detected in the predicted amino acid sequence of the precursor protein. The various structural motifs, active site residues , cysteine residues involved in disulphide bonds and positions of cleavage between pre and propeptide, and propeptide and mature enzymes are indicated in Fig. 3.2.

3.3 Comparison of the FheTCL2 sequence with other *Fasciola* cathepsin L genes.

Alignment of the deduced aminoacid sequence of FheTCL2 with human cathepsin L, demonstrated that it belongs to the same family of cysteine proteinases, the papain superfamily (Fig. 3.3). The sequence of pFheTCL2 exhibits an 86 % identity at the nucleotide level and a 78% identity in amino acids to the cathepsin L1 isolated in our laboratory (Roche *et al.*,1997), confirming the existence of at least two different cathepsin Ls in the E/S of liver fluke.

Heussler and Dobbelaere (1994) previously detected using PCR five cathepsin L-like genes and two cathepsin B-like genes in *F. hepatica*, and advanced the hypothesis of the existence of a multigene family. Other *Fasciola hepatica* cysteine proteases of the papain family have been cloned by Wijffels *et al.* (1994), Yamasaki and Aoki (1993) and Panaccio *et al.* (1994). In order to understand the relationships between all these different clones their DNA and deduced amino acid sequences were aligned (Fig 3.3), and the nucleotide and amino acid identity between them deduced (Table 3.1).

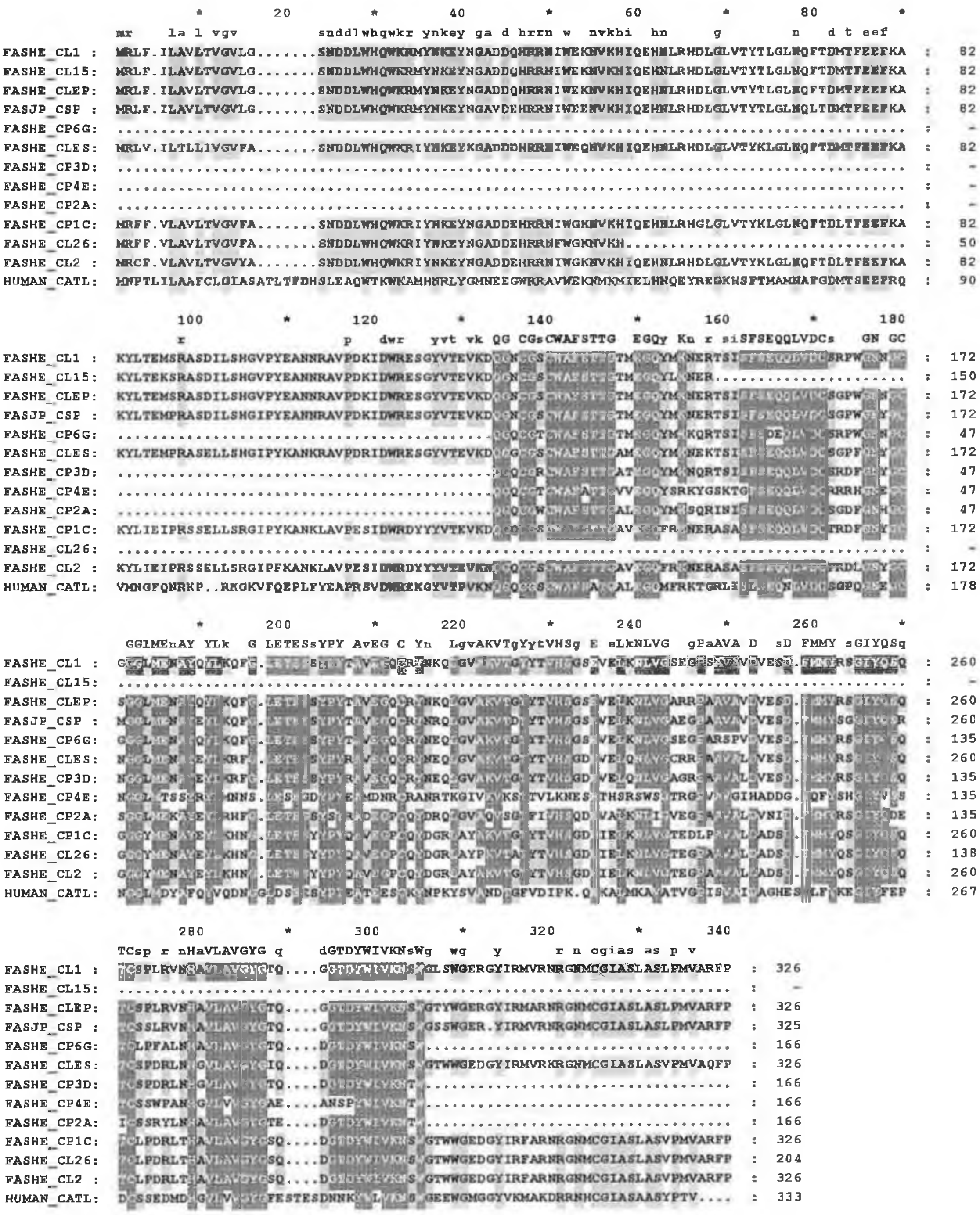


Fig.3.3.

Alignment of the deduced amino acid sequences of the cathepsin L-like genes of *Fasciola hepatica*.

Conserved residues are shaded. A consensus sequence is indicated in the first row of the alignment (residues conserved in all the sequences are indicated in Uppercase). Sequences are FASHE CP1C, FASHE CP2A, FASHE CP3D, FASHE CP4E, FASHE CP6G (Heussler and Dobbelaere, 1994), FASHE CLEP (Wijffels *et al.*, 1994), FASHE CLES (Panaccio *et al.*,1994), FASJP CSP (Yamasaki and Aoki, 1993), FASHE CL1 (Roche *et al.*, 1996), FASHE CL2, FASHE CL15, FASHE CL26 (this work) and HUMAN CATL (Mason *et al.*, 1986).

Cathepsin L2 shows a very high, 98% and 97%, identity at DNA and aminoacid level respectively with the clone Fcp1 (FASHE CP1C) of Heussler and Dobbelaere (1994). In contrast Heussler and Dobbelaere's Fcp4 (FASHE CP4E) showed 55% and 47% identity to cathepsin L2 at the nucleotide and amino acid level respectively. Intermediate values were obtained in the comparison with the other *F. hepatica* sequences . Cathepsin L1 (Roche *et al.* 1997) on the other hand, is clearly very similar to the genes cloned by Wijffels *et al.* (1994) from *F. hepatica* (FASHE CLEP) and Yamasaki and Aoki (1993) from *Fasciola sp.*(FASJP CSP); these showed 97% and 96% identity at the nucleotide level and 98 and 94% identity at the amino acid level respectively. A partial sequence, Fcp6 (FASHE CP6G), is the closest relative to cathepsin L1 of the Heussler and Dobbelaere's clones (93% and 90% identity in DNA and amino acid sequence, respectively). Partial sequences of two other clones purified at the same time as cathepsin L1 and L2 (FASHE CL1-5 and FASHE CL2-6) were obtained, showing very high homology with their complete counterparts (Table 3.1).

Only 8 out of 14 residues coincide between the deduced amino acid sequence of clone FheTCL2 and the N-terminal sequence obtained for the native liver fluke purified cathepsin L2 (Dowd *et al.*, 1994) (Fig. 3.4). On the other hand, only one difference can be detected between the deduced N-terminus of FheTCL1 clone (Roche *et al.*, 1997) and the sequence obtained for the native liver fluke cathepsin L1 (Smith *et al.*, 1995). The presence of a proline in position 7 (corresponding to a conserved tryptophan residue in other cathepsin L-like enzymes) in the sequence presented by Smith *et al.* (1995) is most probably explained by an error in the amino acid sequencing; alternatively a polymorphism can account for the differences at this site. Remarkably, this position was the only difference detected between both N-terminal sequences, suggesting that cross-contamination of the samples should not be ruled out.

Aminoacid identity

	CL1	CL15	CLEP	CSP	CP6G	CLES	CP3D	CP4E	CP2A	CP1C	CL26	CL2	HCATL
CL1	-	0.987	0.972	0.942	0.904	0.856	0.849	0.488	0.717	0.779	0.789	0.782	0.455
CL15	0.989	-	0.987	0.947	0.800	0.847	0.760	0.600	0.720	0.793	0.840	0.787	0.412
CLEP	0.976	0.989	-	0.938	0.880	0.868	0.861	0.482	0.723	0.776	0.789	0.779	0.458
CSP	0.961	0.960	0.959	-	0.849	0.852	0.837	0.500	0.729	0.766	0.773	0.772	0.443
CP6G	0.934	0.867	0.919	0.907	-	0.825	0.837	0.476	0.711	0.729	0.776	0.747	0.479
CLES	0.888	0.871	0.891	0.889	0.863	-	0.934	0.488	0.741	0.807	0.794	0.804	0.445
CP3D	0.863	0.773	0.867	0.855	0.873	0.948	-	0.500	0.771	0.783	0.765	0.789	0.497
CP4E	0.555	0.600	0.553	0.563	0.569	0.553	0.571	-	0.470	0.464	0.412	0.476	0.479
CP2A	0.801	0.760	0.797	0.811	0.821	0.811	0.831	0.571	-	0.687	0.672	0.699	0.497
CP1C	0.865	0.876	0.869	0.865	0.823	0.864	0.831	0.545	0.775	-	0.971	0.969	0.414
CL26	0.860	0.880	0.855	0.861	0.837	0.840	0.809	0.522	0.773	0.985	-	0.975	0.407
CL2	0.866	0.871	0.863	0.868	0.833	0.856	0.819	0.551	0.781	0.987	0.990	-	0.423
HCATL	0.506	0.471	0.511	0.504	0.547	0.507	0.549	0.540	0.560	0.510	0.520	0.514	-

Nucleotide identity

Table 3.1
Nucleotide and amino acid identity between the cloned cathepsin L-like proteinases of *Fasciola hepatica*. The gene names are abbreviations from the ones indicated in Fig. 3.3. Human cathepsin L has been included for comparison.

```
                                1       5         10       14

FheTCL2                       A V P E S I D W R D Y Y Y V T
F. hepatica cathepsin L2 A V P D K I D R R E S G Y V -
FheTCL1                       A V P D K I D W R E S G Y V T
F. hepatica cathepsin L1 A V P D K I D P R E S G Y V T
Human cathepsin L               A P R S V D W R E K G Y V T
Chiken cathepsin L              A P R S V D W R E K G Y V T
Rat cathepsin L                 I P K T V D W R E K G C V T
```

Figure 3.4
Comparison of N-terminal amino acid sequence of liver fluke cathepsin L-like proteinases.
The predicted sequence of the clone sequenced (FheCL2) is shown compared to the N-terminal sequence of *F. hepatica* cathepsin L2 (Dowd *et al.* 1994), cloned *F. hepatica* cathepsin L1 (Roche *et al.* 1997), *F. hepatica* cathepsin L1 (Smith *et al.*, 1994a) and cathepsin L from chicken (Dufour *et al.*, 1987), rat (Ishidoh *et al.*, 1987b) and human (Mason *et al.*, 1986).

3.4 Analysis of the *F.hepatica* cathepsin L protease family

A dendrogram depicting the relationships between the different sequences is shown in Figure 3.5. The *Fasciola hepatica* cathepsin L-like genes can be divided in five groups: group I includes FASHE CL1, FASHE CLEP, FASJP CSP, and two partial sequences (FASHE CL1-5 and FASHE CP6G); group II is formed by FASHE CL2, FASHE CP1C and the partial sequence FASHE CL2-6; group III includes FASHE CLES and the partial sequence FASHE CP3D. The other two groups (IV and V) are formed by the partial sequences FASHE CP2A and FASHE CP4E respectively.

Only five aminoacid substitutions over 204 residues can be detected between FASHE CL2-6 and FASHE CL2, and ten replacements occur between FASHE CL2 and FASHE CP1C. This very low level of variation could suggest the presence of allelic variations of a unique *Fasciola hepatica* CL2 gene, or a multicopy family of different, but very closely related genes. A similar situation can be detected in the group I of sequences. Two residues in 150 differ between our two clones (FASHE CL1 and FASHE CL1-5), while 8 replacements are detected when comparing the 326 residues of FASHE CL1 and FASHE CLEP. In spite of being one residue shorter, the sequence obtained by Yamasaki and Aoki (1993) (FASJP CSP) for *Fasciola spp.* , is more than 94% identical to the other members of the group. Hence, the possibility of polymorphic variants of a single gene cannot be ruled out.

In general we can conclude that there are at least five different cathepsin L-like genes in *Fasciola hepatica*, and evidence for polymorphic variation exists for at least two of them.

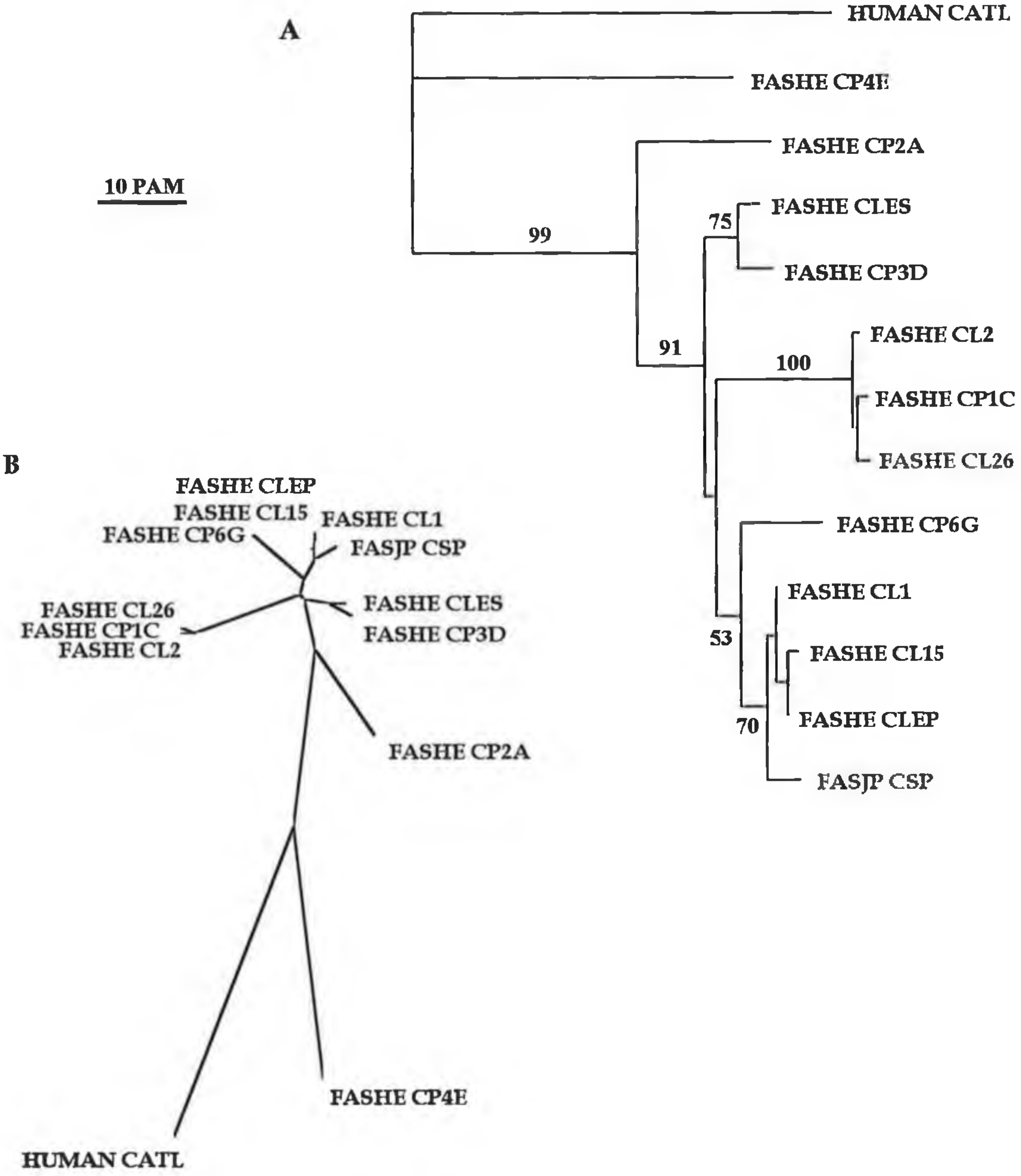


Fig. 3.5

Relationships between the several cathepsin L-like proteinases of*Fasciola hepatica*. From the unrooted dendrogram (A) it results clear the existence of at least 5 different groups of genes. (B)The same tree was rooted with the human cathepsin L . Bootstrap values are indicated.

3.5 Modeling of Cathepsin L1 and cathepsin L2

The general 3D structure of members of the papain superfamily is very conserved as shown by the superimposition of residues of those proteins that have been crystallized (Musil *et al.*, 1991; O'Hara *et al.*, 1995; Coulombe *et al.*, 1996), allowing the generation of models for those members of the superfamily still not crystallized. The models obtained, although might not represent perfectly the actual structure of the protein, are valid approximations that can help in the comprehension of particular features of a given protein. As a first attempt to understand at the structural level the differences observed in the deduced amino acid sequences of liver fluke cathepsin L1 and cathepsin L2, tridimensional models of the proteins were made using the Swiss-Model server (Peitsch, 1996) (http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html). The 3D structures of caricain (1PPO), and actinidin (2ACT, 1AEC) were used as frames for the model-building because they give the highest scores in the alignments.

The models obtained for both sequences showed no major structural differences. Of the 71 different residues between the cathepsin L1 and cathepsin L2 amino acid sequences, 4 map in the signal peptide, 17 in the pro-region, and the remaining 50 in the mature enzyme (Fig 3.3). Only the mature part of the proteins were modeled, because of the absence of crystallographic coordinates at the moment for the modeling of the propeptides. The comparison of the models generated showed that only five residues (Ala^{33}, Val^{34}, Pro^{57}, Ala^{132} and Phe^{196} in cathepsin L2) are buried in the interior of the protein, three (Leu^{136}, Ala^{197}, and Val^{212}) are partially accessible, and the remaining 42 residues that differ between both molecules lie on the surface or are solvent accessible (Fig. 3.6). These represent a 28 % of the total surface of the molecule. Two of the different residues lie in the active site cleft, those in positions 69 and 160 of the models (being Leu^{69} and Val^{160} in CL1 and Tyr^{69} and Leu^{160} in CL2).

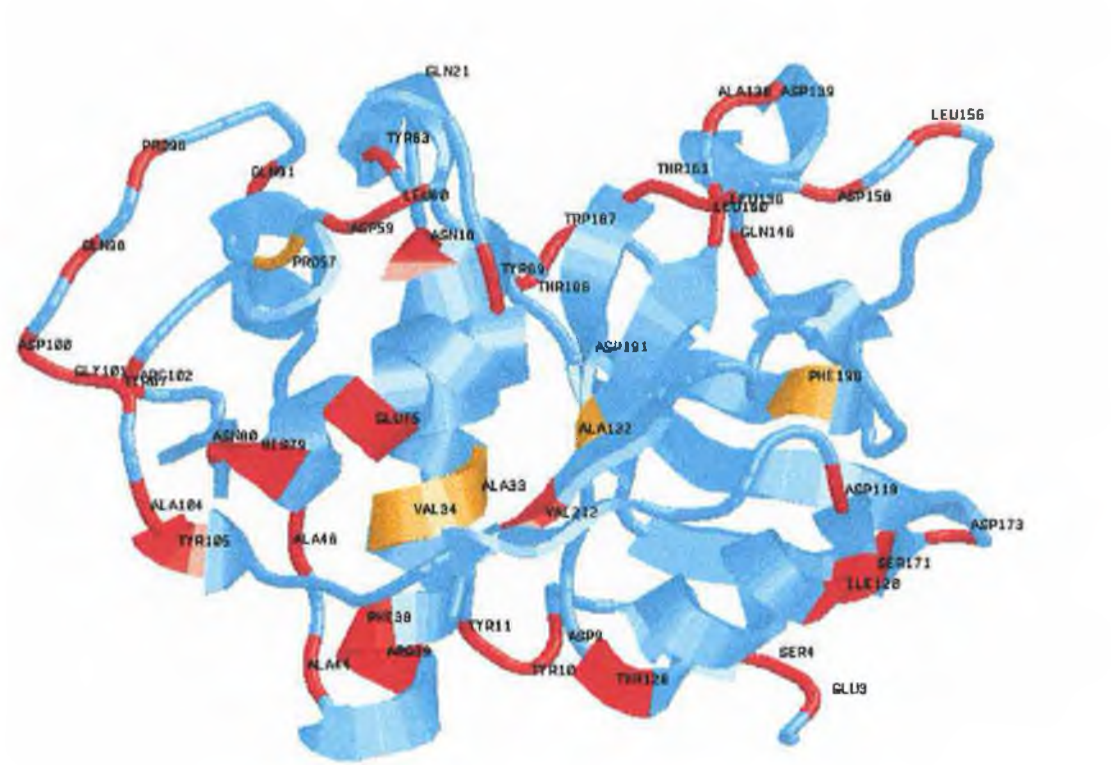


Fig.3.6

Model of the *Fasciola hepatica* cathepsin L2.
The modeled structure is shown as a cartoon, with the residues that differ between cathepsin L1 and cathepsin L2 indicated. Very few residues that account for differences in the two proteins are buried in the interior of the molecule (yellow), while the majority of the differences are in surface areas (red).

4. Yeast expression of *Fasciola hepatica* Cathepsin L2

4.1 Cloning in pAAH5 shuttle expression vector.

In order to analyze the biochemical properties of the enzyme encoded by the isolated cathepsin L2 cDNA, this was subcloned into the *Saccharomyces cerevisiae* expression plasmid pAAH5. This plasmid is a shuttle vector with the yeast replication region of the 2-micron circle and the *E. coli* replication origin of pBR322, allowing autonomous replication both in yeast and in bacteria. The yeast LEU2 gene for leucine synthesis provides a selective marker in *S. cerevisiae*, and the ampicillin resistance gene can be used for selection of transformants in *E.coli*. The vector contains a unique *Hin* dIII cloning site, located between the control regions of the alcohol dehydrogenase I gene (ADC1) from *S. cerevisiae*. At the 5' side of the restriction site the promoter and the untranslated leader of the yeast gene are located, and it is flanked at the 3' side by the C-terminal and neighboring region of the same gene which contains a transcription termination signal and other sequences that may contribute to RNA stability (Ammerer, 1983). The translation signals for initiation and termination, however, in the case of the cathepsin L2 are provided by the insert, while the ribosomal binding site is included in the 5' region from ADC1 (see Figure 2.1).

The coding region of FheCL2 was amplified by PCR using two primers with *Hin* dIII adapters, designed based on the sequences across the start and stop codons of *Fasciola hepatica* cathepsins L1 and L2 (these primers were termed URUF and IREB). An internal *Hin* dIII site is present in the cathepsin L2 cDNA, therefore it was impossible to obtain full length inserts upon digestion of the PCR product. To overcome this problem, the whole PCR product was first subcloned into pGEMT (which has a thymidine overhang, and is suitable for cloning PCR products irrespective of the ends present in them) The resulting plasmid (termed pFheHCL2), was subjected to controlled partial digestions with *Hin* dIII. A minor band of the expected size (1Kb) corresponding to the full length insert was detected, purified and ligated to *Hin* dIII linearized phosphorylated pAAH5 plasmid. The complete sequence of both strands of pFheHCL2 was determined in order to confirm that no mutation occurred due to reading errors in the PCR amplification step.

The orientation of the insert in the pFheHCL2 plasmid was assessed by restriction mapping with *Bam* HI and *Eco* RI double digestions, and by PCR using the primers URUF, IREB and a pAAH5 primer (termed PBR). A clone, termed pFheYCL2, was selected for further characterization.

4.2 Profile of cathepsin L2 activity correlating with the yeast culture growth curve.

Yeast DBY746 cells were transformed with the pFheYCL2 and control plasmid pAAH5 and cultured in buffered selective minimal medium. Samples were taken at time intervals and assayed for cathepsin L2 activity in cell extracts (CE) and culture media supernatants (CS). The fluorogenic peptide substrate benzyloxy carbonyl-phenylalanine-arginine-7-amino-4-methylcoumarin (Z-Phe-Arg-NHMec) was used as a substrate. This substrate is efficiently hydrolyzed by cathepsins Ls in the presence of the activator reducing agent dithiothreitol (DTT), releasing the NHMec, which can be quantified by its intense fluorescence (Barret and Kirschke, 1981). Z-Phe-Arg-NHMec cleaving activity was found in CS and to a lesser extent in CE. Yeasts transformed with pAAH5 control plasmid showed no activity. This activity was demonstrated to correspond to a cystein-proteinase as it was enhanced 10-fold by 0.5 mM DTT and was undetectable in the presence of the specific inhibitor 50 mM E-64.

The growth curve of *S. cerevisiae* DBY746 cells transformed with pFheYCL2 is shown in Fig. 4.1. The growth curve of the cells transformed with the control plasmid pAAH5 is similar to the one shown. The doubling time was calculated to be 3.7 h. The similarity in the growing patterns of yeast cells transformed with the recombinant cathepsin L2 and the control plasmid indicates that the expression of the recombinant protein has no detectable deleterious effects on the yeast strain used.

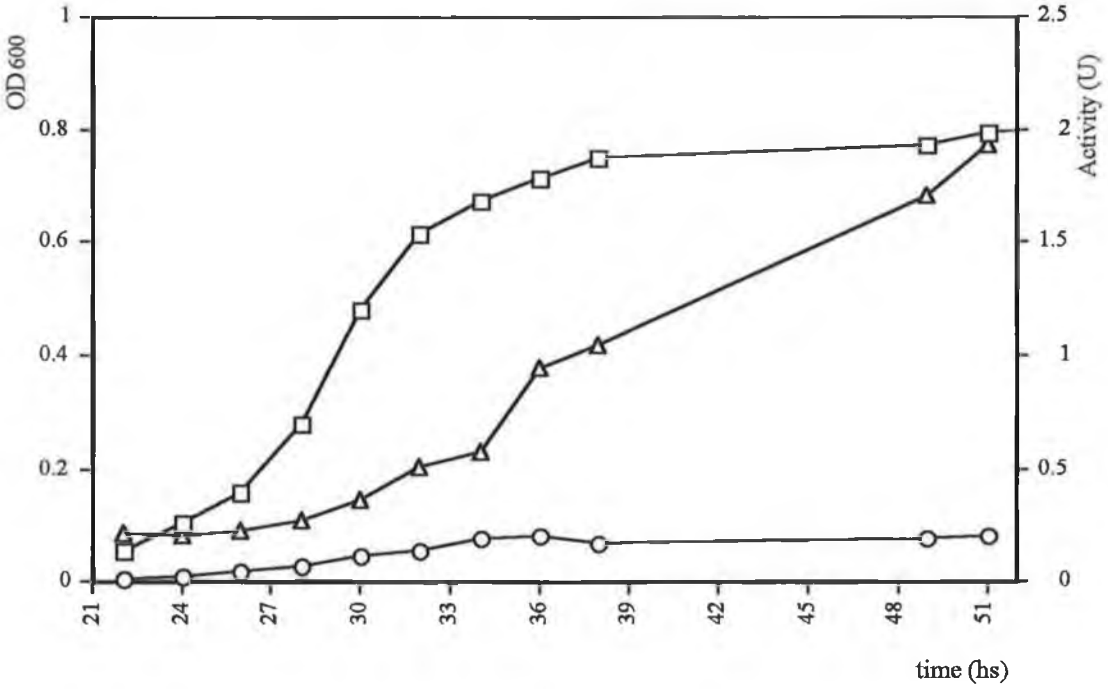


Fig 4.1

Time course production of recombinant cathepsin L2 at pH 6.5

Yeast cells transformed with pFheYCL2 were cultured in buffered medium (pH 6.5) and samples were taken at time intervals. The yeast growth was monitored by the absorbance at 600 nm (rectangles). The cysteine proteinase activity of culture supernatants (triangles) and cell extracts (circles) was assayed using the fluorogenic substrate Z-Phe-Arg-NHMec.

The production of active recombinant cathepsin L2 monitored over a typical culturing period is shown in Fig 4.1. The profile of the activity in the culture media (culture supernatants) follows a similar curve to that of the cell growth. The extra-cellular cathepsin L activity per ml of culture was approximately three times or greater than the intracellular activity after 36 hs. of culture. The early stages of the stationary phase of DBY746 cultures (after 48 hs.) were chosen for routinely harvesting cells and culture media for biochemical characterization and enzyme purification.

4.3 Immunoblotting of cathepsin L2 in yeast cell extracts and culture media.

To confirm that the protein expressed by the yeast was the liver fluke cathepsin L2, CE and CS were subjected to western blot analysis using anti-cathepsin L1 and L2 antisera . Native cathepsin L1 and L2 purified from mature *F. hepatica* excretory/secretory (E/S) products, and CE and CS of yeasts transformed with pFheYCL2 and pFheYCL1 were analyzed in the same gels (Fig 4.2). A major band of approximately 24,000 Da was detected by anti-cathepsin L2 sera in the lanes corresponding to the pFheYCL2 clone. Minor bands of higher molecular weight were also apparent, including a band of approximate molecular mass 37,000 Da that could correspond to the precursor pro-cathepsin L. Minor bands were detected on the lanes corresponding to the pFheYCL1 clone. On the contrary, when probed with anti-cathepsin L1, bands could be detected in the pFheYCL1 lanes but not on the pFheYCL2 lanes. This pattern was consistent in several western blots. The product of pFheYCL2 can be detected sometimes with anti-cathepsin L1 as a faint band, confirming the reaction pattern obtained in the purification procedure (section 3.1). These results clearly demonstrate that the product of the gene expressed in *S.cerevisiae* is cathepsin L2 rather than cathepsin L1 characterized by Roche *et al.* (1997). It is interesting to note that the enzyme is processed in the yeast system to generate a mature polypeptide of the same size (or very similar) to the one produced in the adult liver fluke. This can be considered as an indication of recognition and functionality of the processing signals of the *F.hepatica* gene in the *S. cerevisiae* system.

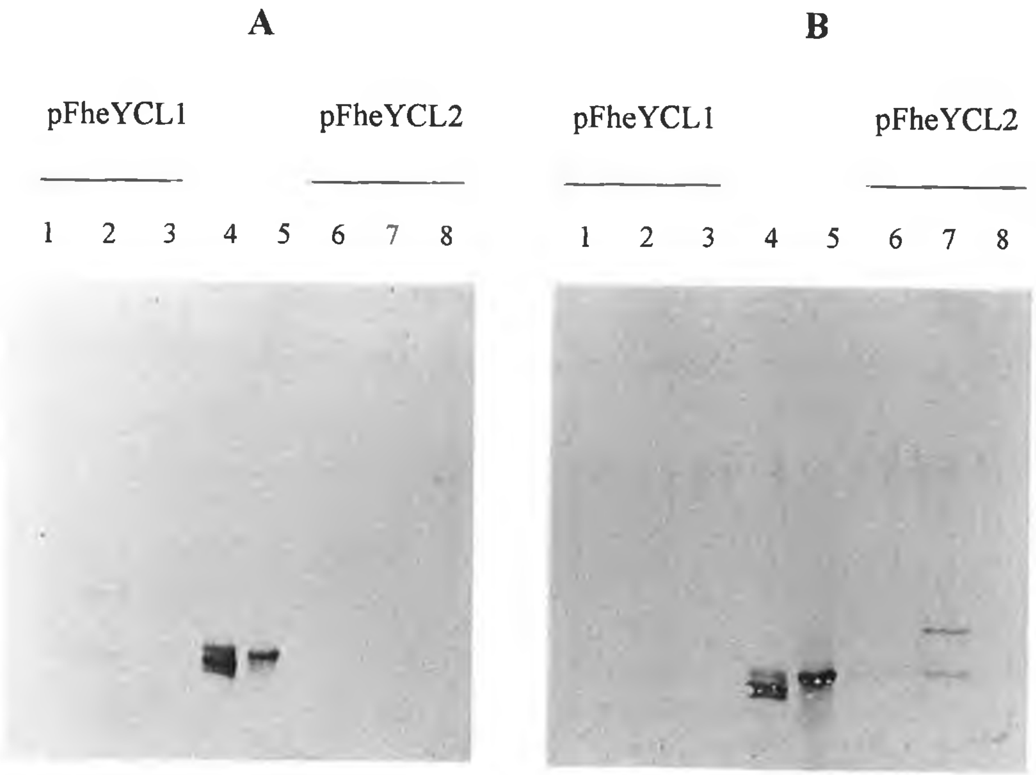


Fig 4.2

Immunoblot analysis of culture supernatants and cell extracts of yeast transformed with pFheYCL1 and PfheYCL2 .

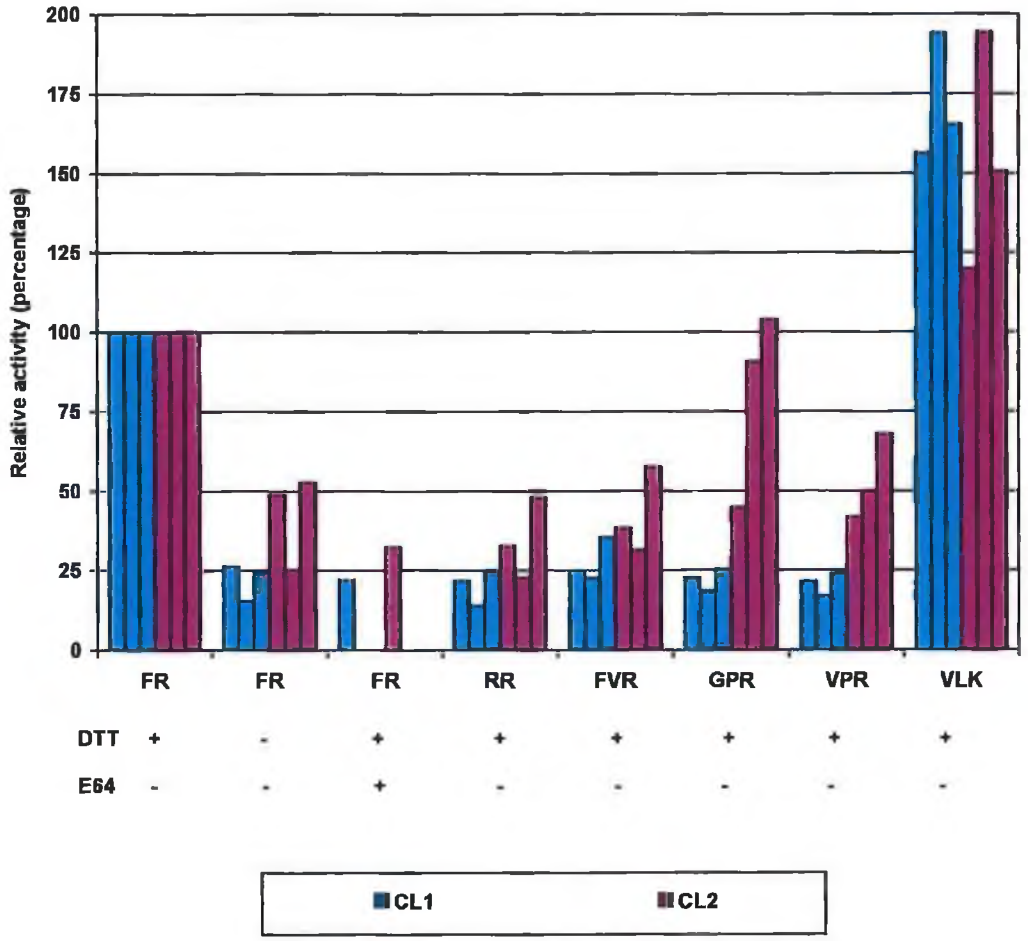
Culture supernatants (lanes 1 and 6), concentrated culture supernatants (lanes 2 and 7), and cell extracts (lanes 3 and 8) of yeast transformed with pFheYCL1 and PfheYCL2 respectively , and purified *F.hepatica* native cathepsin L1 (lane 4) and cathepsin L2 (lane 5) were immunobloted with anti-cathepsin L1 serum (panel A) or anticathepsin L2 serum (panel B).

4.4 Comparison on recombinant Cathepsin L1 and Cathepsin L2 activities on different synthetic substrates

Culture supernatants of yeast DBY746 cells transformed with pFheYCL1 or pFheYCL2 prepared under identical conditions were assayed for proteolytic activity against several synthetic fluorogenic substrates. The relative activities of both enzymes (in culture supernatants, CS) towards different substrates are shown in Fig. 4.3. The activities for both enzymes were enhanced by the addition of dithiothreitol (DTT), a typical characteristic of cysteine proteinases. Furthermore, the addition of the cysteine protease specific inhibitor E64, greatly reduced the activity observed

The preference of the substrate Z-Phe-Arg-NHMec over Z-Arg-Arg-NHMec distinguishes cathepsin L and cathepsin S from cathepsin B enzymes (Barrett and Kirsche, 1981). Cathepsin S and L can be differentiated based on the higher activity of the first enzyme towards the substrate Z-Phe-Val-Arg-NHMec, while cathepsin L has a ten-fold preference for the substrate Z-Phe-Arg-NHMec over the substrate Z-Phe-Val-Arg-NHMec (Bromme *et. al.*, 1989). The CS of both transformants (namely pFheYCL1 and pFheYCL2) showed a preference for the substrate Z-Phe-Arg-NHMec over Z-Arg-Arg-NHMec and Z-Phe-Val-Arg-NHMec, confirming their classification as cathepsins L-like enzymes.

Dowd *et al.* (1994) showed that the *Fasciola hepatica* cathepsin L2 can cleave the substrate Tos-Gly-Pro-Arg-NHMec more efficiently than cathepsin L1. In agreement with their report, the recombinant CL2 showed a two to three fold higher efficiency over Tos-Gly-Pro-Arg-NHMec than the recombinant CL1, confirming that the cloned enzyme was the *Fasciola hepatica* cathepsin L2. Both enzymes have a marked preference for the substrate Boc-Val-Leu-Lys-NHMec with a three-fold higher efficiency over the substrate Z-Phe-Arg-NHMec, an uncommon characteristic for cathepsin Ls (Dowd *et al.* , 1994). The higher activity on the substrate Boc-Val-Leu-Lys-NHMec over the substrate Z-Phe-Arg-NHMec is shared by both recombinant enzymes also indicating that the former is probably a better substrate for the liver fluke enzymes than the latter.



FR	Z-Phe-Arg-NHMec
RR	Z-Arg-Arg-NHMec
FVR	Bz-Phe-Val-Arg-NHMec
GPR	Tos-Gly-Pro-Arg-NHMec
VPR	Boc-Val-Pro-Arg-NHMec

Fig 4.3

Relative acivities of culture supernatants of yeast DBY746 cells transformed with pFheYCL1 or pFheYCL2 prepared under identical conditions towards different synthetic substrates. Three independent experiments are shown.

4.5 Purification of the recombinant Cathepsin L2

Recombinant cathepsin L2 (RCL2) was purified from concentrated yeast culture supernatants by gel filtration on Sephadex S200HR. Fractions were assayed for cathepsin L activity using the fluorogenic substrate Z-Phe-Arg-NHMec. The results of Z-Phe-Arg-NHMec cleaving assays of the factions eluted from the S200HR column show that there are 3 peaks of Z-Phe-Arg-NHMec cleaving activity (peaks I, II and III) (Fig. 4.4.A). Peak III has the majority of the Z-Phe-Arg-NHMec cleaving activity. Only the activities in the second and third peak were inhibited by the cathepsin L inhibitor Z-Phe-Ala-CHN_2, and enhanced by the reducing agent cysteine. The first peak of activity (peak I) which eluted with the main protein peak may represent an endogenous yeast proteinase.

The fractions of peaks I, II and III were separately pooled and aliquots were subsequently analyzed by immunoblotting and SDS-PAGE under reducing and denaturing conditions (Fig. 4.4.B and 4.4.C). Peak I contained most of the proteins present in the culture supernatant, but no reactivity bands were detected in the immunoblot with anti-cathepsin L2 in this fraction. Peak II consisted of two proteins of 28 and 29.5 kDa, but only the latter was reactive with the anti-cathepsin L2 serum. The 29.5 kDa. protein was the only one visualized in the peak III. The 29.5 protein co-migrated with native cathepsin L2 purified from *F.hepatica* ES products. A minor band of approximate molecular weight 37 kDa was visualized in peak II. This band was recognized by the anti-cathepsin L2 serum, and corresponded well with the predicted size for the proform of the enzyme, i.e. approximately 37 kDa +/- 2 kDa.

The recombinant cathepsin L2 proteinase isolated has a specific activity of 0.0026U/mg (peak II) and 0.0226U/mg (peak III) which represents a purification of 1.7-fold and 15-fold for peaks II and III respectively. Peak III contained 2.3 mg protein from 30 l of fermentation broth; therefore, the yield of cathepsin L2 was 0.76 mg/l. Yeast expressed cathepsin L2 peak III was subjected to N-terminal sequencing to determine the exact sequence at the start of the mature active enzyme. The first 3 amino-acids from the N-terminal end of the purified protein were A-P-D which corresponds well with the N-terminal sequence obtained by Dowd *et al* (1994) for the native cathepsin L2 .

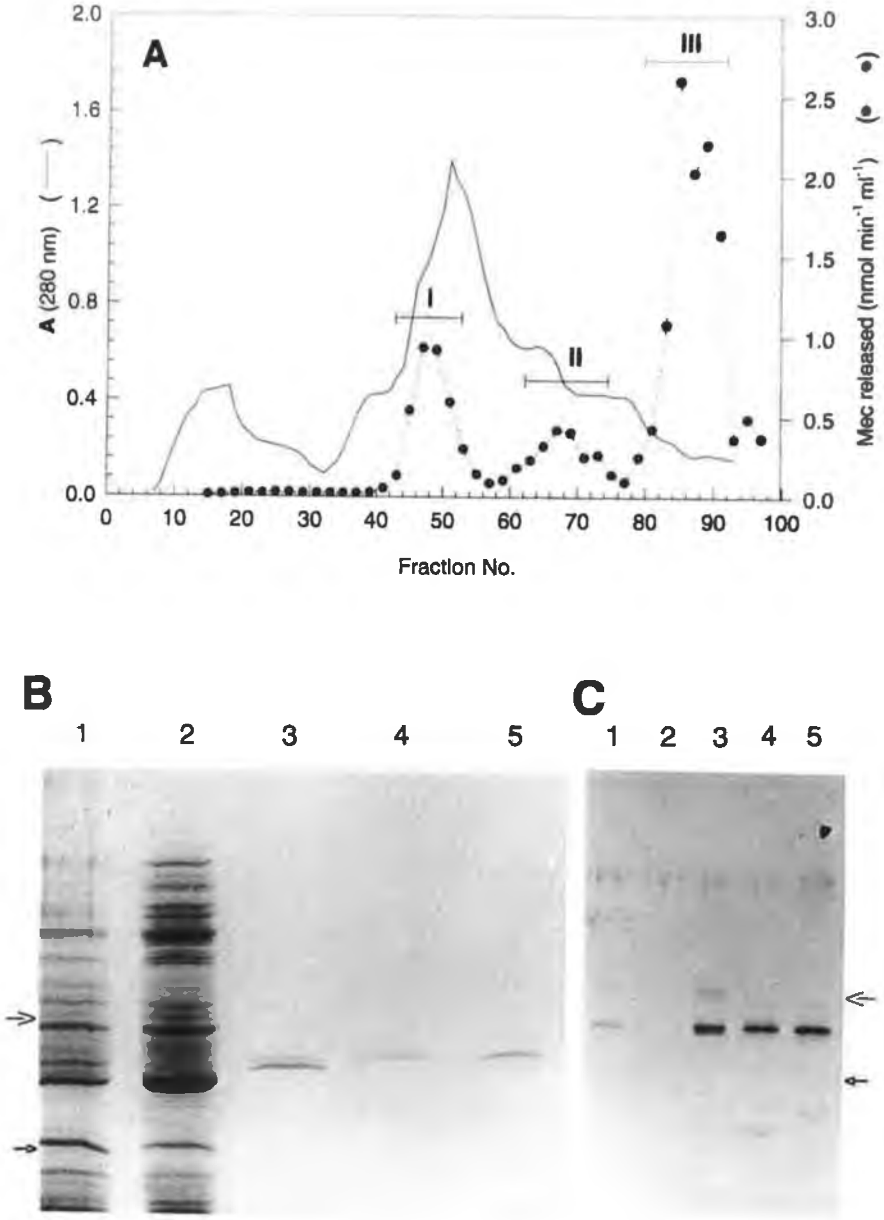


Fig. 4.4
Purification of recombinant cathepsin L2.
(A), profile of protein elution from a Sephacryl S200HR column monitored by absorbance at 280 nm (lines) and cysteine proteinase activity in collected fractions assayed using the fluorogenic substrate Z-Phe-Arg-NHMec (dots). The fractions pooled are indicated.
(B) SDS-PAGE under reducing conditions and immunoblot (C) analysis of culture supernatant of yeast cells transformed with pFheYCL2 (lane 1), peaks I, II and III (lanes 2, 3 and 4 respectively), and purified native cathepsin L2 (lane 5). The position of the markers triose-phospate isomerase (36 kDa) and lysozyme (14.4 kDa) are indicated by arrows.

4.6 Substrate specificy of purified recombinant cathepsin L2

In collaboration with Dowd we analyzed the substrate specificity of the yeast-expressed and native cathepsin L2 using the fluorogenic substrates Z-Arg-Arg-NHMec, Z-Phe-Arg-NHMec and Tos-Gly-Pro-Arg-NHMec (Table 4.1). Both yeast-expressed and native *F. hepatica* enzymes showed a marked preference (as assessed by k_{cat}/K_m) for the cathepsin L substrate Z-Phe-Arg-NHMec over the cathepsin B substrate Z-Arg-Arg-NHMec. Most importantly, both enzymes eficiently cleaved the substrate Tos-Gly-Pro-Arg-NHMec . For all substrates examined, the native and recombinant enzymes exhibited similar k_{cat}/K_m values, confirming that the yeast expressed enzymes is in fact the *F. hepatica* cathepsin L2.

	Native cathepsin L2			Recombinant cathepsin L2		
substrate	K_m	k_{cat}	k_{cat}/K_m	K_m	k_{cat}	k_{cat}/K_m
	(uM)	(s^{-1})	($mM^{-1}s^{-1}$)	(uM)	(s^{-1})	($mM^{-1}s^{-1}$)
Z-Arg-Arg-NHMec	9.3	0.02	2.2	10.2	0.04	3.9
Z-Phe-Arg-NHMec	10.0	0.65	64.8	11.4	0.64	56.1
Tos-Gly-Pro-Arg-NHMec	25.0	1.0	40.0	25.0	1.35	54.0

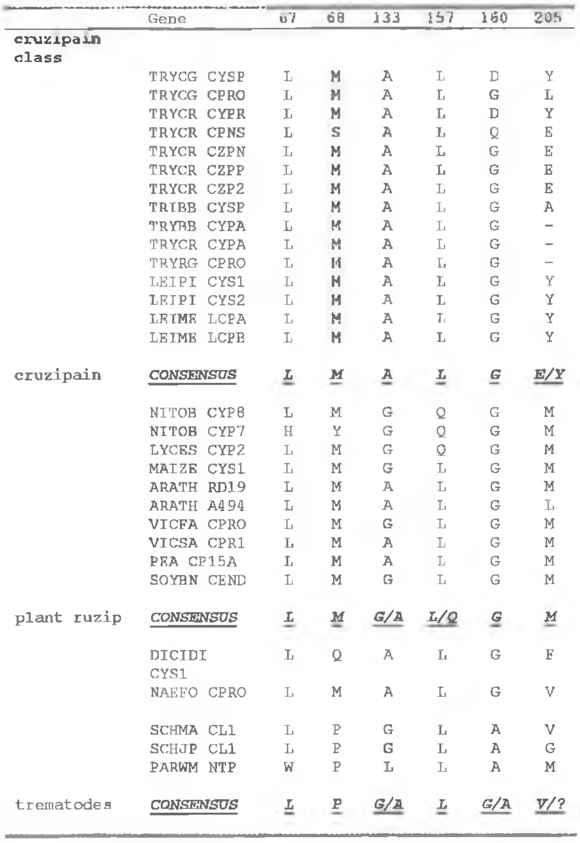
Table 4.1
Reaction kinetics for native cathepsin L2 and yeast expressed recombinat cathepsin L2 on fluorogenic pepetide substrates.

5 Mutagenesis of *Fasciola hepatica* Cathepsin L1

5.1 Mutagenesis of *Fasciola hepatica* Cathepsin L1

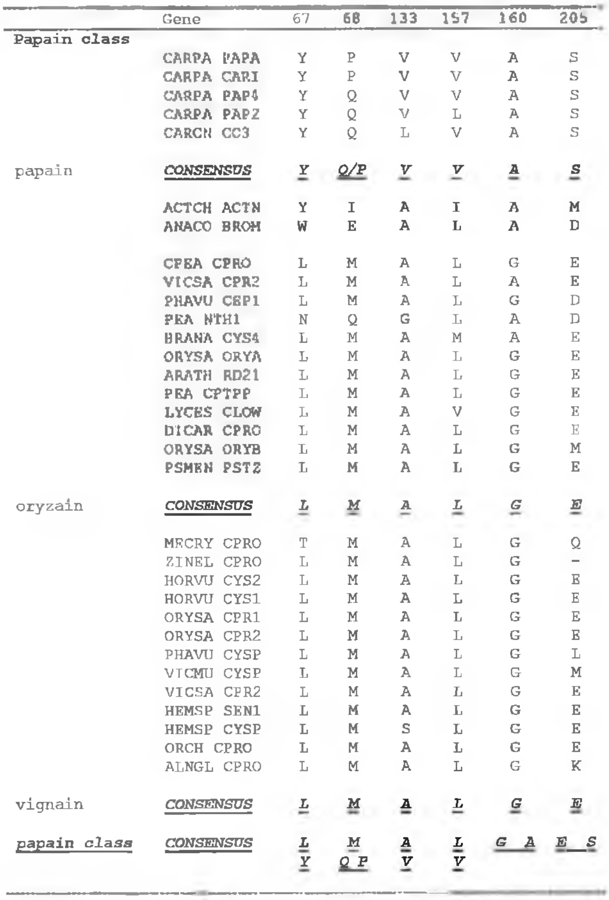
The different substrate specificities of cysteine proteinases of the papain superfamily is determined by the presence of distinct residues in the active site of the enzymes, allowing distinct contacts to be made between the enzyme and the substrate. Schechter and Berger (1967) proposed that the active site of papain can be considered to consist of seven subsites (S_1-S_4 and S_1'-S_3'), each able to accommodate one amino acid residue of a substrate (P_1-P_4 and P_1'-P_3' respectively). The nature of the residue at position P_2 has been shown to be the most significant in terms of determining substrate specificity of cysteine proteinases. Khouri *et al.* (1991) analyzed the crystal structures of papain published by Drenth *et al.*(1976), and identified the residues in the S_2 subsite whose side chains make the most intimate contacts with the P_2 side chain of the substrate; these were at position 67, 68, 133, 157, 160 and 205 (papain numbering). The elucidation of the crystal structure of several other members of the papain family showed that the general structure of these proteins is conserved, with the main differences occurring in external loops away from the active site cleft. The availability of these other structures allowed a more detailed analysis of the residues that lie in the interface of the molecule making contacts with the substrate, and a better alignment of the sequences of other members of the family whose crystal structures still remain unknown (Heinemann *et al.*, 1982; Kamphuis *et al.*; 1984; Baker and Dobson, 1980; Pickersgill *et al.*, 1991; Musil *et al.*, 1991; Jia *et al.*, 1995; Mc Grath *et al.*, 1995) In this study we compared the residues found in the S_2 subsite of the active site of several members of the papain superfamily (Table 5.1and summarized in Table 5.2).

Table 5.1 S$_2$ subsite comparison between members of the papain superfamily



a) Members of the cruzipain class

	Gene	67	68	133	157	160	205
cruzipain class							
	TRYCG CYSP	L	M	A	L	D	Y
	TRYCG CPRO	L	M	A	L	G	L
	TRYCR CYPR	L	M	A	L	D	Y
	TRYCR CPNS	L	S	A	L	Q	E
	TRYCR CZPN	L	M	A	L	G	E
	TRYCR CZPP	L	M	A	L	G	E
	TRYCR CZP2	L	M	A	L	G	E
	TRIBB CYSP	L	M	A	L	G	A
	TRYBB CYPA	L	M	A	L	G	-
	TRYCR CYPA	L	M	A	L	G	-
	TRYRG CPRO	L	M	A	L	G	-
	LEIPI CYS1	L	M	A	L	G	Y
	LEIPI CYS2	L	M	A	L	G	Y
	LEIME LCPA	L	M	A	L	G	Y
	LEIME LCPB	L	M	A	L	G	Y
cruzipain	***CONSENSUS***	***L***	***M***	***A***	***L***	***G***	***E/Y***
	NITOB CYP8	L	M	G	Q	G	M
	NITOB CYP7	H	Y	G	Q	G	M
	LYCES CYP2	L	M	G	Q	G	M
	MAIZE CYS1	L	M	G	L	G	M
	ARATH RD19	L	M	A	L	G	M
	ARATH A494	L	M	A	L	G	L
	VICFA CPRO	L	M	G	L	G	M
	VICSA CPR1	L	M	A	L	G	M
	PEA CP15A	L	M	A	L	G	M
	SOYBN CEND	L	M	G	L	G	M
plant ruzip	***CONSENSUS***	***L***	***M***	***G/A***	***L/Q***	***G***	***M***
	DICIDI CYS1	L	Q	A	L	G	F
	NAEFO CPRO	L	M	A	L	G	V
	SCHMA CL1	L	P	G	L	A	V
	SCHJP CL1	L	P	G	L	A	G
	PARWM NTP	W	P	L	L	A	M
trematodes	***CONSENSUS***	***L***	***P***	***G/A***	***L***	***G/A***	***V/?***



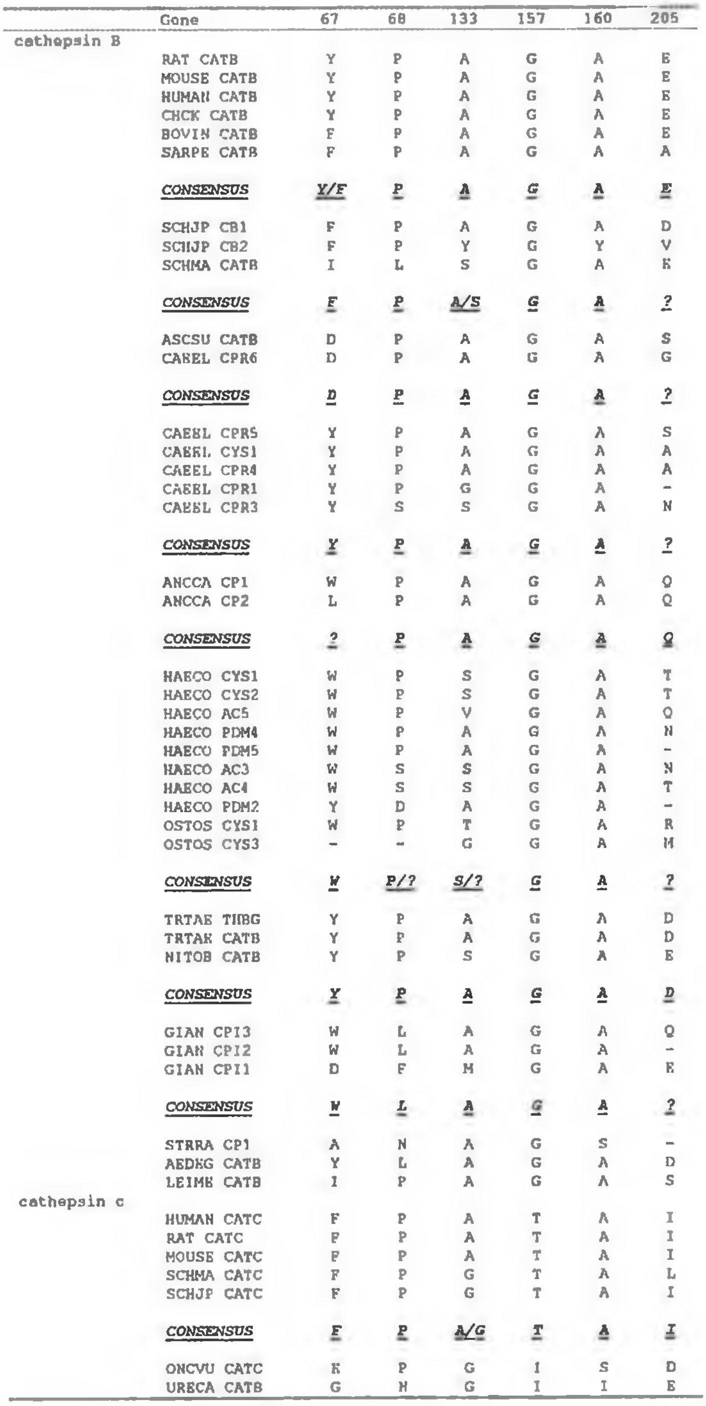
b) Members of the papain class

	Gene	67	68	133	157	160	205
Papain class							
	CARPA PAPA	Y	P	V	V	A	S
	CARPA CARI	Y	P	V	V	A	S
	CARPA PAP4	Y	Q	V	V	A	S
	CARPA PAP2	Y	Q	V	L	A	S
	CARCN CC3	Y	Q	L	V	A	S
papain	***CONSENSUS***	***Y***	***Q/P***	***V***	***V***	***A***	***S***
	ACTCH ACTN	Y	I	A	I	A	M
	ANACO BROM	W	E	A	L	A	D
	CPEA CPRO	L	M	A	L	G	E
	VICSA CPR2	L	M	A	L	A	E
	PHAVU CEP1	L	M	A	L	G	D
	PEA NTH1	N	Q	G	L	A	D
	BRANA CYS4	L	M	A	M	A	E
	ORYSA ORYA	L	M	A	L	G	E
	ARATH RD21	L	M	A	L	G	E
	PEA CPTPP	L	M	A	L	G	E
	LYCES CLOW	L	M	A	V	G	E
	DICAR CPRO	L	M	A	L	G	E
	ORYSA ORYB	L	M	A	L	G	M
	PSMEN PSTZ	L	M	A	L	G	E
oryzain	***CONSENSUS***	***L***	***M***	***A***	***L***	***G***	***E***
	MECRY CPRO	T	M	A	L	G	Q
	ZINEL CPRO	L	M	A	L	G	-
	HORVU CYS2	L	M	A	L	G	E
	HORVU CYS1	L	M	A	L	G	E
	ORYSA CPR1	L	M	A	L	G	E
	ORYSA CPR2	L	M	A	L	G	E
	PHAVU CYSP	L	M	A	L	G	L
	VICMU CYSP	L	M	A	L	G	M
	VICSA CPR2	L	M	A	L	G	E
	HEMSP SEN1	L	M	A	L	G	E
	HEMSP CYSP	L	M	S	L	G	E
	ORCH CPRO	L	M	A	L	G	E
	ALNGL CPRO	L	M	A	L	G	K
vignain	***CONSENSUS***	***L***	***M***	***A***	***L***	***G***	***E***
papain class	***CONSENSUS***	***L*** ***Y***	***M*** ***Q P***	***A*** ***V***	***L*** ***V***	***G A***	***E S***

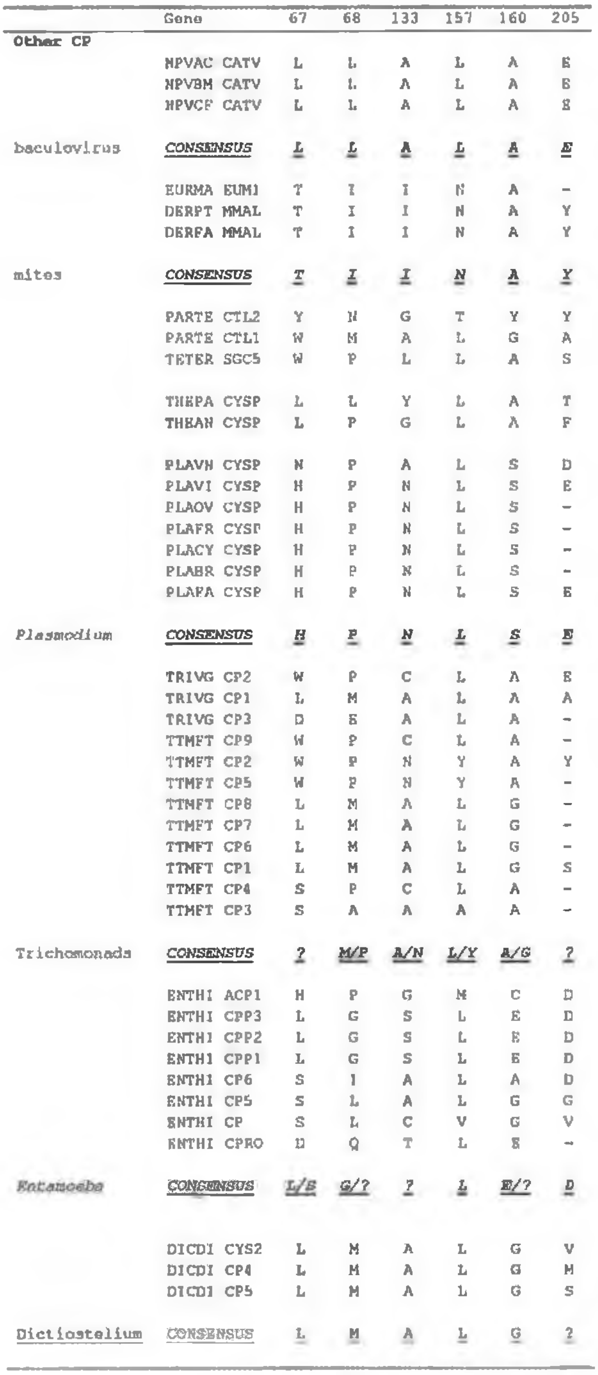
c) Members of the cathepsin L and cathepsin H classes

	Gene	67	68	133	157	160	205
Cathepsin L Class							
	MOUSE CATL	L	M	A	L	G	A
	RAT CATL	L	M	A	L	G	A
	HUMAN CATL	L	M	A	M	G	A
	PIG CATL	L	M	A	L	G	A
	SHEEP CATL	L	M	A	L	G	A
	RAT TEST	F	M	A	L	A	Y
	CAT CATL	L	I	A	V	G	A
	CHCK CATL	L	M	A	L	G	A
	TROUT CATL	L	M	A	L	G	A
	ZFISH CATL	L	M	A	L	G	K
	RAT CATRL2	T	A	A	V	A	C
cathepsin L	*CONSENSUS*	*L*	*M*	*A*	*L*	*G*	*A*
	HUMAN CATK	Y	M	A	L	A	L
	RABIT CATK	Y	M	A	V	A	L
	MOUSE CATK	Y	M	S	V	A	M
	CHCK JTAP	Y	M	G	I	A	L
	TROUT CATK	Y	M	A	L	G	-
cathepsin K	*CONSENSUS*	*Y*	*M*	*A*	*L/V*	*A*	*L*
	BOVIN CATS	F	M	G	V	G	Y
	RAT CATS	F	M	G	M	G	Y
	HUMAN CATS	F	M	G	V	G	F
	MOUSE CATS	Y	M	G	V	G	Y
	CARP CP	L	M	A	I	A	T
cathepsin S	*CONSENSUS*	*F*	*M*	*G*	*V*	*G*	*Y*
	SARPE CATL	L	M	A	L	G	A
	DROMME CPRO	L	M	A	L	G	P
	BOMMO CPRO	L	M	A	L	G	S
	PENVA PCP1	L	M	G	L	G	Q
	PENVA PCP2	L	M	A	L	G	Q
	HEPNO CLE	W	V	A	L	A	D
	HOMAM CYS1	W	V	A	L	A	D
	NEPNO CLS	W	M	A	L	G	E
	HOMAM CYS3	W	M	A	L	G	E
	HOMAM CYS2	W	M	T	L	A	V
decapod DCP	*CONSENSUS*	*L/W*	*M/V*	*A*	*L*	*G*	*?*
	SPIMA CPRO	F	M	G	I	G	V
	SPIER CPRO	L	M	G	I	G	M
	SCHMA CL2	T	M	A	L	G	N
	SCHJP CL2	T	M	G	I	G	N
Schistosoma	*CONSENSUS*	*?*	*M*	*G*	*I*	*G*	*N*
	FASHE CL1	**L**	**M**	**A**	**V**	**A**	L
	FASSP CYP	L	M	A	V	A	L
	FASHE CAT2	L	M	A	L	G	L
	FASHE CP2A	L	M	A	L	A	-
	FASHE CP4B	L	M	G	A	G	-
	FASHE CL2	**Y**	**M**	**A**	**L**	**A**	**L**
Fasciola	*CONSENSUS*	*L*	*M*	*A*	*L/V*	*A/G*	*L*
Cathepsin H Class							
	HUMAN CATH	L	P	A	V	A	C
	RAT CATH	L	P	A	V	A	C
	ORYSA ORYC	L	P	A	V	A	C
	HORVU ALEU	L	P	A	V	A	C
	MAIZE CYS2	L	P	A	V	A	C
	PEA NACP	L	P	A	V	A	C
	LYCES CYP3	L	P	A	V	A	C
	CONSENSUS	*L*	*P*	*A*	*V*	*A*	*C*

d) Members of the cathepsin B class



	Gone	67	68	133	157	160	205
cathepsin B							
	RAT CATB	Y	P	A	G	A	E
	MOUSE CATB	Y	P	A	G	A	E
	HUMAN CATB	Y	P	A	G	A	E
	CHCK CATB	Y	P	A	G	A	E
	BOVIN CATB	F	P	A	G	A	E
	SARPE CATB	F	P	A	G	A	A
	CONSENSUS	*Y/F*	*P*	*A*	*G*	*A*	*E*
	SCHJP CB1	F	P	A	G	A	D
	SCHJP CB2	F	P	Y	G	Y	V
	SCHMA CATB	I	L	S	G	A	K
	CONSENSUS	*F*	*P*	*A/S*	*G*	*A*	*?*
	ASCSU CATB	D	P	A	G	A	S
	CAEEL CPR6	D	P	A	G	A	G
	CONSENSUS	*D*	*P*	*A*	*G*	*A*	*?*
	CAEEL CPR5	Y	P	A	G	A	S
	CAEEL CYS1	Y	P	A	G	A	A
	CAEEL CPR4	Y	P	A	G	A	A
	CAEEL CPR1	Y	P	G	G	A	-
	CAEEL CPR3	Y	S	S	G	A	N
	CONSENSUS	*Y*	*P*	*A*	*G*	*A*	*?*
	ANCCA CP1	W	P	A	G	A	Q
	ANCCA CP2	L	P	A	G	A	Q
	CONSENSUS	*?*	*P*	*A*	*G*	*A*	*Q*
	HAECO CYS1	W	P	S	G	A	T
	HAECO CYS2	W	P	S	G	A	T
	HAECO AC5	W	P	V	G	A	Q
	HAECO PDM4	W	P	A	G	A	N
	HAECO PDM5	W	P	A	G	A	-
	HAECO AC3	W	S	S	G	A	N
	HAECO AC4	W	S	S	G	A	T
	HAECO PDM2	Y	D	A	G	A	-
	OSTOS CYS1	W	P	T	G	A	R
	OSTOS CYS3	-	-	G	G	A	M
	CONSENSUS	*W*	*P/?*	*S/?*	*G*	*A*	*?*
	TRTAE THBG	Y	P	A	G	A	D
	TRTAE CATB	Y	P	A	G	A	D
	NITOB CATB	Y	P	S	G	A	E
	CONSENSUS	*Y*	*P*	*A*	*G*	*A*	*D*
	GIAN CPI3	W	L	A	G	A	Q
	GIAN CPI2	W	L	A	G	A	-
	GIAN CPI1	D	F	M	G	A	E
	CONSENSUS	*W*	*L*	*A*	*G*	*A*	*?*
	STRRA CP1	A	N	A	G	S	-
	AEDEG CATB	Y	L	A	G	A	D
	LEIME CATB	I	P	A	G	A	S
cathepsin c							
	HUMAN CATC	F	P	A	T	A	I
	RAT CATC	F	P	A	T	A	I
	MOUSE CATC	F	P	A	T	A	I
	SCHMA CATC	F	P	G	T	A	L
	SCHJP CATC	F	P	G	T	A	I
	CONSENSUS	*F*	*P*	*A/G*	*T*	*A*	*I*
	ONCVU CATC	K	P	G	I	S	D
	URECA CATB	G	H	G	I	I	E



a) Cysteine proteinases of other organisms

	Gene	67	68	133	157	160	205
Other CP							
	NPVAC CATV	L	L	A	L	A	E
	NPVBM CATV	L	L	A	L	A	E
	NPVCF CATV	L	L	A	L	A	E
baculovirus	*CONSENSUS*	*L*	*L*	*A*	*L*	*A*	*E*
	EURMA EUM1	T	I	I	N	A	-
	DERPT MMAL	T	I	I	N	A	Y
	DERFA MMAL	T	I	I	N	A	Y
mites	*CONSENSUS*	*T*	*I*	*I*	*N*	*A*	*Y*
	PARTE CTL2	Y	N	G	T	Y	Y
	PARTE CTL1	W	M	A	L	G	A
	TETER SGC5	W	P	L	L	A	S
	THEPA CYSP	L	L	Y	L	A	T
	THEAN CYSP	L	P	G	L	A	F
	PLAVN CYSP	N	P	A	L	S	D
	PLAVI CYSP	H	P	N	L	S	E
	PLAOV CYSP	H	P	N	L	S	-
	PLAFR CYSP	H	P	N	L	S	-
	PLACY CYSP	H	P	N	L	S	-
	PLABR CYSP	H	P	N	L	S	-
	PLAFA CYSP	H	P	N	L	S	E
Plasmodium	*CONSENSUS*	*H*	*P*	*N*	*L*	*S*	*E*
	TRIVG CP2	W	P	C	L	A	E
	TRIVG CP1	L	M	A	L	A	A
	TRIVG CP3	D	E	A	L	A	-
	TTMFT CP9	W	P	C	L	A	-
	TTMFT CP2	W	P	N	Y	A	Y
	TTMFT CP5	W	P	N	Y	A	-
	TTMFT CP8	L	M	A	L	G	-
	TTMFT CP7	L	M	A	L	G	-
	TTMFT CP6	L	M	A	L	G	-
	TTMFT CP1	L	M	A	L	G	S
	TTMFT CP4	S	P	C	L	A	-
	TTMFT CP3	S	A	A	A	A	-
Trichomonads	*CONSENSUS*	*?*	*M/P*	*A/N*	*L/Y*	*A/G*	*?*
	ENTHI ACP1	H	P	G	M	C	D
	ENTHI CPP3	L	G	S	L	E	D
	ENTHI CPP2	L	G	S	L	E	D
	ENTHI CPP1	L	G	S	L	E	D
	ENTHI CP6	S	I	A	L	A	D
	ENTHI CP5	S	L	A	L	G	G
	ENTHI CP	S	L	C	V	G	V
	ENTHI CPRO	D	Q	T	L	E	-
Entamoeba	*CONSENSUS*	*L/S*	*G/?*	*?*	*L*	*E/?*	*D*
	DICDI CYS2	L	M	A	L	G	V
	DICDI CP4	L	M	A	L	G	M
	DICDI CP5	L	M	A	L	G	S
Dictiostelium	CONSENSUS	L	M	A	L	G	?

Gene group	67	68	133	157	160	205
papain	Y	p/q	v	v	A	S
vignain	L	M	a	L	G	e
oryzain	I	m	a	I	g	e
cathepsin H	L	P	A	V	A	C
decapod DCP	l/w	m	a	I	g	g/d/e
cathepsin L	L	m	A	I	g	a
cathepsin S	f	M	g	v	g	y
cathepsin K	Y	M	a	v/l	a	I
Cestodes	f/l	M	G	I	G	v/m
Schistosoma	T	M	g/a	l/l	G	N
Fasciola	I	M	a	l/v	a/g	I
Cruzipain	L	M	A	L	g	e/y
Plant Cruzip	I	m	g/a	l/q	G	m
Trematode .cruzip	l/w	p	g	I	a	g/m/v
Entamoeba 1	I	g	s	I	e	d
Entamoeba 2	s	I	a	I	g/a	g/d/v
Triplomonad 1	**I**	**m**	**A**	**I**	**g/a**	**-**
Triplomonad 2	**w**	**p**	**n/c**	**l/y**	**a**	**-**
Plasmodium	h	P	n	I	s	d/e
Giardia	**w**	**I**	**a**	**G**	**A**	**q/e**
trematodes B	f	p	a/s	G	A	d/e
nematodes B1	w	p	s/a	G	A	q/n/t
nematodes B2	y/d	p	a	g	a	a/s/n
Plant cath B	Y	P	a	G	A	d/e
cathepsin B	y/f	P	A	G	A	E
cathepsin C	F	P	a/g	T	A	i
mites	T	I	I	N	A	Y
baculovirus	L	L	A	L	A	E

Table 5.2

Comparison of the consensus S2 subsite residues of the papain superfamily.

Complete conservation is indicated in capital letters, while preferences are indicated in lower case. The amitochondriate taxa (Triplomonads and Diplomonads) are indicated in bold, as they probably represent the oldest members of the family (see section 6). No single consensus exist in triplomonads, but the existing sequences can be divided in two broad groups according to the residues in the S_2 subsite. Similarly, the sequences of *Entamoeba* were subdivided. The nematode cathepsin Bs were subgrouped in: nematodes 1 (sequences from *H. contortus, O. ostertagi* and *A. caninum*), and nematodes 2 (sequences from *C.elegans* and *A. suum*). Cathespin L-like sequences of platyheminths were considered separately (Cestodes, *Schistosoma, Fasciola*).

Although the residue at position 205 (papain numbering) plays an important role in cathepsin B activity, this residue is not very well conserved. It has been shown that this position does not contribute major interactions with the substrate in papain, and that could be the case in all the enzymes that present short side chains in this position. The Gly^{66} residue constitutes part of the left wall of the S_2 subsite, and is absolutely conserved in all the cysteine proteinases, so it was not included in the tables above. Is interesting that position 68 is a Pro in the cathepsins B, cathepsins C, cathepsins H, papain and caricain while a Met is found in the remaining sequences from metazaoans. Position 160 is constrained to small non polar residues (Ala or Gly), while at position 133, although small non polar residues are preferred the constrains seem to be weaker. The non-polar character of the residue at position 157 is well conserved in all the cysteine proteinases of the papain superfamily.

5.2 Generation of *Fasciola hepatica* cathepsin L1 Leu-70-Tyr mutant

Dowd *et al.* (1994) showed that cathepsin L1 and cathepsin L2 from *Fasciola hepatica* have different substrate specificities. Most particularly, cathepsin L2 can cleave substrates with Pro in the P_2 position (e.g. Z-Gly-Pro-Arg-NHMec) whereas these were poorly cleaved by cathepsin L1.

The analysis of the amino acid residues present in the active site cleft of cathepsin L1 and cathepsin L2 showed a couple of interesting variations, that may account for the differences in substrate specificity observed between the enzymes. The residue at position 70 (67 in papain numbering) is a leucine in cathepsin L1, while a tyrosine is present at the equivalent position in cathepsin L2. Furthermore, at position 161 (157 in papain) valine and leucine are present in cathepsin L1 and L2 respectively. The substitution between aliphatic side chains in position 161 is probably less relevant than the substitution of the aliphatic non-polar Leu^{70} present in cathepsin L1 to the aromatic polar Tyr^{70} in the cathepsin L2, where not only the geometry but also the equilibrium of charges at the active site cleft might be modified (Fig 5.1).

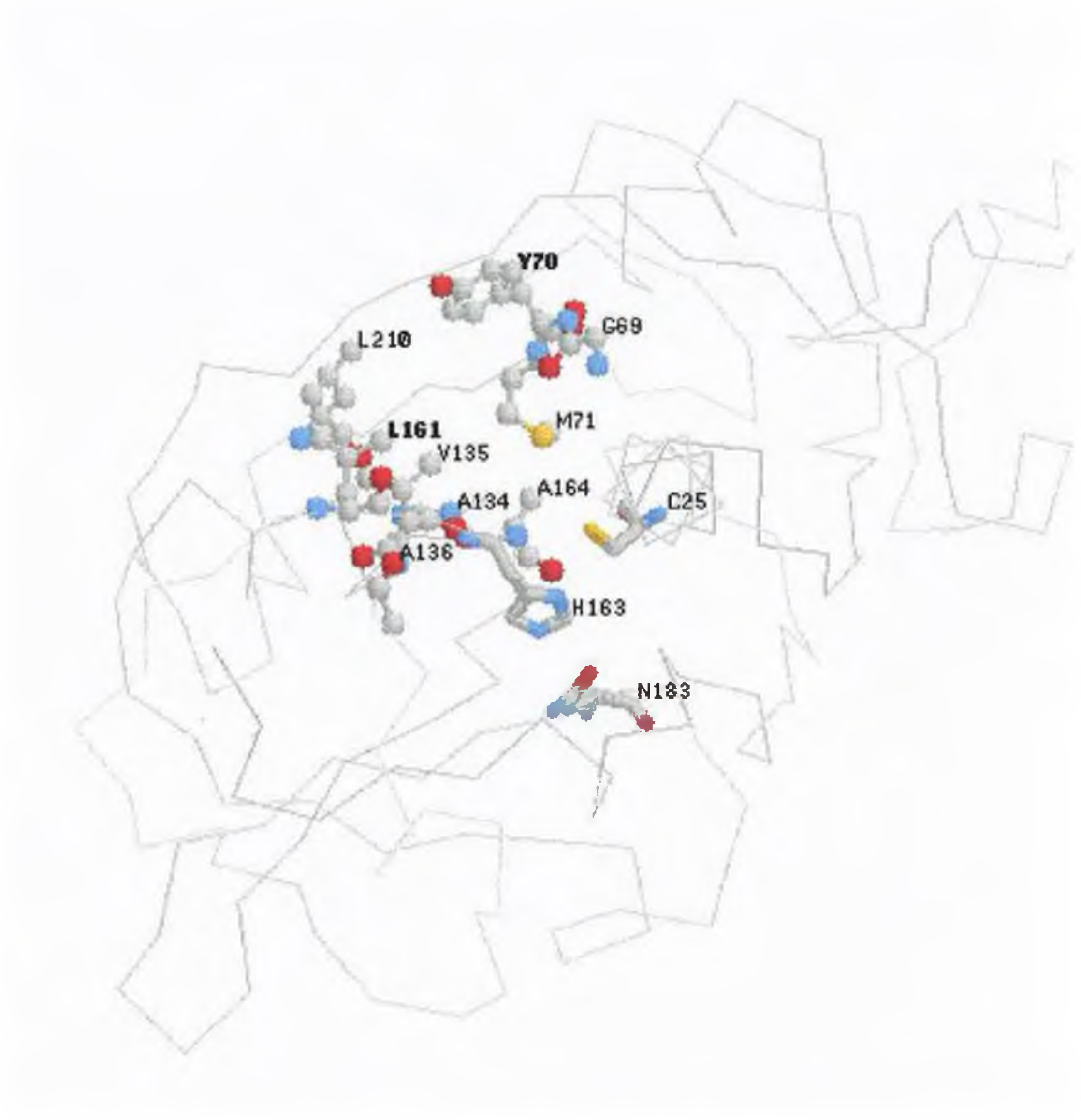


Fig. 5.1
Model of *F.hepatica* cathepsin L2 indicating the residues forming the S_2 binding pocket.

The molecule is shown from the top, with the catalytic triad (represented as sticks) in the center. The residues lining the S_2 binding pocket are represented as ball and sticks. The two residues that differ between *F.hepatica* cathepsin L1 and cathepsin L2 are indicated in bold type.

In order to investigate the role of the residues at position 70 in cathepsins L1 and L2 of *Fasciola hepatica*, we decided to mutate this position in cathepsin L1 to the residue present in cathepsin L2. A PCR based mutagenesis protocol derived from the overlap extension method (Horton and Pease, 1995) was developed. Two overlapping oligonucleotide primers with opposite orientations covering the region centered around residue 70 were designed based on the *Fasciola hepatica* cathepsin L1 sequence, but containing a Tyr codon at position 70. These primers were termed Forward Tyr (FoTyr) and Backward Tyr (BaTyr).

In a first amplification round, each of these primers and an external (vector derived) primer are used to generate two products that comprises the amino and carboxy portion of the procathepsin L1. These products are then combined in a second reaction, denatured and reannealed. As an overlapping region exists, intermolecular byproducts would be formed, some of them non-productive (5'overlapping strands), others (3' overlapping strands) can be used as template for the amplification with the external (vector derived) primers to generate a full length mutated procathepsin L1 product The inclusion of the desired mutation is monitored by restriction enzyme digestion with *Nsi* I, as a restriction site for this enzyme was introduced in the forward mutagenic primer (FoTyr) (Fig 5.2 and Fig.5.3).

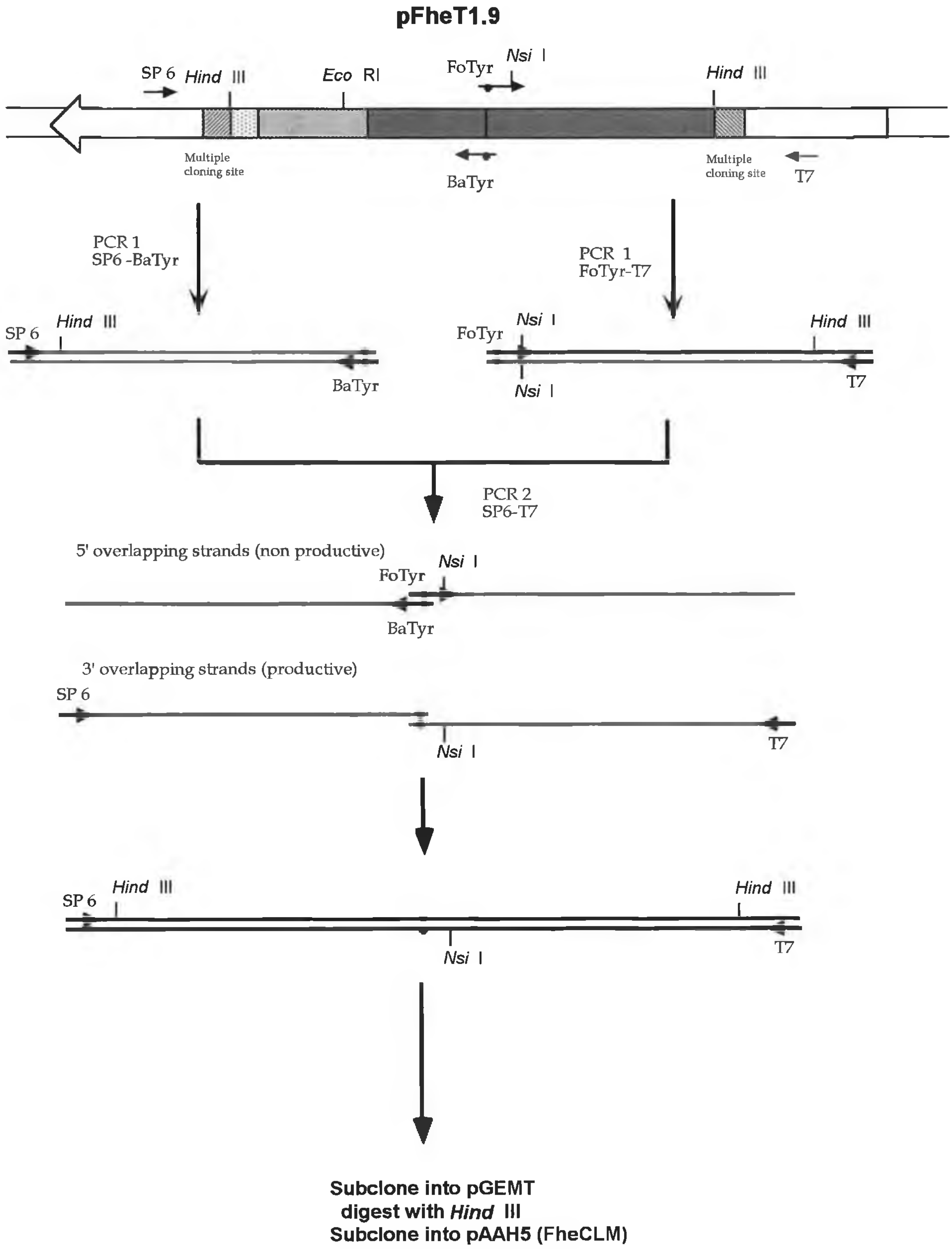


Fig. 5.2

Schematic representation of the strategy for the mutagenesis of FheCL1 and the cloning of the mutant product into the yeast shuttle vector pAAH5.

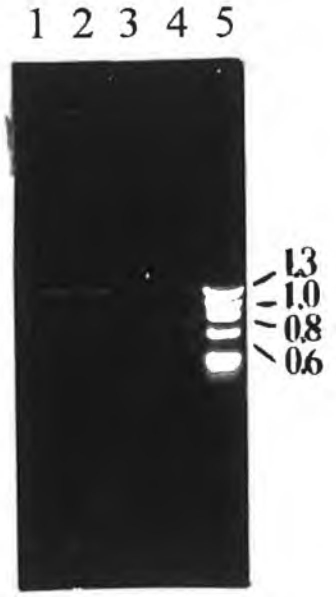


Fig.5.3

TAE agarose gel of PCR products of the secondary aplification for the generation of the *Fasciola hepatica* cathepsin L1 Leu-70-Tyr mutant.

Lanes 1 and 2 undigested product. Lanes 3 and 4 *Nsi* I digested product. Lane 5 molecular weight marker (ϕX174 *Hae* III) The expected size of the bands upon *Nsi* I digestion are 800 and 692 bp.

5.3 Cloning in pAAH5 shuttle expression vector

In order to analyze the biochemical properties of the mutated procathepsin L1 enzyme, the coding sequence carrying the mutation was subcloned into the *Saccharomyces cerevisiae* expression plasmid pAAH5. Although it was theoretically possible to digest the final PCR product with *Hin* dIII and subclone it directly into pAAH5 shuttle vector, the purification of this fragment proved to be difficult, due to the persistent contamination with undigested products (of very similar molecular size to the expected digested product). Consequently, an alternative approach was taken. The final product of the overlap mutagenesis procedure, was gel purified and subcloned into pGEMT vector and the resulting plasmid was digested with *Hin* dIII. A band of the expected size (1Kb) corresponding to the full length insert was detected, purified and ligated to *Hin* dIII linearized phosphorylated pAAH5 plasmid. A plasmid with the insert in the correct orientation was selected (termed pFheCLM) and used to transform yeast DBY746 cells.

The complete sequence of the insert was determined in order to confirm the inclusion of the desired mutation and to check for other changes that might be introduced due to Taq errors during the amplification steps. Analysis of the determined sequence showed that the desired mutation had been achieved. Twelve other differences were detected in the sequence of the mutant product. However, only five of these generated an amino acid change, three of those in the mature enzyme. One of these substitutions (Arg for Ala in position 207) lay close to a residue that might form the bottom of the active site cleft (Leu^{209} in cathepsin L2) (Fig. 5.4).

```
        Hind III
         M   R   L   F   I   L   A   V   L   T   V   G   V   L   G   S   N   D   D   L   W   H
        aag ctt atg cga ttg ttc ata tta gcc gtc ctc aca gtc gga gtg ctt ggc tcg aat gat gat ttg tgg cat
        --- --- --- --- --a --- g-- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
                         V

 Q   W   K   R   M   Y   N   K   E   Y   N   G   A   D   D   Q   H   R   R   N   I   W   E   K   N   V
caa tgg aag cga atg tac aac aaa gaa tac aat ggg gct gac gat cag cac aga cga aat att tgg gaa aag aat gtg
--- --- --a --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

 K   H   I   Q   E   H   N   L   R   H   D   L   G   L   V   T   Y   T   L   G   L   N   Q   F   T   D
aaa cat att caa gaa cat aac cta cgt cac gat ctc ggc ctc gtc acc tac aca ttg gga ttg aac caa ttc acg gat
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --g --- ---
                                                                                             L

 M   T   F   E   E   F   K   A   K   Y   L   T   E   M   S   R   A   S   D   I   L   S   H   G   V   P
atg aca ttc gag gaa ttc aag gcc aaa tat cta aca gaa atg tca cgc gcg tcc gat ata ctc tca cac ggt gtc ccg
--- --- --- --- --- --- --- --a --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

 Y   E   A   N   N   R   A   V   P   D   K   I   D   W   R   E   S   G   Y   V   T   E   V   K   D   Q
tat gag gcg aac aat cgt gcc gta ccc gac aaa att gac tgg cgt gaa tct ggt tat gtg acg gag gtg aaa gat cag
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

 G   N   C   G   S   C   W   A   F   S   T   T   G   T   M   E   G   Q   Y   M   K   N   E   R   T   S
gga aac tgt ggc tcc tgt tgg gca ttc tca aca acc ggt act atg gag gga cag tat atg aaa aac gaa aga act agt
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

                                                                                FoTyr
                                                                        ggt tgc ggt ggt gga tAT atg gaa
 I   S   F   S   E   Q   Q   L   V   D   C   S   R   P   W   G   N   N   G   C   G   G   G   L   M   E
att tca ttc tct gag caa caa ctg gtc gat tgt agc cgt cct tgg gga aat aat ggt tgc ggt ggt gga ttg atg gaa
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -at --- ---
                                                                                             Y
                                                                            tgc ggt ggt gga tAT atg gaa
                                                                                 BaTyr
 Nsi I
aat gcA tac caa t
 N   A   Y   Q   Y   L   K   Q   F   G   L   E   T   E   S   S   Y   P   Y   T   A   V   E   G   Q   C
aat gct tac caa tat ttg aaa caa ttt gga ttg gaa acc gaa tcc tct tat ccg tac acg gct gtg gaa ggt cag tgt
--- --a --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---
aat

 R   Y   N   K   Q   L   G   V   A   K   V   T   G   F   Y   T   V   H   S   G   S   E   V   E   L   K
cga tac aat aaa cag tta gga gtt gcc aaa gtg act ggc ttc tat act gtt cat tct ggc agt gag gta gaa ttg aaa
--- --- --- --g --- --- --- --- --- --- --- --- --- --t --- --- --g --- --- --- --- --- --- --- --- ---

 N   L   V   G   S   E   G   P   S   A   V   A   V   D   V   E   S   D   F   M   M   Y   R   S   G   I
aat cta gtc ggt tcc gaa gga cct tcc gcg gtc gct gtg gat gtg gaa tct gac ttc atg atg tac agg agt ggt att
--- --- --- --- g-- --- --- --- g-- --- --- --- --- --- --- --- --- --- --t --- --- --- --- --- --- ---
                 A               A

 Y   Q   S   Q   T   C   S   P   L   R   V   N   H   A   V   L   A   V   G   Y   G   T   Q   G   G   T
tat cag agc caa act tgt tca ccg ctt cgt gtg aac cat gca gtc ttg gct gtc ggt tat gga aca cag ggt ggt act
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

 D   Y   W   I   V   K   N   S   W   G   L   S   W   G   E   R   G   Y   I   R   M   V   R   N   R   G
gac tat tgg att gtg aaa aat agt tgg gga ttg tcg tgg ggt gag cgc ggt tac att cga atg gtt agg aac cga ggt
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- ---

 N   M   C   G   I   A   S   L   A   S   L   P   M   V   A   R   F   P   *   *
aac atg tgt gga att gct tcg ctg gcc agt ctc ccg atg gtg gca cga ttt ccg tga taa gct t
--- --- --- --- --- cg- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -
                     R

                                                                                Hind III
```

Fig. 5.4

Sequence of cathepsin L1 and the mutant product.

Substituted nucleotide bases are indicated (identical residues indicated by a dash beneath the cathepsin L1 sequence). The position of the primers used is indicated (arrows). The residues of the catalytic triad are encircled, and the residues suspected to form the S2 subsite are boxed. The desired mutation is indicated (rectangle).

5.4 Functional Expression of the *Fasciola hepatica* cathepsin L1 Leu-70-Tyr mutant

Supernatant of the culture of yeast DBY746 cells transformed with the plasmid pFheCLM was collected and subjected to analysis by SDS-PAGE, and Western blots. No significant differences could be detected between the mutant product and the recombinant yeast produced cathepsin L1. This indicates that the variations introduced did not modified the sorting of the recombinant enzyme in the yeast expression system (Fig 5.5).

The CS was assayed for cysteine proteinase activity with synthetic fluorogenic peptide substrates. Efficient cleaving activity towards Z-Phe-Arg-NHMec, Z-Val-Leu-Lys-NHMec and Z-Phe-Val-Arg-NHMec, was observed (Fig.5.6). The activity towards the substrates Z-Val-Pro-Arg-NHMec and Z-Gly-Pro-Arg-NHMec was low, similar to the activity of the cultures expressing the recombinant cathepsin L1. The activity towards all the substrates was enhanced by the addition of the reducing agent DTT. These results indicate that, although substitutions have been introduced in the cathepsin L1 gene, active enzyme is produced in yeast transformed with the plasmid pFheCLM. The substitution Leu-70-Tyr, seems to have no effect on the ability of cathepsin L1 to cleave substrates with proline in the S_2 position.

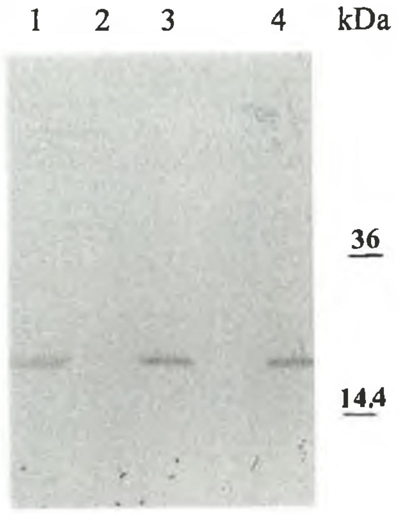


Fig 5.5
Immunoblot analysis of culture supernatants of yeast DBY 486 cells transformed with pFheYCL1, pFheYCL2 and PfheYCLM .
Culture supernatants (lanes 1 2 and 3 respectively), and purified *F.hepatica* native cathepsin L1 (lane 4) detected with anti-cathepsin L1 serum .

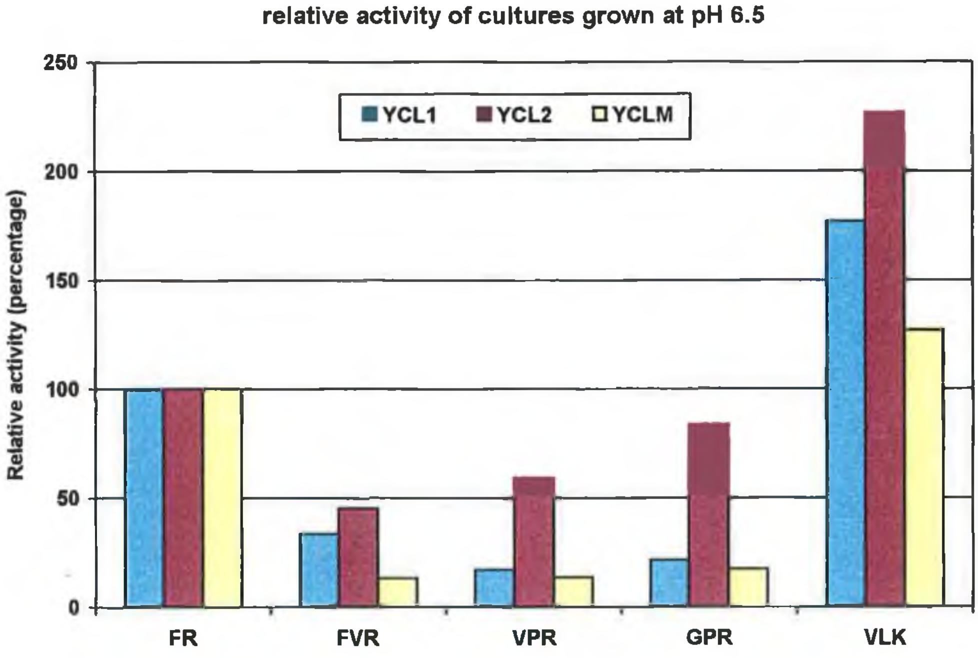


Fig 5.6
Relative activity of culture supernatants of yeast DBY746 cells transformed with pFheYCL1, pFheYCL2, and pFheYCLM (prepared under identical conditions) towards different synthetic substrates in the presence of DTT. Substrate abbreviations as in figure 4.3.

5.5 Purification of the recombinant Cathepsin L1 Leu-70-Tyr mutant

In order to have a better understanding of the kinetic properties of the mutant enzyme, we have to rule out the existence of any contaminating activity. To achieve this goal, a purification scheme was developed.

Recombinant cathepsin L1 Leu-70-Tyr mutant (RCLM) was purified from concentrated yeast culture supernatants by gel filtration on Sephadex S200HR. Fractions were assayed for cathepsin L activity using the fluorogenic substrate Z-Phe-Arg-NHMec. The results of Z-Phe-Arg-NHMec cleaving assays of the fractions eluted from the S200HR column show that there are 2 peaks of Z-Phe-Arg-NHMec cleaving activity (peaks I, and II) (fig. 5.7.A). The activities of both peaks were enhanced by the reducing agent cysteine. The first peak of activity (peak I) eluted with a minor protein peak that appears immediately after the major protein peak, and the second peak (peak II) eluted with a minor protein peak that appeared further down in the elution.

The fractions of peaks I and II were separately pooled and aliquots were subsequently analyzed by immunoblotting and SDS-PAGE under reducing and denaturing conditions (Fig. 5.7.B and 5.7.C). Pool I contained several proteins bands, with two major bands of approximately 29 kDa, that are recognized by anti-cathepsin L1 serum. Most of the other bands present were of lower molecular weight. Pool II consisted of two proteins of 28 and 29.5 kDa, both reactive with the anti-cathepsin L1 serum. The bands comigrate with native cathepsin L1 purified from *F.hepatica* Es products. A minor band of approximate molecular weight 37 kDa. was visualized in both pool samples. This band was recognized by the anti-cathepsin L1 serum, and corresponded well with the expected size of the procathepsin L.

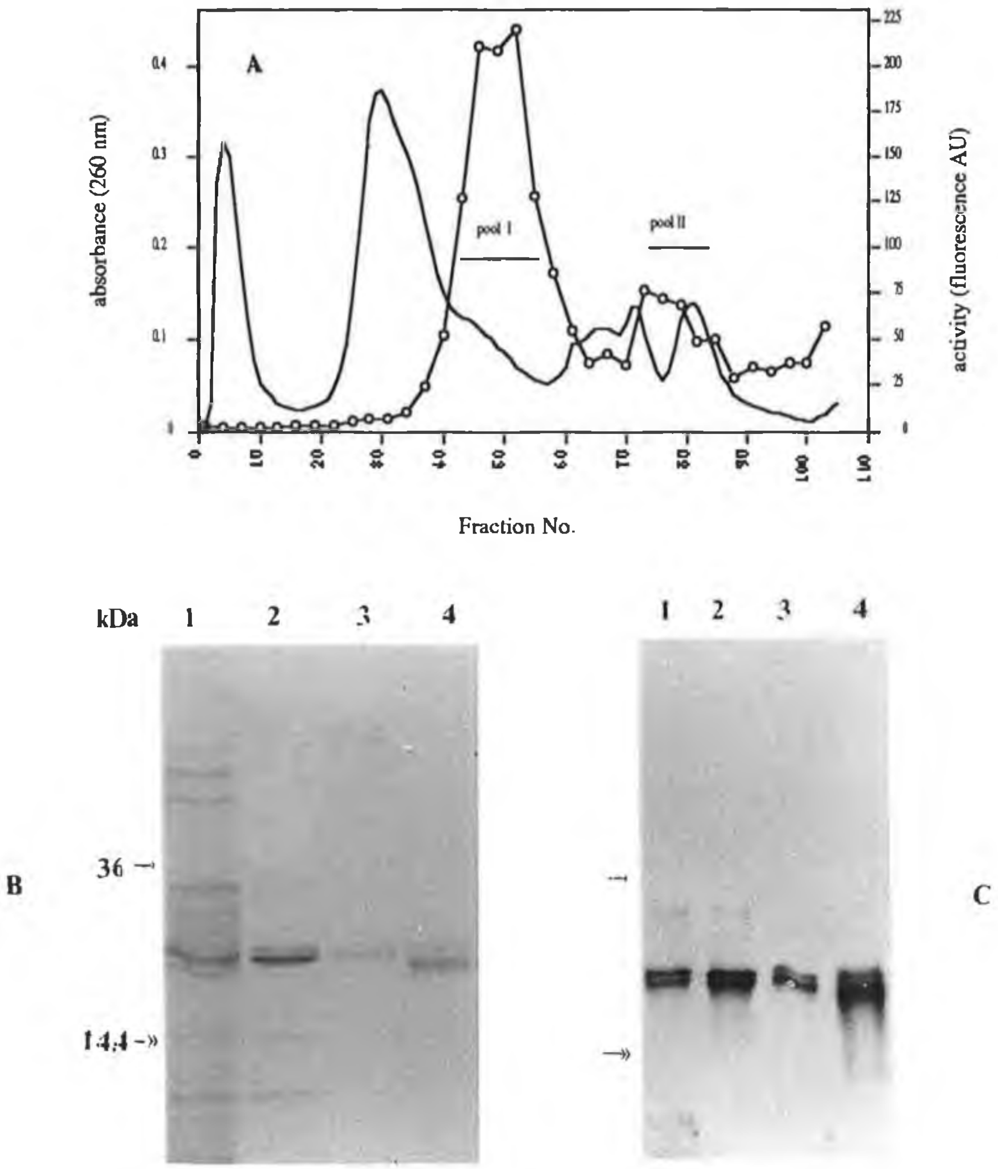


Fig. 5.7

Purification of recombinant cathepsin L1.Leu 70 Tyr mutant.

(A), profile of protein elution from a Sephacryl S200HR column monitored by absorbance at 280 nm (lines) and cysteine proteinase activity in collected fractions assayed using the fluorogenic substrate Z-Phe-Arg-NHMec (circles). The fractions pooled are indicated.

SDS-PAGE under reducing conditions (B) and immunoblot (C) analysis of culture supernatant of yeast cells transformed with pFheYCLM (lane 1), pools I and II (lanes 2 and 3), and purified native cathepsin L1 (lane 4). The position of the markers triose-phospate isomerase (36 kDa) and lysozyme (14.4 kDa) are indicated by arrows.

5.5 Substrate specificity of the recombinant cathepsin L1 Leu-70-Tyr mutant

In order to confirm the data obtained previously a continuous recording of the hydrolysis of different synthetic substrates was performed. The substrate concentration was kept far below the estimated K_m. In these conditions, and with a low concentration of enzyme (to allow the substrate hydrolysis to be recorded), a first-order curve for the product generation is obtained (Fig. 5.8).

The mutant product can cleave the substrates Z-Phe-Arg_NHMec, and Z-Val-Leu-Lys-NHMec, as efficiently as the non mutated recombinant enzyme. However, none of the enzymes were capable of cleaving efficiently the substrates Tos-Gly-Pro-Arg-NHMec or Boc-Val-Pro-Arg-NHMec, indicating that the Tyr residue at position 70 is not responsible for the preference for proline in the P_2 position.

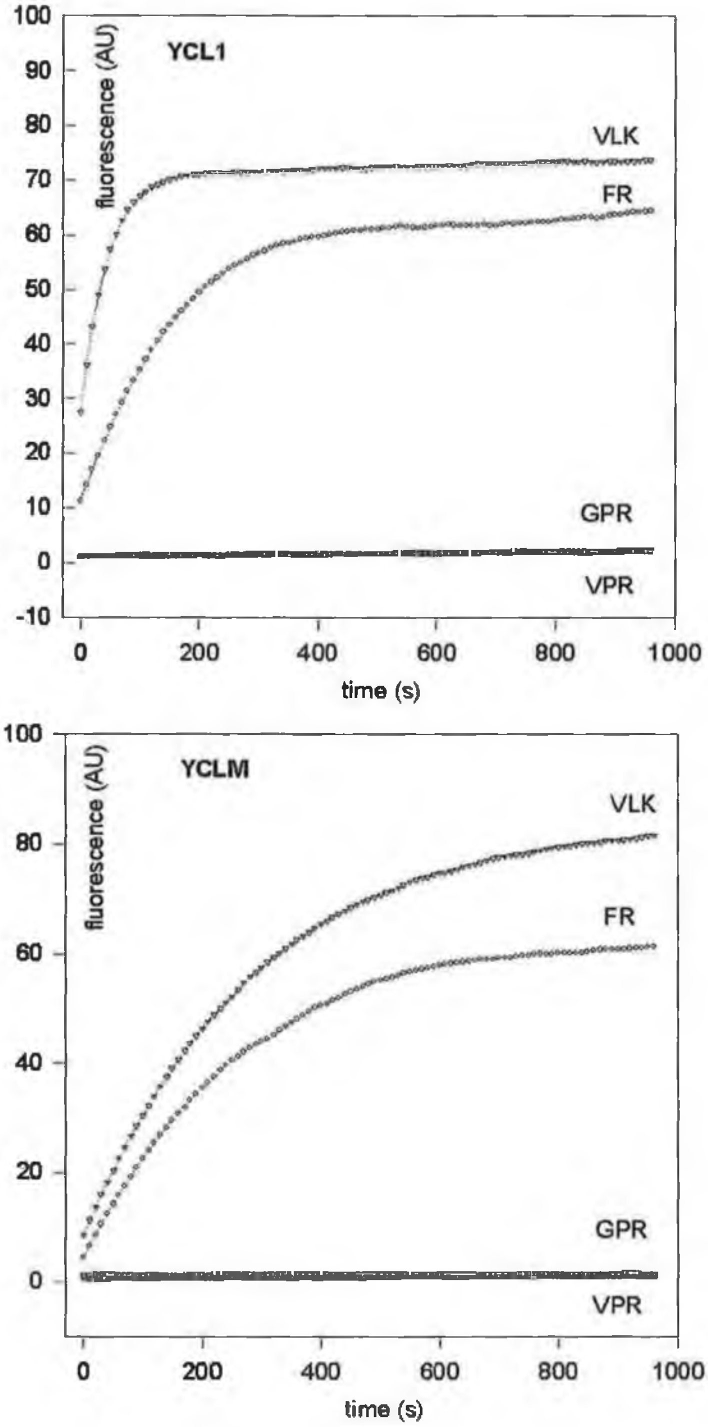


Fig.5.8

Hydrolysis of synthetic substrates by purified recombinant *F.hepatica* cathepsin L1 (A) and purified recombinant *F.hepatica* cathepsin L Leu-70-Tyr mutant (B). The substrate concentration was 2 uM, a value much lower than the K_m.

6 Phylogeny of the papain-like cysteine proteinases

6.1 Alignment of sequences

In order to understand the relationships of the trematode cysteine proteinase genes with other members of the papain superfamily, an alignment of the sequences of cysteine proteinases of this superfamily obtainable in the public databases was made. Hughes (1994) analyzed a restricted set of sequences and provided evidence for an early divergence of cathepsin B-like enzymes. Later Berti and Storer (1995) carried out a similar study, and were able to recognize three different classes in the non cathepsin-B complex. The cathepsin B-like genes of *Schistosoma* were the only platyhelminth sequences considered in these studies. In the last couple of years, many new genes have been sequenced, several of them from parasitic organisms, duplicating the amount of available sequences analized by Berti and Storer (1995). Furthermore, cysteine proteinases of amitochondriate protists have been isolated, opening new avenues for the understanding of the evolutionary relationships of members of the papain superfamily.

A progressive alignment method using the program ClustalW was used for sequence alignment. The alignment procedure took into account the available data on secondary structure of several members of the family, which was used to produce a profile, assigning higher gap penalties to discontinuities in the alignment that interrupt the secondary structure features. The total sequences, the regions corresponding to the mature enzyme and the propeptides were aligned separately. Alignments were performed with or without the secondary structure profiles, and the effects of different sequences in the topology of the resulting trees were analyzed by adding sequences or profiles of aligned sequences to previously aligned sets.

Phylogentic trees were calculated using the Neighbor joining method of Saitou and Nei (1987) based on a matrix of "distances" between all sequences. The distances were corrected using the Kimura formula for correcting the distance data for multiple substitutions (Kimura, 1983). This correction stretches distances (especially large ones) to try to correct for the fact that observed distances (mean number of differences per site) greatly underestimate the actual number that happened during evolution. Confidence limits on the resulting trees were calculated by bootstrapping. Different sets of sequences may give different topologies (branching orders) for parts of a tree that are weakly supported by the data, so it is useful to have an indication of the degree of error in the tree. The Bootstrapping method developed by Felsenstein (1985) is a general approach that can be applied to any tree drawing method. The method works by taking random samples of positions from the alignment. If the alignment has N positions, each bootstrap sample consists of a random sample of N positions, taken with replacement (in any given sample, some sites may be sampled several times, others not at all). Then, with each sample of sites, a distance matrix is calculated and a tree is drawn. For each grouping on the tree, the number of times this grouping occurs in the sample trees is recorded. This give an estimate of the support for the grouping, given the data set and the method used to draw the tree. Subsets of optimized alignments were also analyzed by parsimony methods using the program PROTPARS from the Phylip package. This method infers an unrooted phylogeny, based on the maximum parsimony principle, and considering the genetic code. The principle of maximum parsimony involves the identification of a tree that requires the smallest number of evolutionary changes to explain the differences observed among the elements being analyzed. In this sense, as the genetic code is considered, not all substitutions are allowed, and some substitutions are counted as two or more changes (for example, Phe can change to Lys, but not to Gln, but the sequence of changes Phe→ Lys→ Gln is possible and it is counted as two changes). In this case there is also one synonymous change involved [Phe (UUY) to either Leu (UUR) / Leu (CUY), then Leu (UUR) / Leu (CUY) to Leu (CUR) and finally Leu (CUR) to Gln (CAR)]. However, these silent substitutions are not counted based on the assumption that these changes are faster and more common that non-synonymous substitutions.

The prokaryotic aminopeptidases C of *Lactobacillus* and *Streptococcus*, although included in the papain superfamily, and even in the C1 family (which includes the plant enzyme papain, that gives name to the group) showed restricted similarity (only in the region surrounding the active site residues), and hence were considered no further. The group of the bleomycin hydrolases is in a similar situation, being well conserved at only those residues in the immediate vicinity of the Cys, His and Asp residues of the active site, and some hydrophobic regions. Due to the ample divergence in the rest of the sequences, this group was not further analyzed. These two groups of enzymes could probably be better considered as two different families belonging to the CA clan, as the cytoplasmic calcium dependent proteases (calpains, family C2) already are.

The remainder of the sequences analyzed belong to the C1 family (papain superfamily) of cysteine proteinases, namely the papain group (Barret and Rawlings, 1996). Several groups can be recognized after the alignment and construction of phylogenetic gene trees (some of them with very high bootstrap probability) as reliable families. In this situation are the cathepsins B, C, L, H, K and S, the cathepsin-like enzymes of decapods, some of the enzymes from platyhelminths, the cysteine proteinases from kinetoplastids, the *Entamoeba* cysteine proteinases, and two groups of sequences from plant genes. Some deeper branches also received good bootstrap values, so it was possible to group families into classes (Fig 6.1).

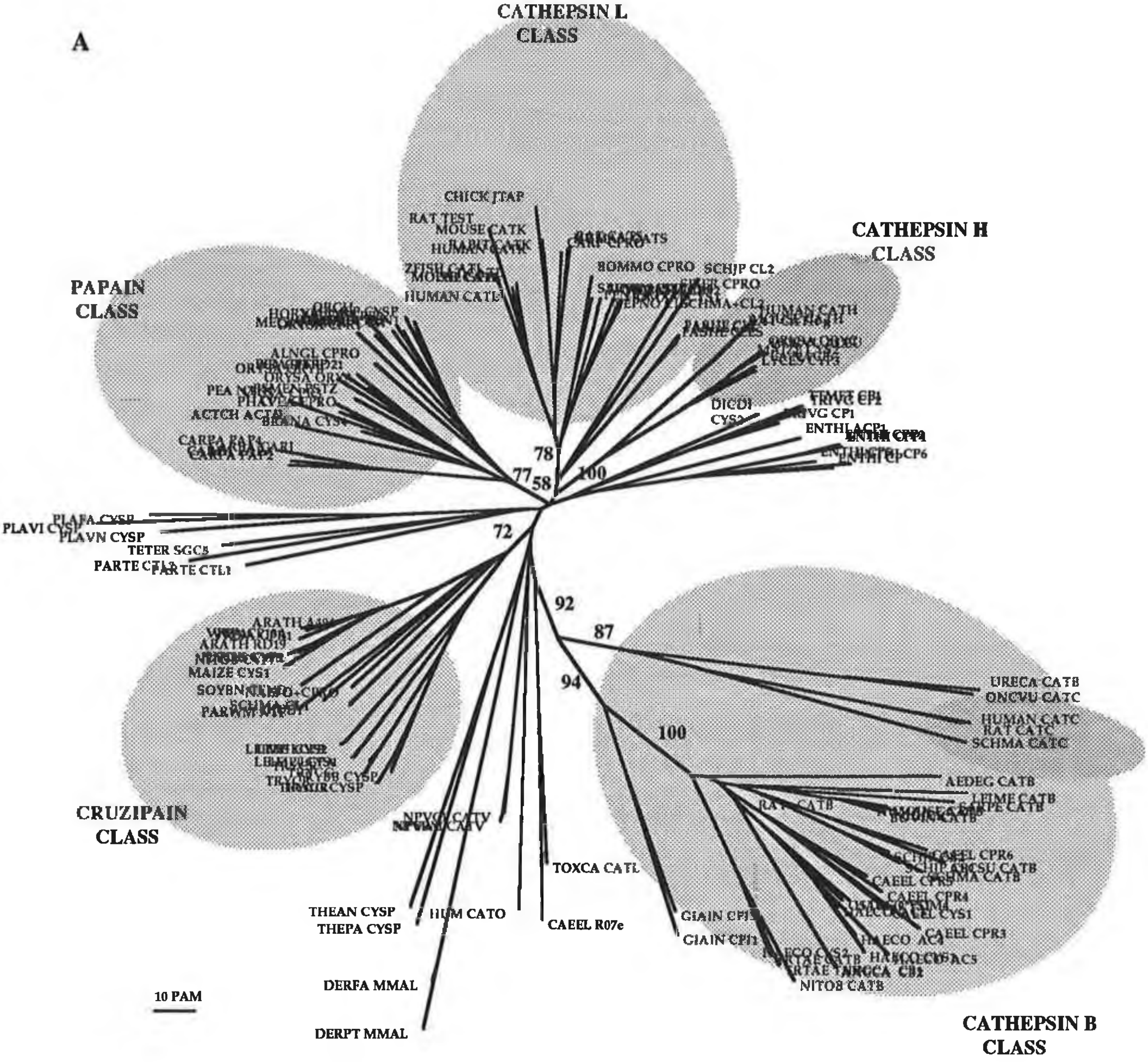
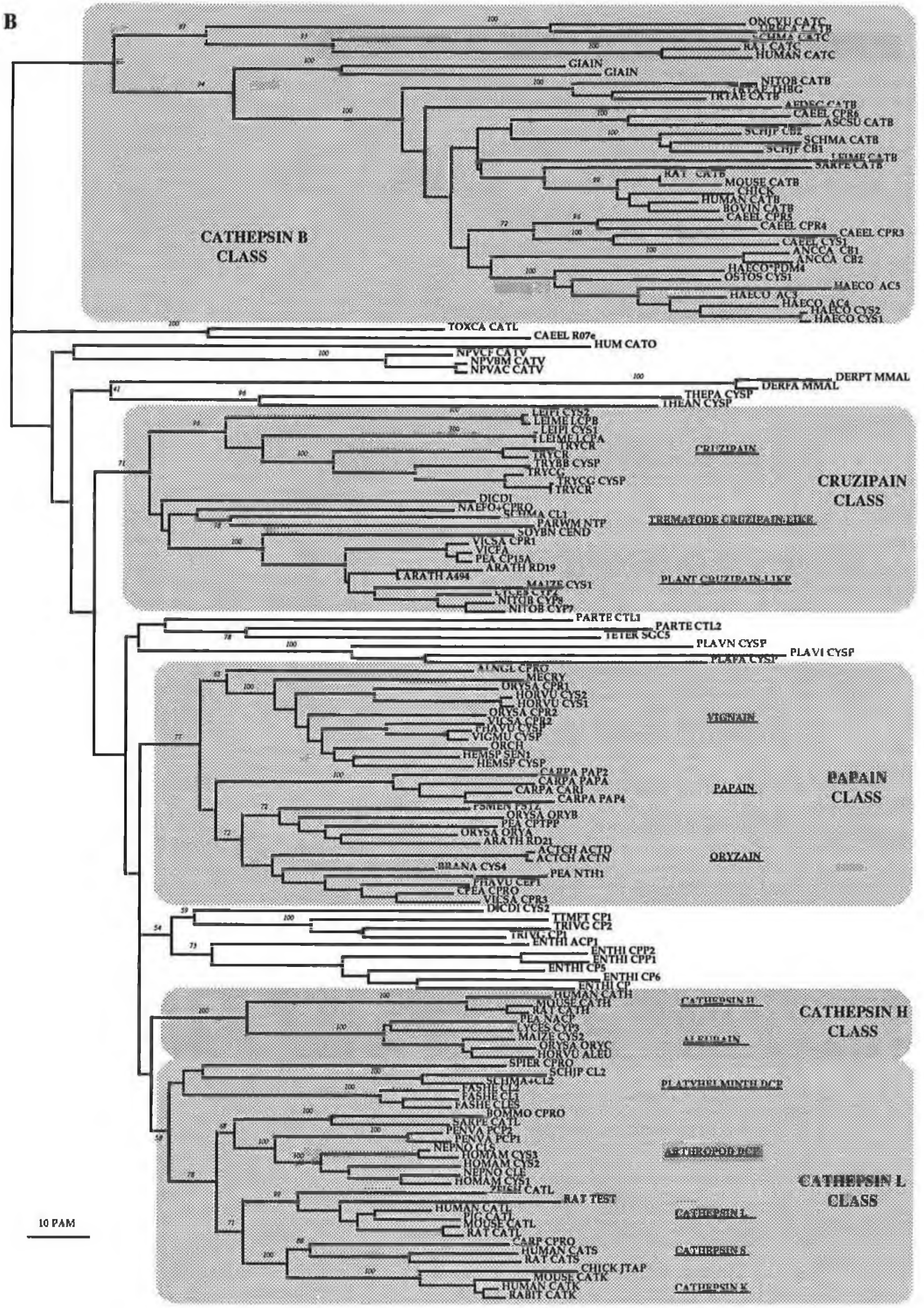


Fig 6.1
A) Neighbor joining unrooted tree of complete coding sequences of cysteine proteinases of the papain superfamily. B) Rooted representation of the same tree.
Different classes can be recognized and are indicated by shaded areas.
Bootstrap values of the main branches are indicated.


B
CATHEPSIN B CLASS
ONCVU CATC
DREFA CATB
SCHMA CATC
RAT CATC
HUMAN CATC
GIAIN
GIAIN
NITOB CATB
TRTAE THBG
TRTAE CATB
AEDEC CATB
CAEEL CPR6
ASCSU CATB
SCHJP CB2
SCHMA CATB
SCHJP CB1
LEIME CATB
SARPE CATB
RAT CATB
MOUSE CATB
CHICK
HUMAN CATB
BOVIN CATB
CAEEL CPR5
CAEEL CPR4
CAEEL CPR3
CAEEL CYS1
ANCCA CB1
ANCCA CB2
HAECO PDM4
OSTOS CYS1
HAECO AC5
HAECO AC3
HAECO AC4
HAECO CYS2
HAECO CYS1
TOXCA CATL
CAEEL R07e
HUM CATO
NPVCF CATV
NPVBM CATV
NPVAC CATV
DERPT MMAL
DERFA MMAL
THEPA CYSP
THEAN CYSP
LEIPI CYS2
LEIME LCPB
LEIPI CYS1
LEIME LCPA
TRYCR
TRYCR
TRYBB CYSP
TRYCG
TRYCG CYSP
TRYCR
CRUZIPAIN
CRUZIPAIN CLASS
DICDI
NAEFO CPRO
SCHMA CL1
PARWM NTP
SOYBN CEND
TREMATODE CRUZIPAIN-LIKE
VICSA CPR1
VICFA
PEA CP15A
ARATH RD19
ARATH A494
PLANT CRUZIPAIN-LIKE
MAIZE CYS1
LYCES CYP2
NITOB CYP8
NITOB CYP7
PARTE CTL1
PARTE CTL2
TETER SGC5
PLAVN CYSP
PLAVI CYSP
PLAFA CYSP
AENGL CPRO
MECRY
ORYSA CPR1
HORVU CYS2
HORVU CYS1
ORYSA CPR2
VICSA CPR2
PHAVU CYSP
VIGMU CYSP
VIGNAIN
ORCH
HEMSP SEN1
HEMSP CYSP
CARPA PAP2
CARPA PAPA
CARPA CARI
CARPA PAP4
PAPAIN
PAPAIN CLASS
PSMEN PST2
ORYSA ORYB
PEA CPTPP
ORYSA ORYA
ARATH RD21
ACTCH ACTD
ACTCH ACTN
ORYZAIN
BRANA CYS4
PEA NTH1
PHAVU CEP1
CPEA CPRO
VICSA CPR3
DICDI CYS2
TTMFT CP1
TRIVG CP2
TRIVG CP1
ENTHI ACP1
ENTHI CPP2
ENTHI CPP1
ENTHI CP5
ENTHI CP6
ENTHI CP
HUMAN CATH
MOUSE CATH
RAT CATH
CATHEPSIN H
CATHEPSIN H CLASS
PEA NACP
LYCES CYP3
MAIZE CYS2
ALEURAIN
ORYSA ORYC
HORVU ALEU
SPIER CPRO
SCHJP CL2
SCHMA CL2
PLATYHELMINTH DCP
FASHE CL2
FASHE CL1
FASHE CLES
BOMMO CPRO
SARPE CATL
PENVA PCP2
PENVA PCP1
NEPNO CLS
HOMAM CYS3
HOMAM CYS2
ARTHROPOD DCP
NEPNO CLE
HOMAM CYS1
ZEISH CATL
CATHEPSIN L CLASS
RAT TEST
HUMAN CATL
PIG CATL
MOUSE CATL
RAT CATL
CATHEPSIN L
CARP CPRO
HUMAN CATS
RAT CATS
CATHEPSIN S
CHICK JTAP
MOUSE CATK
HUMAN CATK
RABIT CATK
CATHEPSIN K
10 PAM

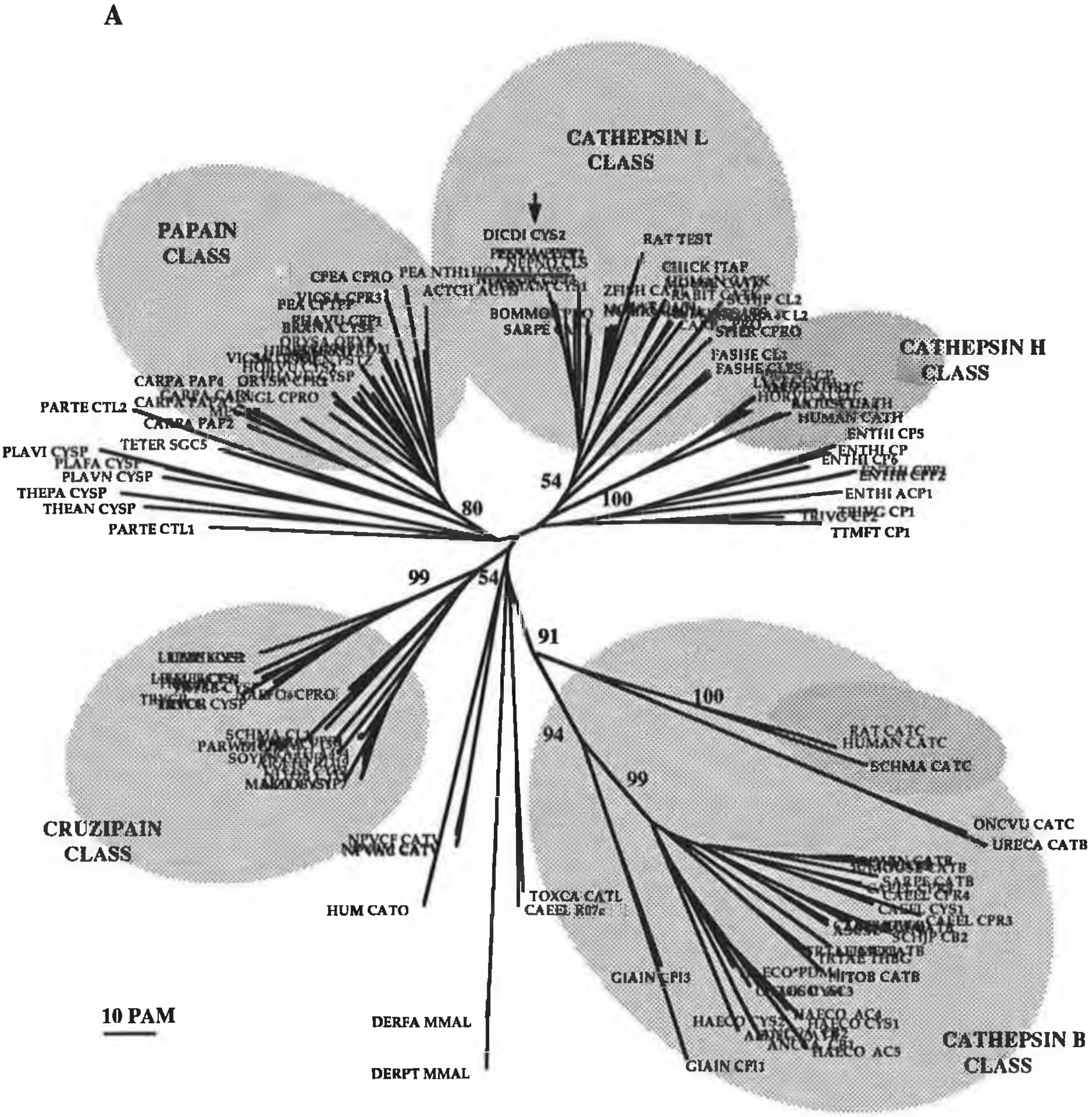
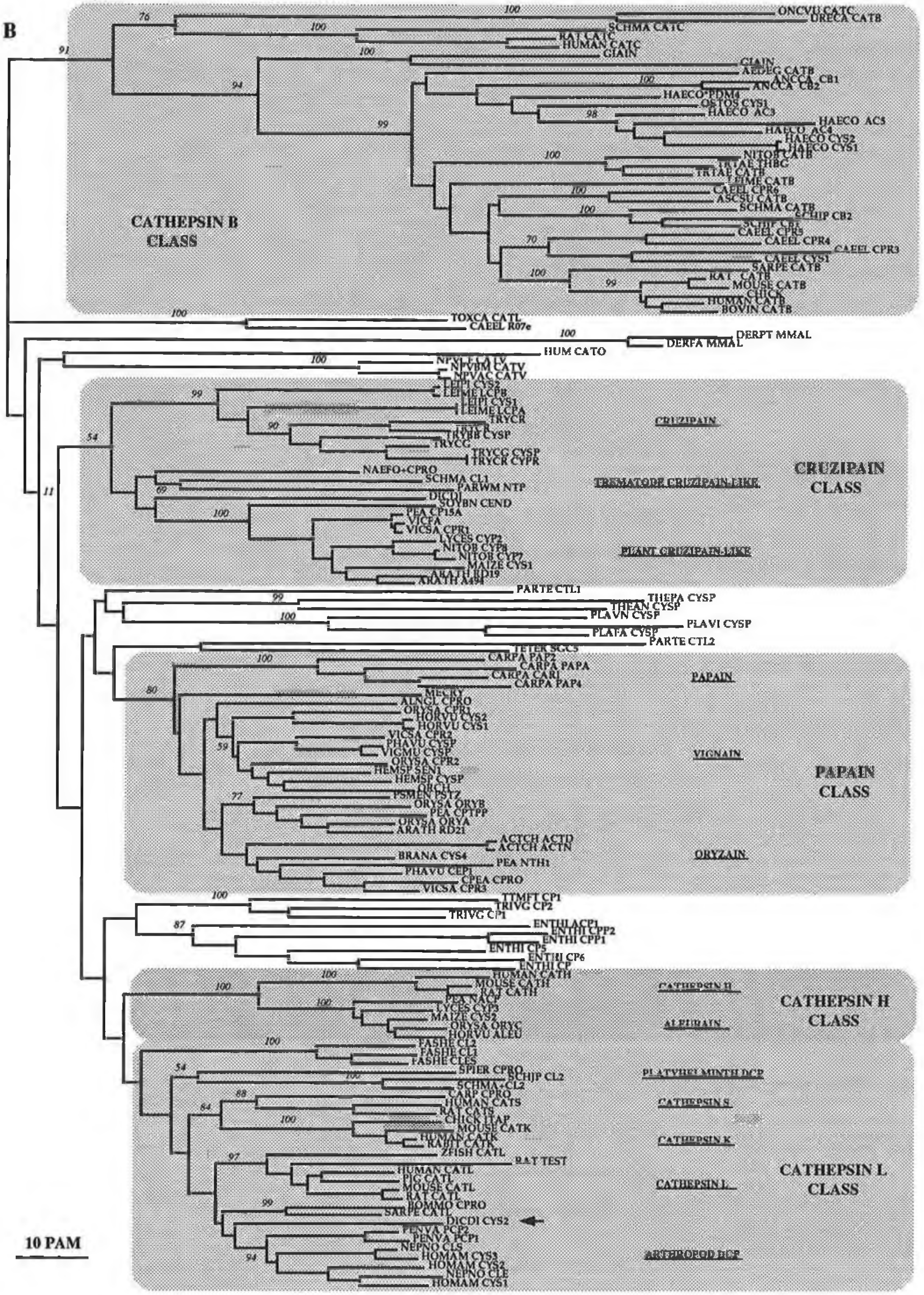


Fig 6.2

A) Neighbor joining unrooted tree of mature regions of cysteine proteinases of the papain superfamily. B) Rooted representation of the same tree. Different classes can be recognized and are indicated by shaded areas. Bootstrap values of the main branches are indicated.


B
CATHEPSIN B CLASS
91
76
100
ONCVU CATC
URECA CATB
100
SCHMA CATC
RAT CATC
HUMAN CATC
100
GIAIN
94
GIAIN
AEDEG CATB
100
ANCCA CB1
ANCCA CB2
HAECO PDM4
OSTOS CYS1
98
HAECO AC3
99
HAECO AC5
HAECO AC4
HAECO CYS2
HAECO CYS1
100
NITOB CATB
TRTAE THBG
TRTAE CATB
LEIME CATB
100
CAEEL CPR6
ASCSU CATB
100
SCHMA CATB
SCHJP CB2
SCHJP CB1
CAEEL CPR5
70
CAEEL CPR4
CAEEL CPR3
CAEEL CYS1
100
SARPE CATB
99
RAT CATB
MOUSE CATB
CHICK
HUMAN CATB
BOVIN CATB
100
TOXCA CATL
CAEEL R07e
100
DERPT MMAL
DERFA MMAL
HUM CATO
100
NPVLF CATV
NPVBM CATV
NPVAC CATV
11
54
99
LEIPI CYS2
LEIME LCPB
LEIPI CYS1
LEIME LCPA
TRYCR
90
TRYCR
TRYBB CYSP
TRYCG
TRYCG CYSP
TRYCR CYPR
NAEFO+CPRO
SCHMA CL1
69
PARWM NTP
DICDI
SOYBN CEND
100
PEA CP15A
VICFA
VICSA CPR1
LYCES CYP2
NITOB CYP8
NITOB CYP7
MAIZE CYS1
ARATH RD19
ARATH A494
CRUZIPAIN
TREMATODE CRUZIPAIN-LIKE
PLANT CRUZIPAIN-LIKE
CRUZIPAIN CLASS
PARTE CTL1
99
THEPA CYSP
THEAN CYSP
100
PLAVN CYSP
PLAVI CYSP
PLAFA CYSP
PARTE CTL2
TETER SGC5
100
80
CARPA PAP2
CARPA PAPA
CARPA CARI
CARPA PAP4
MECRY
ALNGL CPRO
ORYSA CPR1
HORVU CYS2
HORVU CYS1
VICSA CPR2
59
PHAVU CYSP
VIGMU CYSP
ORYSA CPR2
HEMSP SEN1
HEMSP CYSP
ORCH
77
PSMEN PST7
ORYSA ORYB
PEA CPTPP
ORYSA ORYA
ARATH RD21
ACTCH ACTD
ACTCH ACTN
BRANA CYS4
PEA NTH1
PHAVU CEP1
CPEA CPRO
VICSA CPR3
PAPAIN
VIGNAIN
ORYZAIN
PAPAIN CLASS
100
TTMFT CP1
TRIVG CP2
TRIVG CP1
87
ENTHI ACP1
ENTHI CPP2
ENTHI CPP1
ENTHI CP5
ENTHI CP6
ENTHI CP
100
100
HUMAN CATH
MOUSE CATH
RAT CATH
100
PEA NACP
LYCES CYP3
MAIZE CYS2
ORYSA ORYC
HORVU ALEU
CATHEPSIN H
ALEURAIN
CATHEPSIN H CLASS
100
FASHE CL2
FASHE CL1
FASHE CLES
54
SPIER CPRO
100
SCHJP CL2
SCHMA+CL2
88
84
CARP CPRO
HUMAN CATS
RAT CATS
100
CHICK ITAP
MOUSE CATK
HUMAN CATK
RABIT CATK
ZFISH CATL
97
RAT TEST
HUMAN CATL
PIG CATL
MOUSE CATL
RAT CATL
99
BOMMO CPRO
SARPE CATL
DICDI CYS2
PENVA PCP2
PENVA PCP1
NEPNO CLS
94
HOMAM CYS3
HOMAM CYS2
NEPNO CLE
HOMAM CYS1
PLATYHELMINTH DCP
CATHEPSIN S
CATHEPSIN K
CATHEPSIN L
ARTHROPOD DCP
CATHEPSIN L CLASS
10 PAM

6.2 The cathepsin B class

The cathepsin B enzymes of vertebrates can be reliably clustered with similar enzymes from invertebrates and with some plant relatives of this group, since they obtained a bootstrap probability of 100% (Figs. 6.1 and 6.2). The branching order varied only at the terminal nodes of the tree when they were constructed using full length coding sequences, only mature enzymes, or only conserved regions, but the general topology was maintained. The deeper branch in the cathepsin B genes separates two sequences from the diplomonad *Giardia muris* (GIAN_CP1 and GIAN_CP3) from the rest of the cathepsin B genes, with very good bootstrap values. An additional partial sequence also exists from this organism (not included in the tree), that clustered well with the two full length sequences. The diplomonads are probably one of the earliest eukaryotic lineages, as stated by the absence of mitochondria in this parasitic protists. The presence of cysteine proteinases of the papain superfamily in these organisms points to an origin of the papain superfamily very early during eukaryote evolution, probably just after the eukaryote/prokaryote divergence (Ward *et al.*, 1997).

A second group is the cluster is formed by the cathepsin B-like enzymes from tobacco and wheat (NITOB_CATB, TRTAE_CATB, TRTAE_THBG) (Fig 6.1, 6.2 and 6.3). The position of this group, although supported by good bootstrap values, was not considered reliable due to the existence of a cathepsin B-like gene from the parasitic kinetoplastid *Leishmania mexicana* (LEIME_CATB) further down in the cathepsin B clade. It is expected that the enzyme from the protist would separate from the rest of the cathepsin B genes at the deepest branches of the tree, but its position was poorly resolved although it tends to group with the mammalian enzymes. The consistent location of the plant enzymes at deeper branches than the metazoan counterparts could account for an early divergence of the plant enzymes, or could be due to different substitution rates in plants and metazoans.

In all the topologies obtained, the cysteine proteinases from the parasitic nematodes *Haemonchus conturtus* (HAECO_CYS1, HAECO_CYS2, HAECO_AC3, HAECO_AC4, HAECO_AC5,HAECO_PDM4), *Otertagia ostertagi* (OSTOS_CYS1) and *Ancylosotoma caninum* (ANCCA_CB1, ANCCA_CB2) clustered together with high bootstrap values. Partial sequences of other cathepsin Bs from these organisms and from the parasitic nematode *Strongyloides ratti* also clustered in the same node (data not shown). The several cathepsin B-like enzymes from the free living nematode *Caenorabdithis elegans* were found in a well defined node, placing almost all the nematode sequences together. The only exceptions are one of the *C.elegans* enzymes (CAEEL_CPR6) and an enzyme from the parasitic nematode *Ascaris suum* (ASCSU_CATB), that seem to have diverged earlier. The cathepsin B-like enzymes of the trematodes *Schistosoma mansoni* and *Schistosoma japonicum* constitute a well defined node. Another well defined internal cluster comprises the vertebrate enzymes (RAT_CATB, MOUSE_CATB, HUMAN_CATB, BOVIN_CATB, CHICK_CATB) and a representative from the flesh fly *Sarcophaga peregrina* (SARPE_CATB). A second insect enzyme from *Aedes aegypti* (AEDEG_CATB) also tends to cluster with these, although it is not well supported by the bootstrap data.

The cathepsin B enzymes possess a major insertion in the primary sequence that is not present in other cysteine proteinases of the papain superfamily. This insertion accounts for the occluding loop, and it is responsible for the carboxy-terminal dipeptidylpeptidase activity typical of the mammalian cathepsin B enzymes. This loop is conserved in the cysteine proteinases from the common flesh fly *S. peregrina*, the parasitic nematodes *H.contortus, O. ostertagi* , *A. caninum, Ascaris suum*, three of the cysteine proteinases produced by the free living nematode *Caenorhabitis elegans*, the parasitic trematodes *Schistosoma mansoni* and *Schistosoma japonicum*, and the parasitic trypanosomatid *Leishmania mexicana*. The two histidine residues related to the exopeptidase activity (His^{110}, His^{111} cathepsin B numbering) are conserved in the aforementioned proteinases (even in a non catalytic form of the *S.japonicum* enzyme (SCHJP-CB2) that bears a serine residue replacing the active site cysteine), with the only exception being two members of the *C.elegans* cathepsin B gene family. In these two proteinases (CAEEL_CYS1 and CAEEL CPR3) these residues have been substituted by a Glu and a Thr respectively. A similar, but shorter loop, was observed in the remaining members of the *C.elegans* family of cysteine proteinases, in the yellow fever mosquito vector *Aedes aegypti*, and in the plant enzymes of this group from wheat and tobacco. The plant homologues maintain only one of the histidine residues responsible for the exopeptidase activity (Fig 6.4).

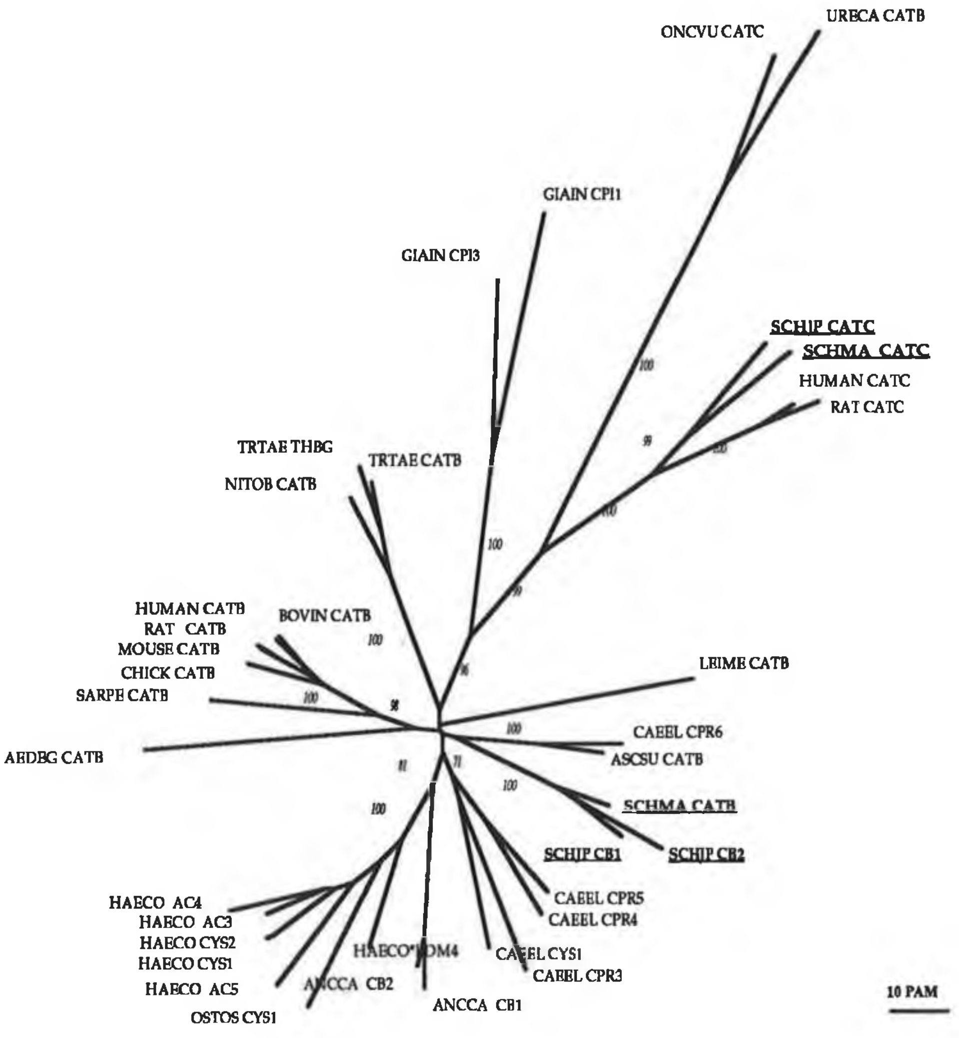


Fig. 6.3

Neighbor joining unrooted tree of members of the cathepsin B class.
Bootstrap values of the main branches are indicated.
The platyhelminth sequences are underlined

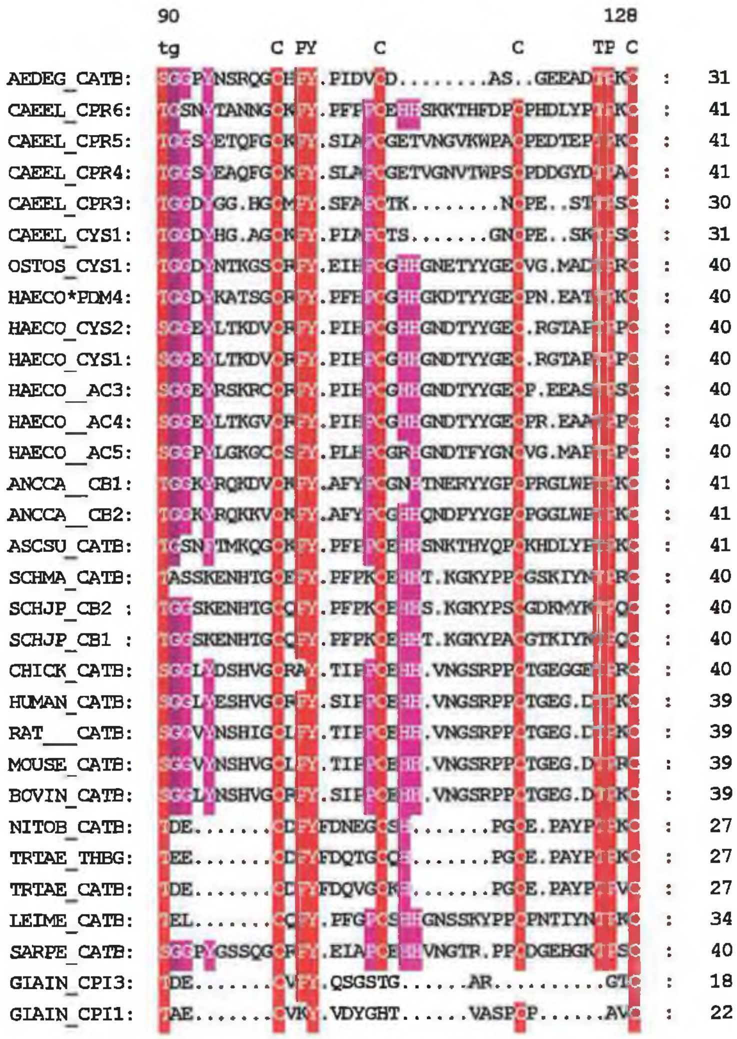


Fig.6.4

Alignment of the region of the occluding loop of members of the cathepsin B family. Conserved residues are indicated (red 90%, purple 75 %, pink 60 %). The numbering on top corresponds to the positions in the rat cathepsin B sequence, and on the right the length of the region for each sequence.

A second well defined cluster is constituted by the lysosomal dipeptidyl peptidases cathepsin C of mammals, and similar enzymes from invertebrates (Fig. 6.1, 6.2 and 6.3). A first node contains the mammalian enzymes and counterparts from the parasitic trematodes *Schistosoma mansoni* and *S. japonicum*, and is supported by a bootstrap value of 100. These enzymes are characterized by the presence of a very long propeptide. The function of this long propeptide has not been elucidated, and it remains to be seen if it can inhibit the mature enzyme in the same way as the propeptides of the other members of the superfamily do. However, we have detected similar conserved motifs in the carboxy-terminal end of this long propeptide to the ones present in the non cathepsin B members of the superfamily (see below). Two other enzymes group with the cathepsin C and cathepsin B nodes better than with any other node. The first has been described as a cathepsin C that takes part in the molting process of the larvae of the parasitic nematode *Onchocerca volvulus*, and the second is a gene product from the echiurid worm *Urechis caupo*. Although the last one was originally termed as cathepsin B-like gene, our alignment shows that it has a much closer relationship to the *O. volvulus* gene, with a bootstrap probability of 100%. These two sequences have a very short propeptide but in spite of this difference, they cluster with the cathepsin Cs with good bootstrap value (87% when considering the whole protein sequence, 76% if only the mature enzyme is considered) within the cathepsin B class.

The cathepsins B and C families constitute adjacent nodes, occurring with a bootstrap confidence value over 90%, so the grouping of cathepsins B and C into a cathepsin B class was considered reliable (Fig. 6.3). Furthermore, the cathepsins C have an extra pair of cysteine residues that could correspond well with a disulphide bridge in the corresponding position in cathepsin B (residues C100-C132). The *S. mansoni* sequence has a substitution in the first of these positions, although it is not substituted in the closely related *S. japonicum* . The *O. volvulus* and *U. caupo* genes also share with cathepsins B a short loop in the same position of the occluding loop. The two cysteine residues that hold together this loop in cathepsins B are also present in the *O.volvulus* and the *U.caupo* sequences, making them interesting as they have features common to the two families.

6.3 The ERFNIN complex

Karrer *et al.* (1993) defined two families within the papain superfamily based on the presence or absence of a conserved interspersed motif, the ERFNIN motif, in the propeptide of several members of the papain superfamily. For example, this motif is present in the cathepsin L family but is absent in the cathepsin B family. The structural basis of this difference was elucidated when the structure of proenzymes was determined. It was evident that this ERFNIN motif comprised a conserved alpha helix, but also that it corresponded to a similar but shorter one in the propeptide of cathepsin B-like enzymes. The ERFNIN complex of enzymes comprises the majority of the proteins in the papain superfamily, and can be subdivided into several families. Berti and Storer (1995) pointed out the existence of at least three classes within this complex, although the group is poorly resolved. This seems to be due to a period of rapid gene duplication and divergence. Furthermore the whole papain superfamily is a mixture of paralogous and orthologous sequences, which could be subjected to different selective pressures, making the relationships between groups more difficult to establish. The existence of long propeptides in several sequences (namely the apicomplexans *Plasmodium* and *Theileria*), and carboxyterminal extensions in several sequences (*Trypanosoma, Leishmania*, several plant enzymes) made difficult the alignment of the full length proteins, obscuring the possible relationships between groups. However, a much better result was obtained by the alignment of the more conserved mature proteinase sequences. The general position of the different groups was maintained, with some variation in the situation of the protist genes, that, in general, tend to separate at the deeper branches of the tree. Based on the previous work of Berti and Storer (1995) and our own analysis we have separated the ERFNIN complex into four classes.

6.3.1 The cruzipain class

The cruzipain class is the first well sustained class (72% bootstrap value) and include sequences from several protists, helminths and plants(Fig. 6.1, 6.2 and 6.5). All the sequences from the parasitic kinetoplastid protists *Leishmania* and *Trypanosoma* (with the only exception of *Leishmania*'s cathepsin B-like enzyme) constituted a well defined node supported by a bootstrap probability of 100%, and are the first to diverge in the group. Sequences from two other protist are present in this group, the heterolobosean amoeba *Naegleria fowleri* (NAEFO_CPRO), and the cellular slime mold *Dictiostelium discoideum* (DICDI_CYS1). The *N. fowleri* gene diverges previously to the genes for the *Schistosoma mansoni* cathepsin L1 (SCHMA_CL1) and the neutral thiol protease of the trematode *Paragonimus wetermani* (PARWM_NTP), while the cysteine proteinase 1 from the slime mold *Dictiostelium discoideum* branch diverges before a well defined group of sequences from plants (NITOB_CYP7, NITOB_CYP8, LYCES_CYP2, MAIZE_CYS1, ARATH_A494, ARATH_RD19, PEA_CP15A, VICFA_CPRO, VICSA_CPRO, and SOYBN_CEND) (Fig 6.1 and 6.5).

A short loop (nine residues in plants, 10 in the protist sequences) is present exclusively in the slime mold, *Naegleria* and plant enzymes. Two cysteine residues at the ends of this loop are conserved, and correspond to the residues Cys^{63} and Cys^{67} of cathepsin B, involved in a disulphide bond. By comparison with the crystal structure of cathepsin B, this loop would be on the surface of the left domain, and away from the active site, and possibly stabilized by a disulphide bond between the two cysteines. Another extra pair of cysteine residues exist in the plant enzymes, at positions equivalent to residues 150 and 190 of papain, and they are at sufficient distance to establish an extra disulphide bridge exclusive to this subfamily of plant enzymes. No similar residues exist in the slime mold or the amoeba genes, or in any other member of the superfamily, suggesting that these substitutions have occurred after the divergence of plants.

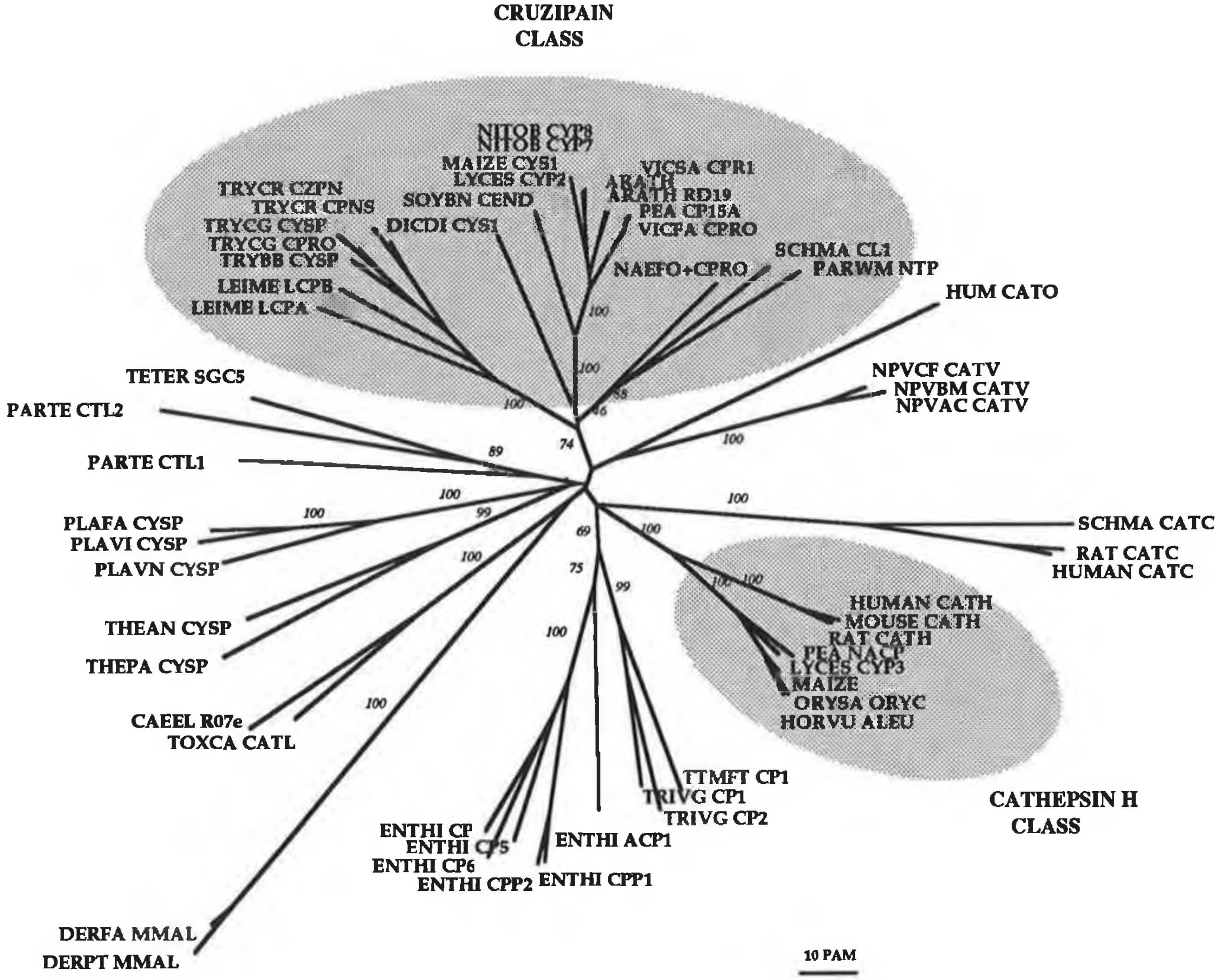


Fig 6.5
Neighbor joining unrooted tree of complete coding sequences of the cruzipain and cathepsin H classes, and the cysteine proteinases of other protists, virus and the unusual proteinases of nematodes, mites, and human cathepsin O. Cathepsins C were included for comparison. The different recognized classes are indicated by shaded areas. Bootstrap values of the main branches are also indicated. All positions where gaps existed in at least one sequence were not included. The topology of the tree was not influenced by the inclusion in the analysis of the positions with gaps.

So far the sequences of the platyhelminths *Schistosoma* and *Paragonimus* are the only metazoan representatives of this class, although the fact that representatives of this class occur in kinetoplastids, slime molds, and plants, suggests that orthologous members of this class should be present in other metazoans. It is noteworthy that the *Schistosoma mansoni* gene was originally described as cathepsin L-like (Smith *et al.*, 1994), based on sequence comparison with mammalian cathepsins L, and on activity towards synthetic substrates. A sequence for the mature region of a similar gene from *Schistosoma japonicum* has been isolated (*S.japonicum* cathepsin L1, Day *et al.* 1995) and is highly similar to the *S.mansoni* sequence. Other genes from this organism (*S.mansoni* and *S. japonicum* cathepsins L2) exist and they cluster in a different class (see below), suggesting that *Schistosoma* cathepsins L1 and L2 are in fact paralogous genes.

Although Berti and Storer (1995) recognized this class as a reliable one, they proposed the name Ddis1 class to it. Due to the presence of kinetoplastid sequences in this group, that diverged earlier in the mitochondrial eukaryote lineage than the Dictiostellida, we propose to rename this group (after the most well studied member) as the Cruzipain class.

6.3.2 The papain class

A well supported node of enzymes from plants constitute the papain class (Fig 6.1 and 6.6). Within this class, a distinct clade is constituted by the papaya enzymes papain (CARPA_PAPA), chymopapain (CARPA_PAP2), glycyl endopeptidase (CARPA_PAP4) and caricain (CARPA_CARI). The placement of this node within the papain class was not well resolved. The remaining plant enzymes of the group have a mature enzyme that is longer than the ones existing in the papaya enzymes (due to a carboxy-terminal extension), by at least twelve residues, and can be subdivided in two groups.

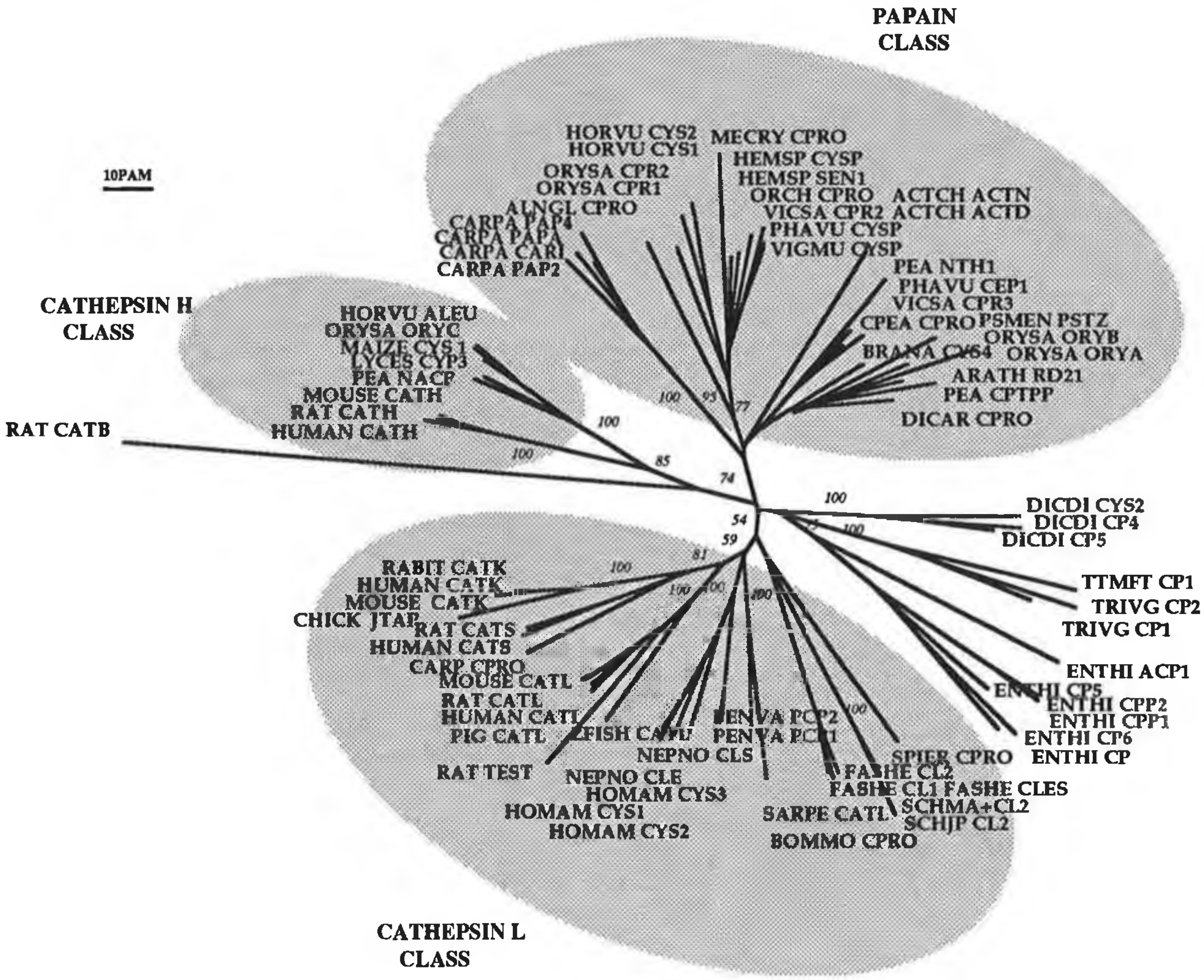


Fig 6.6
Neighbor joining unrooted tree of complete coding sequences of the papain, cathepsin L and cathepsin H classes. The cysteine proteinases of *Entamoeba*, Trichomonads, *Dictiostelium* and rat cathepsin B were included for comparison. The different classes are indicated by shaded areas. Bootstrap values of the main branches are also indicated. All positions where gaps existed in at least one sequence were not included. The topology of the tree was not influenced by the inclusion in the analysis of the positions with gaps.

One group of enzymes with a good bootstrap value (77%) include a dehydration responsive gene from *A. thaliana* (ARATH_RD21), a senescense related proteinase from pea (PEA_CPTPP), and germination related proteases from rice (oryzains, ORYZA_ORYA, ORYZA_ORYB), and douglas fir(PSMEN_PSTZ). These genes have a very long carboxy-terminal extension rich in cysteine residues, and homologous to the granulin-like mammalian growth factors, and not present in other cysteine proteinases, but shared by other pea mRNAs with the same pattern of induction as the pea cysteine proteinase. We propose to call this group the oryzain family, after the presence of oryzains alpha and beta in this group.

Several other plant enzymes (ORYSA_CPR1, ORYSA_CPR2, PHAVU_CYSP, VIGMU-CYSP, VICSA_CPR2, HEMSP_SEN1, HEMSP_CYSP and ORCH_CYSP) showed a carboxy-terminal KDEL or RDEL motif associated with retention in the endoplasmic reticulum. As in the previous group, the suspected function and tissue expression of these proteinases is diverse, so it is not clear if these proteinases are orthologous. However, this group possess a conserved motif (WYEWVR) in the propeptide (discussed below) different from the rest of the members of the class, so it is reasonable to assume that they constitute a family of orthologous genes. We propose to call this the vignain family after one of its members (PHAVU_CYSP).

There is no non-plant sequences in the papain class, however the class shows general similarities with the cruzipain class, the cathepsin L class, the cathepsin H class and several proteinases from protists, but no significant clustering with these could be made. The complete absence of metazoan sequences in this group could be due to an origin of the class after the divergence of plants and animals, an early loss of the members of this class in the metazoan evolution, or more simply could be an effect of sampling bias.

6.3.3 The cathepsin L class

The cathepsin L class is composed of several different families including the mammalian cathepsins L, K and S, the digestive cysteine proteinases from decapods, the larval digestive cysteine proteinase of the fruit fly, the imaginal disc proteinase of the flesh fly, the egg cysteine proteinase of the silkmoth, the cathepsin L-like enzymes of trematodes (including the liver fluke CL1 and CL2, and the *S.mansoni* and *Sjaponicum* CL2), and probably *Dictiostelium* cysteine proteinases. No sequences from plants were allocated to this class (Fig. 6.1 and 6.6)

Several families received good support according to the high bootstrap values. The vertebrate cathepsins L constituted a very well defined node. The rat testin (RAT_TEST) gene which has the active site cysteine substituted by a serine is also included in this group. The proteinases from arthropods cluster well with this group with good bootstrap values. Cathepsins S and cathepsins K constitute well defined families, and are clustered together with very good confidence values, indicating a recent duplication.

When only the mature enzyme or conserved regions were analyzed, the *Dictiostelium* cysteine proteinases (DICDI_CP2, DICDI_CP4 and DICDI_CP5) clustered within the cathepsin L class, but if the whole coding region was taken into account the slime mold genes separated at a deeper branch closer to *Entamoeba* genes and the diplomonads (Fig. 6.1, 6.2 and 6.6). This was not dependent on the long insertion at position 179 (cathepsin L numbering, 164 in papain numbering, and corresponding to the long serine-rich domain in DICDI_CP4 and DICDI_CP5) previous to the active site asparagine residue, since considering the mature enzyme or only conserved regions (excluding the insertion) gave similar results. An insertion in a similar position exist in the *Plasmodium* genes. The addition of the propeptide in the analysis seem to be responsible for this behavior. Although the alignment of the propeptides of several members of the cathepsin L class is possible, the *Dictiostelium* cysteine proteinases propeptides are more similar to the proregion of the members of the papain class. This could be the reason for its positioning in a deeper branch when the whole coding region is considered.

6.3.3.1 The platyhelminth genes

The sequences from the trematodes *Fasciola hepatica, Schistosoma mansoni, Schistosoma japonicum* and the cestode *Spirometra erinacei* tend to form a cluster, although not well resolved. The *F. hepatica* enzymes constitute a well defined node, but the relationships with the *S.mansoni* and *S.japonicum* cathepsin L2 are not well supported. The latter sequences showed more homologies with the sequence from the cestode *S. erinacei* , in all the possible topologies. These two nodes join when we consider the whole protein, but the *F.hepatica* genes separate first when only the mature region, or conserved regions are considered. An additional sequence of the mature region of a cysteine proteinase from the cestode *Spirometra mansonoides* is very closely related to the *S.erinacei* gene (data not shown).

In all situations analyzed (complete coding sequence, mature enzyme regions, conserved regions, or propeptides) the platyhelminth genes tend to group with the cathepsin L class, although with poor bootstrap values (58% when considering the whole sequences). No alternative positioning of the platyhelminth sequences was observed in any of the situations analyzed, so this clustering was considered as the most probable. Furthermore, a parsimony analysis restricted to the mature regions of the cathepsin L class (including the platyhelminth sequences) was performed, showing a good correlation for an early branching of the platyhelminths from the cathepsin L class (Fig. 6.7).

In any of the situations analyzed, the platyhelminth genes separated before the divergence of the mammalian cathepsins L, K or S. The distances of the *F.hepatica* genes to the vertebrate cathepsins L, K and S, indicate that the trematode genes should not be considered as cathepsins L-like, as they are as related to them as to the cathepsins K or cathepsins S (table 6.1).

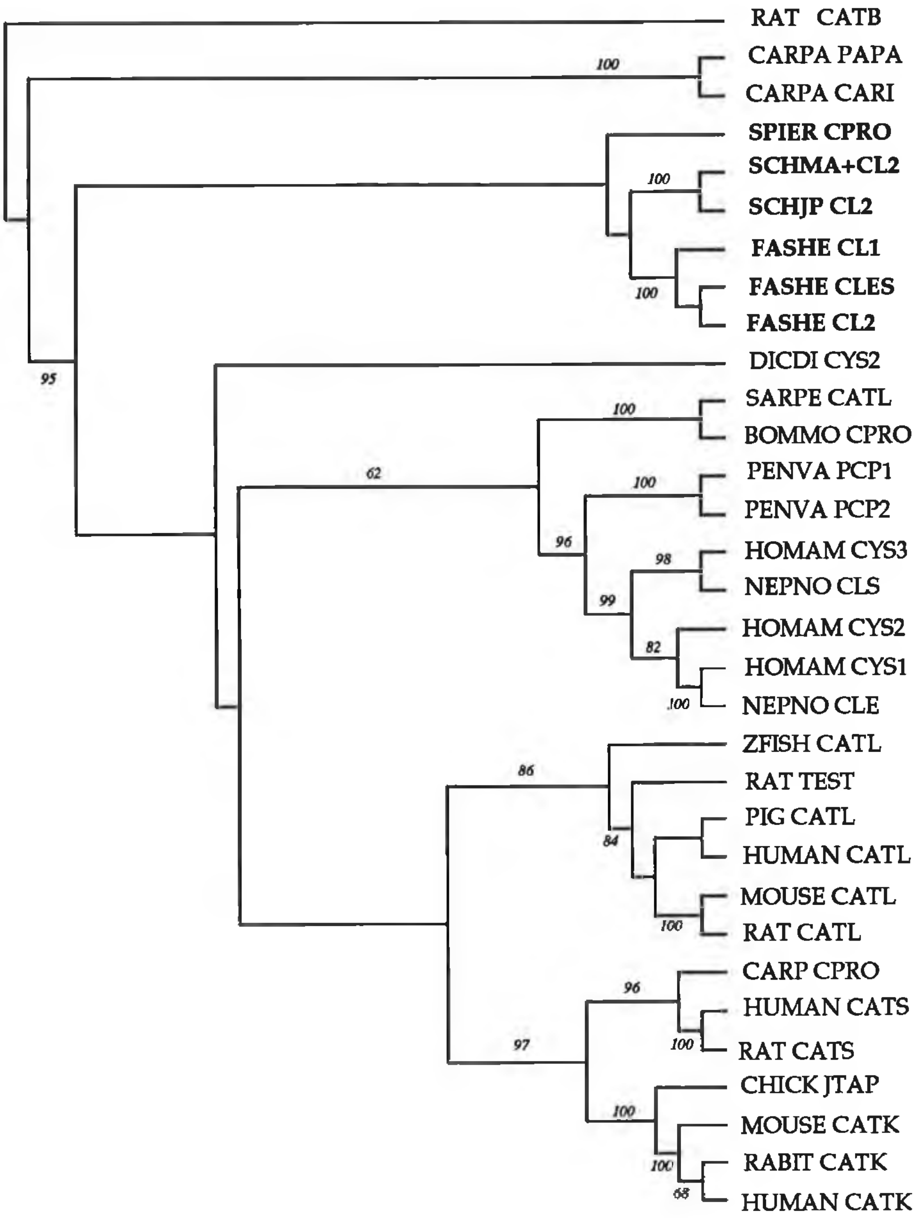


Fig. 6.7
Parsimony analysis of the mature region of members of the cathepsin L class. Rat cathepsin B , papain and caricain were used as outgroups. The sequences from platyhelmints are indicated in bold type.

	HL	ML	RL	NL	D2	HK	MK	HS	RS	FL1	FL2	HH	MH	RH
HUMAN_CATL	-													
MOUSE_CATL	0.268	-												
RAT_CATL	0.255	0.057	-											
NEPNO_CLE	0.515	0.519	0.509	-										
DICDI_CYS2	0.548	0.567	0.576	0.530	-									
HUMAN_CATK	0.467	0.502	0.511	0.567	0.594	-								
MOUSE_CATK	0.487	0.511	0.511	0.585	0.616	0.134	-							
HUMAN_CATS	0.493	0.502	0.505	0.571	0.598	0.454	0.474	-						
RAT_CATS	0.479	0.506	0.512	0.569	0.584	0.455	0.469	0.280	-					
FASHE_CL1	**0.554**	0.579	0.579	0.591	0.584	**0.562**	0.567	**0.575**	0.585	-				
FASHE_CL2	**0.570**	0.576	0.576	0.588	0.606	**0.556**	0.574	**0.588**	0.591	0.227	-			
HUMAN_CATH	0.619	0.623	0.619	0.630	0.634	0.623	0.639	0.654	0.657	**0.630**	**0.611**	-		
MOUSE_CATH	0.599	0.604	0.613	0.617	0.632	0.601	0.614	0.645	0.654	0.604	0.611	0.207	-	
RAT_CATH	0.599	0.604	0.607	0.624	0.622	0.604	0.624	0.639	0.654	0.598	0.601	0.204	0.069	-
CARPA_PAPA	0.638	0.638	0.644	0.633	0.648	0.615	0.631	0.644	0.670	0.640	0.646	0.605	0.631	0.627

Table 6.1
Distances of several members of the cathepsin L class of cysteine proteinases
The distances of *F. hepatica* cathepsins L1 and L2 to human cathepsins L, S, K and H are highlighted.

6. 3.4 The cathepsin H class

Two well defined families constitute the cathepsin H class. The lysosomal cathepsins H from vertebrates, constitute the first family with a 100 % bootstrap probablility. The barley aleurain (HORVU_ALEU), the oryzain gamma from rice (ORYSA_ORYC), and cysteine proteinases from maize (MAIZE_CYS2), tomato (LYCES_CYP3) and pea (PEA_NACP) constitute the aleurain family. These two families constitute adjacent nodes occurring with a bootstrap confidence of 100%, so their inclusion into a class was considered very reliable. The cathepsin H class separates at a deep branch in the noncathepsin B group, and no relation with the previously described classes or with the protist genes received significant support (Fig. 6.5 and 6.6).

This class is characterized by the exclusive presence of an extra pair of cysteine residues. The first of these residues is on the carboxy-terminal end of the propeptide, at a position equivalent to residue 82 in cathepsin L. The second cysteine residue conserved in this family is at the carboxy-terminus of the protein in a position equivalent to residue 214 in cathepsin L. Analysing the crystal structure data of procathepsin L and procaricain, we concluded that these residues in the cathepsin H protein are located within a short distance of each other, but it is not clear if they are able to establish a disulphide bridge. A disulphide bridge in this position could have important consequences as the propeptides of members of the papain superfamily have been shown to act as potent inhibitors of the mature enzyme.

6.3.5 The protists genes and families

The protist proteinases tend to separate at the deeper branches of the tree and no significant relation with the previously described classes can be established (Fig. 6.1, 6.5 and 6.6). The several cysteine proteinases of *Entamoeba* constitute a well defined family. A second well defined clade include the virulence factor proteinases of *Trichomonas vaginalis* and *Tritrichomonas foetus*. These two families tend to cluster together, and although the bootstrap values are low, no alternative topologies were detected. The trichomonads are amitocondriate organisms, and probably one of the earliest eukaryotic lineages. The presence of two different cysteinases from amitocondriate taxa (the cathepsin B-like enzymes of *Giardia*, and the ERFNIN-complex genes of the trichomonads), as well as confirming an early origin of the superfamily probably just after the eukaryote/prokaryote divergence, also indicate that at least two different genes exist in all mitochondriate taxa.

The several cysteine proteinases from the apicomplexan *Plasmodium* cluster together as expected, but not with the two genes of *Theileria* cysteine proteinases, while the proteinases from the ciliates *Paramecium* and *Tetrahymena* are not well resolved. In general all the cysteine proteinases from Alveolata (phylum that comprises apicomplexans and ciliates) were poorly resolved, but in all the cases they are located within the ERFNIN-complex.

6.4 Miscellaneous proteins

Two particular ERFNIN containing groups are the proteinases of baculovirus, and the mite allergens. Although these sequences clearly belong to the papain superfamily, the relationships of these enzymes are obscure, due to a tendency of their clusters to separate at deep branches of the tree.

A third group include two nematode sequences, one from the parasite *Toxocara canis* and the other from the free living nematode *C.elegans*, constituting the first non cathepsin B-like enzymes from nematodes. These enzymes were very poorly resolved in all the cases. In a similar situation is the human cathepsin O cloned from breast carcinoma.

6.5 Comparison of the propeptides of the papain superfamily

The propeptides of several proteinases of the papain superfamily have been demonstrated as powerful specific inhibitors of their cognate enzyme. At the same time that the mature enzymes evolve to allow different activities, the propeptides must change to keep the inhibitory activity. The analysis of the crystal structures of several proenzymes allowed the recognition of a possible mechanism for this inhibition, as the propeptide folds over the active site cleft. The conservation in the mode of this interaction allows the prediction (after alignment) of the regions that might play a role as effective inhibitors. The alignment of members of the different groups shows a striking level of conservation within close related enzymes in the region predicted to make these contacts (Fig 6.8).

The alignment of the propeptides of the cysteine proteinases of the papain superfamily was possible in the two main groups, namely, the cathepsin B class and the ERFNIN complex. As expected, the propeptides fall into the same groups as the full length enzymes or the mature portions, although some variations could be detected (Fig 6.8). The constraints on the mature region of the gene are obviously tighter, but the fact that the propeptides are involved in the stabilizing and folding of the enzyme suggest that this region too would exhibit structural conservation. In this sense we searched for conserved patterns in the propeptides of members of the different classes that could serve as identification tags for the class. The alignment of the propeptides was possible in the regions of predicted secondary structure, but the C-terminal regions of the propeptides (i.e. the region just before the cleavage point between propeptide and mature enzyme), were difficult to align due to extreme variation. This is not surprising, since the crystallographic data on procathepsin B, procathepsin L and procaricain showed that this region makes no major contacts with the mature enzyme and is not tightly structured. The propeptides of the cathepsin B enzymes range from 62 to 80 amino acids, whereas those in the remaining families range from 90 to 120 amino acids. The length differences observed between the cathepsin B families and the remaining cysteine proteases of the superfamily is accounted for by differences in the N-terminal region of the propeptide. The minor length differences observed in the propeptides within the cathepsin B families and within the ERFNIN complex could be explained by variations in the C -terminal non structured region.

6.5.1 The cathepsin B propeptide

In the cathepsin B family several conserved residues could be detected, that correspond to the conserved secondary structures in the proregion (Fig. 6.8). The first alpha helix of the cathepsin B propeptide is well conserved, and although few residues are strictly conserved, an interspersed motif of residues of similar biochemical properties could be identified. The general formula of the consensus motif is

Leu - Ser/Thr - $X_{2\text{-}3}$ - Leu - Val - X - Tyr - Aliph - Asn

The leucine residues are sometimes substituted by valine or isolecucine, so the aliphatic condition is maintained. Two residues separate the Ser / Thr residue from the next aliphatic residue in the sequences from vertebrates, plants, and *Schistosoma spp.*, while the remaining members of the family have three residues in this region. The final Asn residue is substituted by an Arg in the *Haemonchus* and *Ostertagia* sequences. This LLVYN motif is structurally equivalent to the C-terminal half of the ERFNIN motif present in the rest of the members of the papain superfamily.

The short beta strand detected in the crystallographic analysis of procathepsin B corresponds to a stretch characterized by an aromatic residue, a polar residue and a conserved Ala, substituted by a Val in the *Haemonchus* sequences.

Finally the short helix that in the crystal structure of procathepsin B competes with the occluding loop shows four well conserved positions : Lys - Leu - Met - Gly/Asn . The Met residue has been substituted by a Cys in the vertebrate cathepsins B. This cysteine residue interacts with the active site cleft in the rat procathepsin B. The next position is a conserved Gly substituted by an Asn in all the nematode sequences. The strong sequence conservation in this region, involved in interactions with the active site in the proenzyme, suggest that the contacts might be fundamental for the proper inhibition of the mature enzime.

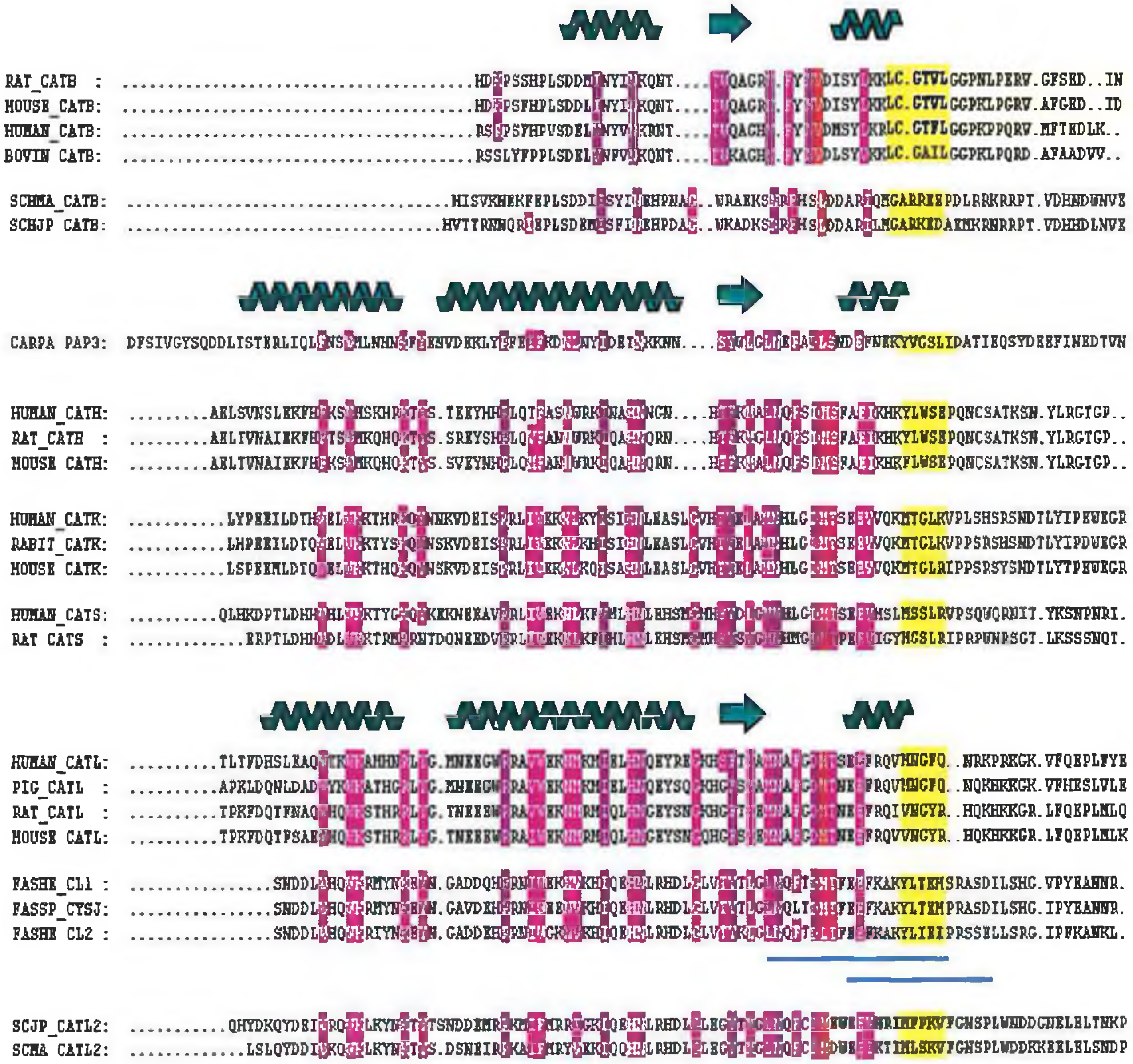


Fig. 6.8
Alignment of the propetide of cysteine proteinases of the papain superfamily. Conserved residues corresponding to the described motifs are indicated (red 90%, purple 75%, and pink 60%). The secondary structures of rat cathepsin B, human cathepsin L, and procaricain are depicted over the respective sequences. The region of contacts with the active site cleft is indicated in yellow, and by homology, extended to the other families. The blue lines indicate the region of the liver fluke propeptides chosen to analyze as a prospective specific inhibitor.

6.5.2 The ERFNIN complex

A) The ERFNIN motif has been identified as a characteristic tag of the non cathepsin B members of the papain superfamily (Karrer *et al.* 1993). This conserved motif corresponds with the second alpha helix of the propeptide of the members of this complex. This second helix constitutes the core of the prodomain, with the C-terminus of the first helix packing against one side and the amino terminus of the third helix packing against the other side (Coulombe *et al.* 1996, Groves *et al.* 1996). This structural feature is very well conserved along the whole complex. However, some variations in this ERFNIN motif between families and classes exist, as indicated in Table 6.2. A slightly different conserved motif can be recognized in each class (in brackets in the table), even when we consider only those residues strictly conserved. The neutral thiol protease of the lung fluke *P.westermani* is the only member of the group that seems not to have a typical ERFNIN motif. However, an analysis of the DNA sequence in this study shows that a sequencing error led to two frameshifts disrupting the alignment in this region. The sequence included in our analysis is corrected in this region.

B) Some other motifs can be ascribed to the other conserved structural features of the propeptide, and here again variations in these between families and classes can be established. The first helix of the propeptide can be characterized by a general motif

Phe/Trp - X2 - Phe/Trp - X5 - Lys/Arg - X - Tyr/Phe

with the only exception being the vignain family which shows a bigger and slightly modified motif :

Asp - Leu - X - Ser - X3- Leu - Trp - X - Leu - Tyr - Glu - X - Trp - X4 -Val - X - Arg

(the underlined residues are those corresponding to the general motif). Several variations in the general motif were found, and these correspond well with the classes and families proposed (table 6.3).

Cruzipain class **(ERFKNAQ)**	**E**	**e**	x2		**R**	x3			**F**	**k**	**x**	**N**	x3			**A**	x3			**q/a**		
Cruzipain family	E	E	x	x	R	f/l	x	v/a	F	k/r	q/e	N	l/m	x	x	A	k/r	x	x	A	a	a
Cruzipain like plant	E	E	H	x	R	f/l	x	V	F	K	x	N	L	x	r/k	A	k/r	x	h	Q	x	l
Dicdi-cys1 & Naegleria	E	E	x	x	R	f/y	x	l	F	K	x	N	I/v	x	K	x	x	x	x	N		
Trematode	e	D	x	x	R	f	x	I	F	k	x	n	i/l	x	r/k	A	Q	x	x	Q	x	r
Cathepsin H class **(ERFNIN)**	**E**	x3			**R**	x3			**F**	x2		**N/s**	x3			**I/v**	x3			**N**		
Cathepsin H	E	Y	x	H	R	L	Q	x	F	A	x	N	W	R	K	I	x	A	H	N		
Aleurains	E	x	x	x	R	F	r/k	I	F	S	e	n/s	L	x	x	i/v	R	S	x	N	r/k	
Cathepsin L class **(ERWNIHN)**	**E**	**E**	x2		**R**	x3			**w**	x2		**N**	x3			**I**	x2		**H**	**N**		
Cathepsin L family	E	E	x	x	r	R	a	i/v	W	E	k	N	m	k/r	m	I	q/e	I	H	N		
Cathepsin K & S	e/d	E	x	x	R	R	x	I	W	E	K	N	L	K	x	I	x	x	H	N		
Arthropod CPs	E	E	x	Y	R	x	x	V	f	e/q	Q	N	x	Q	x	I	x	x	h/f	N		
Platyhelminth CP	e/d	e/d	x	x	R	r/k	x	I	f/w	x	x	n	v	x	x	I	Q	x	H	N		
Papain class **(ERFIFKNIN)**	**E**	**K**	x2		**R**	**F**	x	**I**	**F**	**K**	x	**N**	x3			**I**	x3			**N**		
Papain family	E	K	x	Y	R	F	E	I	F	K	D	N	L	x	Y	I	D	E	x	N	K	
Vignain family	E	K	x	r/q	R	F	x	i/v	F	K	x	N	x	x	f/y	I	H	x	x	N	k	
Oryzain family	E	k	x	x	R	F	x	I	F	K	d	N	L	x	F	I	D	e	H	N	a	
Entamoeba family * **(ERIFNVN)**	**E**	x3			**R**	x2		**I**	**f**	x2		**N**	x3			**V**	x3			**N**		
Entamoeba	E	x	L	r	R	r	A	I	F	n	m	N	x	r/k	f	V	x	x	f	N		
Triplomonads	E	Y	x	x	R	f	G	I	y/w	I	S	N	k	R	i	V	Q	e	H	N		
Dictiostelium	E	F	x	x	R	y	x	I	F	k	S	N	M	D	Y	V	x	Q	W	N		
Alveolates # **(EKFNIKHN)**	**E**	x3			**k/r**	x3			**F**	x2		**N**	**y**	x2		**i**	**k**	x	**h**	**N**		

Table 6.2
The conserved residues corresponding to the second helix of the propeptide (the ERFNIN motif). Fully conserved residues are in capitals, while partial conservation is indicated in lower case.Consensus for classes is indicated in bold.
* Under Entamoeba family we included the sequences of cysteine proteinases from *Entamoeba, Dictiostelium* cysteine proteinases 2, 4 and 5, and the cysteine proteinases from *Trichomonas* and *Tritrichomonas*
# Under Alveolates we included cysteine proteinase sequences from *Plasmodium, Theileria, Paramecium* and *Tetrahymena.*

General consensus									f/w	x2		w/f			x5			k/r	x	y/f		
Cruzipain class									F	x	x	F	K	x	k	x	x	k	x	Y		
Cathepsin H class									F	x	x	w/f	x	x	x	x	x	K	x	Y	x	S
Cathepsin L class									W	x	x	w/f	K	x	x	x	x	k/r	x	Y		
Papain class	S		x3		L	x2		l	y	x2		W		x3		h	x	k	x	y		
Papain family	S	x	x	x	L	x	x	L	F	x	x	W	x	x	K	H	H	K	x	Y		
Vignain family	S	x	x	x	L	w	x	L	Y	E	x	W	x	x	x	H	V	x	x	R		
Oryzain family								M	Y	x	x	W	L	x	x	x	x	K	x	Y		
Entamoeba family *									F	x	x	W	x	x	x	x	N	k/r	x	f/y	t	
Alveolates #									f	x	x	w/f	x	x	k	y	n	k/r	x	y	x	n

Table 6.3
The conserved residues corresponding to the first helix of the propeptide (the FWKY motif).

Fully conserved residues are in capitals, while partial conservation is indicated in lower case.Consensus for classes are indicated in bold.

* Under Entamoeba family we included the sequences of cysteine proteinases from *Entamoeba, Dictiostelium* cysteine proteinases 2, 4 and 5, and the cysteine proteinases from *Trichomonas* and *Tritrichomonas*

# Under Alveolates we included cysteine proteinase sequences from *Plasmodium, Theileria, Paramecium* and *Tetrahymena.*

C) A third conserved region in the propeptide comprises the short beta strand (that makes contacts with the mature enzyme), the loop leading to the third helix, and the third helix (Fig. 6.8). Again, some variations can be seen within the classes and families. The most striking difference is in the cruzipain class, that clearly differentiates from the rest, in the region corresponding to the beta strand, which is involved in stabilizing the interactions of the propeptide with the mature enzyme. While the rest of the members of the ERFNIN group have the motif Phe/Tyr - X - Aliph - X - Aliph - Asn , this class only maintain the second aliphatic group, substituting the aromatic residue for a conserved Ala, an His or Phe in the position of the first aliphatic group, and a conserved Thr in the position of the Asn. Interestingly the beta strand on the mature enzyme that interacts with this region of the propeptide (known as prosegment binding loop PBL) is very well conserved among the whole superfamily, including the cruzipain class.

The residues corresponding to the third helix are conserved among the whole ERFNIN complex, and an interspersed conserved motif can be established with the general formula F x D M/L T/S x2 E F (Table 6.4). The Asp residue is conserved in all the members of the different classes, with the only exception being the trichomonads where it is substituted by a His, and *Entamoeba* where it is substituted by an Ala.

According to the crystallographic data, the terminal portion of the propeptides is rather unstructured, and this appears on the alignments as a very variable region both in residues and length. However, this region is generally rich in charged residues. It is noteworthy that the trematode sequences show a Asn residue at a terminal position, 3 or 4 residues before the actual start of the mature enzyme. The recent characterization of Sm32 as an aspariginyl endopeptidase in *Schistosoma mansoni*, similar to legumains might suggest that this enzyme is responsible for the cleavage of the proenzyme generating the mature active form. Similar terminal or near to terminal Asn residues can be detected in several of the plant cysteine proteinases, and in the allergens of mites.

General consensus	**F**	**x**	**D**	**m/l**	**t/s**	*x2*		**E**	**F**						
Cruzipain class	F	S	D	L	T	x	x	E	F	r	x	x	y/f	l/h	g
Cathepsin H class	F	X	D	m	s	w/f	e/a	E	f	x	x	x	x	l	g/w
Cathepsin L class	F	G	D	m	t	x	e	E	f/v	x	x	x	m	x	g
Papain class	F	a/g	D	l/m	t	n	x	E	f/y	r/k	x	x	y/f	x	g
Entamoeba family *	f	A	d/h/a	l/m/i	t	n/p	x	E	Y	x	x	x	l	l	g
Alveolates #	F	s	D	y/l/m	s	x	x	e	f	x	x	x	f	x	x

Table 6.4
The conserved residues corresponding to the third helix of the propeptide.
Fully conserved residues are in capitals, while partial conservation is indicated in lower case.Consensus for classes are indicated in bold.

* Under Entamoeba family we included the sequences of cysteine proteinases from *Entamoeba, Dictiostelium* cysteine proteinases 2, 4 and 5, and the cysteine proteinases from *Trichomonas* and *Tritrichomonas.*

# Under Alveolates we included cysteine proteinase sequences from *Plasmodium, Theileria, Paramecium* and *Tetrahymena.*

6.5.3 The Cathepsin C propeptide

Although the cathepsins C cluster with the cathepsins B, the propeptide of these enzymes is unusual for its length (close to 200 residues). The only other cysteine proteinases of the papain superfamily with very long propeptides are the cysteine proteinases of *Plasmodium*. The propeptide of cathepsins C is similar in the C-terminal region to the propeptide of the ERFNIN complex proteinases (Fig.6.9)

A motif with a general formula W X_2 F X_2 K, is present and corresponds well both in sequence and position to the first helix motif present in the ERFNIN complex proteinases. However this could be part of a bigger alpha helix, as conserved interspersed hydrophobic residues extend a further 17 amino acids towards the amino terminus. A conserved motif related to the ERFNIN is also present with the general formula

E/Q X_5 L Y X_2 D/N X_2 F V X_2 I N

that corresponds well in sequence and position with the ERFNIN motif. At the position corresponding to the third helix, some conservation exists with a general formula Y X_2 L S/T. This could indicate that at least the C-terminal region of the long propeptide of cathepsins C can adopt a similar fold to the remaining members of the papain superfamily.

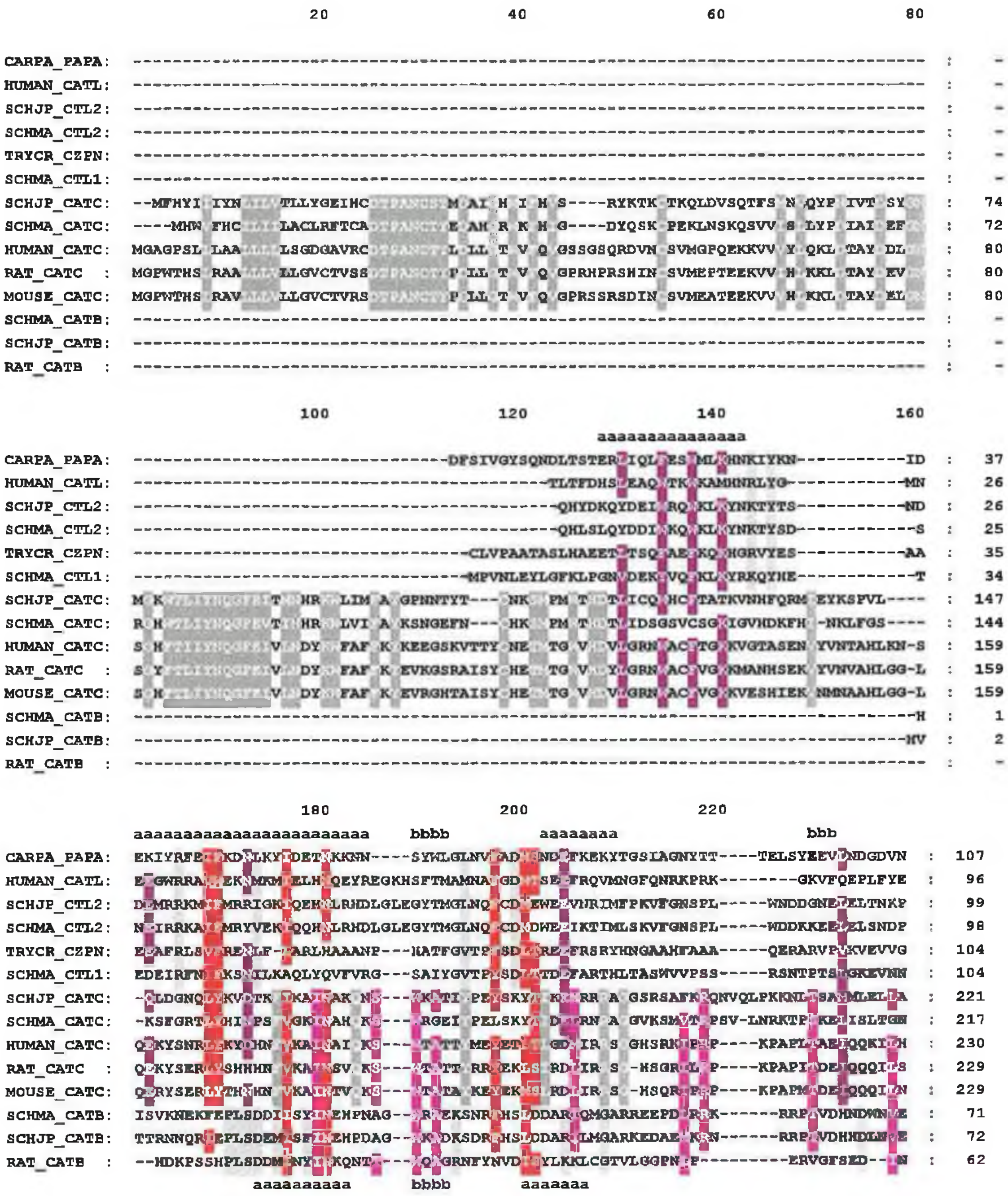


Fig. 6.9

Alignment of propetides of cathepsin Cs with representative members of other families.

Conserved residues within each group (cathepsin Cs, cathepsin Bs, cathepsin L and cruzipain) are indicated in gray shades. Conserved residues (over 70 %) between all the families are in red. Conserved positions between cathepsin B and cathepsin C are indicated in pink , while conserved positions between cathepsin C and the ERFNIN group are indicated in purple.

7 Discussion

7.1 Isolation of the cDNA encoding *Fasciola hepatica* cathepsin CL2

Secreted proteinases have been implicated in vital functions of parasitic trematodes including immune evasion, tissue penetration and nutrition (reviewed in Dalton and Brindley, 1997). Several reports indicated that cysteine proteinases are the major components of the secretions in *Fasciola hepatica*. These enzymes have been immunolocalized to granules in the gut epithelium (Smith *et al*., 1993). The physicochemical characterization of this enzymatic activity has proved to be difficult. The enzyme activity migrates as multiple bands in electrophoresis and purification attempts by different laboratories have yielded products with heterogeneous activities (Dalton and Heffernan, 1989; Yamasaki *et al*., 1989, Wijffels *et al*. 1994). Two enzymes with distinct physicochemical properties and substrate specificity, but similar to vertebrate cathepsin Ls, were identified in our laboratory as the major components of the secretion products of the adult *F. hepatica* and termed cathepsin L1 and cathepsin L2 (Smith *et al*, 1993, Dowd *et al*., 1994).

Because of their involvement in crucial biological functions these secreted cysteine proteinases have been considered candidates to which immuno-prophylaxis could be directed. Recently, Dalton *et al.* (1996) demonstrated that vaccination of cattle with cathepsin L1 and cathepsin L2 elicited high levels of protection against reinfection and induced significant reduction in egg output and viability. Wijffels *et al.* (1994) have also shown that vaccination of sheep with a preparation of *F. hepatica* cysteine proteinases, which presumably contained both cathepsin L1 and cathepsin L2, could induce high anti-fecundity effects. The cathepsin L-like enzymes secreted by liver fluke have been demonstrated as valuable tools in immunodiagnosis of human fasciolosis (O'Neill *et al*., 1997), and in fasciolosis in cattle and sheep (Buchon *et al*., pers. comm.). Fagbemi and Guobadia (1995) successfully used a cysteine proteinase of 28 kDa from *Fasciola gigantica* for immunodiagnosis of fasciolosis in cattle, sheep and goat.

In order to obtain a better understanding of the differences of the two secreted cathepsin L-like enzymes purified in our laboratory, and to further analyze their involvement in the interaction with the host, we decided to clone the cDNAs encoding these enzymes. The screening of an adult cDNA library with anti-cathepsin L1 antibodies allowed us to identify the gene encoding cathepsin L1 (Roche *et al.*, 1997). In the present work we used a similar approach to purify a clone encoding cathepsin L2. Using an antiserum prepared against purified cathepsin L2 to screen an adult *F. hepatica* cDNA library, a cDNA clone encoding a complete preprocathepsin L was isolated. The sequence obtained for the clone FheTCL2 is 86 % identical to the previously cloned cathepsin L1 at the nucleotide level, showing a 78 % homology in the deduced amino acid sequence (Fig. 3.3 and Table 3.1). A comparison of the immunological and physicochemical properties of the recombinant enzymes by expressing the cDNAs in yeast, has shown that both enzymes are different (Figs. 4.2 and 4.3). This clearly indicates that the two different cathepsin L-like activities present in the secretions of adult *F. hepatica* are encoded by different genes.

An alignment of all the sequences of cathepsin L-like enzymes of *F. hepatica* allowed us to analyze the relationships between the members of the *F. hepatica* cysteine proteinase gene family. At least five different genes bearing sequence similarities to vertebrate cathepsin L are present in the liver fluke (Figs. 3.3, 3.5 and Table 3.1). Very similar proteinases to cathepsin L1 have been cloned independently by different groups (FASJP-CSP : Yamasaki and Aoki, 1993, FASHE-CLEP: Wijffels *et al.*, 1994, FASHE-CP6G : Heussler and Dobbelaere , 1994). The FheTCL2 sequence here obtained, is 96 % identical to the clone CP1C reported by Heussler and Dobbelaere. A third clone obtained by Heussler and Dobbelaere (FASHE-CP3D) is highly homologous to a sequence obtained by Panaccio *et al.* (1994) (FASHE-CLES). Two other partial sequences show a greater divergence from the previous three groups and from each other, constituting independent groups (FASHE-CP2A and FASHE-CP4E). Yamasaki and Aoki (1993) reported the cloning of a second cysteine proteinase 86 % identical at the nucleotide level and 76 % at the amino acid level to FASJP-CSP, but they did not present the corresponding sequence. The identity values reported correlate well with the ones obtained in the comparison of cathepsin L2 and CSP (86% and 77%), suggesting that the cloned enzyme might correspond to a cathepsin L2 homologue. Wijffels *et al.* (1994) reported two different N-terminal sequences, one corresponding to the clone purified (FASHE-CLEP), and a second distinct polypeptide that clearly corresponds to a cathepsin L-like proteinase. This sequence does not correspond to any of the mature N-terminal regions known so far, but only partial sequences exist for two of the liver fluke cysteine proteinases groups (FASHE CP2A and FASHE CP4E).

Southern blot analysis showed a complex pattern of hybridization for all the cathepsin L-like genes, with the exception of CP4E that seems to be encoded by a single gene (Heussler and Dobbelaere, 1994). Although cross hybridization was not completely ruled out, it is feasible to think that the genes encoding the homologues to our cathepsins L1 and L2, and CP3D and CP2A are in multiple copies. We detected slight variations in the sequences of cathepsin L1 clones (FASHE-CL1 and FASHE-CL15) and cathepsin L2 (clones FASHE-CL2 and FASHE-CL26) originated from the same cDNA library. Two-dimensional SDS-PAGE of excretion/secretion cysteine proteinases indicated heterogeneity, with multiple proteins of different isoelectric point and similar molecular mass (Wijfeels *et al.*, 1994). These data taken together, and the very little differences observed between the clones isolated in different laboratories, indicate that intraspecific genetic variation exists for at least cathepsin L1 and cathepsin L2 genes. This polymorphism could be due to the presence of different, but very closely related genes belonging to a multicopy family, or could be due to allelic variation on a single gene. Further studies on the genomic organization of the cathepsin L-like genes of *F. hepatica* are needed to clarify this point.

All the genes cloned so far have been selected from adult cDNA, indicating that they are expressed by the adult liver fluke. However, Heussler and Dobbelaere (1994) showed by Northern blot experiments that the levels of expression of the various genes were different. The highest expression was detected for the homologues to our cathepsin L1 and L2, while strong hybridization is also found with the CP3D and CP2A probes, and weak signals are obtained with the CP4E probe. If cathepsin L1 and L2 are the more actively transcribed genes in the adult fluke, is not surprising that these clones have been repeatedly isolated by screening of adult cDNA libraries by different groups.

The different clones obtained by Heussler and Dobbelaere were obtained by PCR of adult cDNA with primers designed from cysteine proteinase consensus sequences. While this study gave valuable information on the genomic organization of the *F. hepatica* cathepsin L-like genes, the assignment of the different genes to functions and/or tissue distribution was not possible. Some of the products of these genes could be involved in normal cellular processes, including protein breakdown, in a similar manner to the mammalian cathepsins, while others may represent the secreted forms that play important roles in the interaction with the host, being valuable tools for immunodiagnosis and immunoprophylaxis. Although Yamasaki *et al.* (1992) localized a 28 kDa enzyme to the gut of adult *Fasciola sp.* flukes, these authors used PCR with conserved cysteine proteinase primers for the cloning of the enzyme, and obtained two different sequences (Yamasaki and Aoki, 1993). Using a sheep antiserum against a band of 28 kDa for the screening of a cDNA library the clones FASHE-CLEP (Wijffels *et al.*, 1994) and FASHE-CLES (Panaccio *et al.*, 1994) were obtained. Immunolocalization with the sheep antiserum identified cathepsin L-like activity in the gut microvilli and in the Mehlis gland. Considering the heterogeneity observed by the same authors in the 28 kDa region when analyzing the samples on two-dimensional SDS-PAGE, is difficult to assess which of the clones correspond to the secreted form.

In order to clone the cDNA encoding for the secreted cysteine proteinases of *F. hepatica,* an expression library was screened in our laboratory with specific anticathepsin L1 and anticathepsin L2 antibodies prepared against purified proteins from excretion/secretion products of adult flukes. A clone was isolated and unquestionably assigned to the secreted cathepsin L1. The recombinant protein obtained by expressing this gene in yeast has physicochemical and immunological properties indistinguishable to the cathepsin L1 purified from excretion/secretion products of adult liver fluke (Roche *et al.* 1997).

Here we present evidence that the FheTCL2 cDNA undoubtedly corresponds to the secreted form of cathepsin L2, the second major excreted proteinase of adult liver fluke. This is confirmed by: (1) the differences in sequence to the previously cloned cathepsin L1 gene (Fig. 3.3), (2) reactivity with specific anticathepsin L1 and anticathepsin L2 antisera (Fig. 4.2), and (3) similar reactivity towards synthetic substrates as the cathepsin L2 purified from excretion/secretion products of adult liver fluke (Fig. 4.3). Unpublished results by Dalton *et al.* localize cathepsin L2 to the digestive tract of adult flukes (Dalton, pers. comm.).

Based on previous data and on the sequence similarities detected in our analysis of the cysteine proteinase genes of liver fluke, it is evident that a complex of cysteine proteinases of similar mass (around 28-30 kDa) is present in liver fluke. The fact that almost identical genes have been isolated from a japanese strain *Fasciola sp.* where *F. hepatica* and *F. gigantica* coexist and interbreed , is a good indication that the 28 kDa cysteine proteinases detected by Fagbemi and Guaobadia (1995) in *F. gigantica* could correspond to any of the genes purified in *F. hepatica*. Some of these activities are encoded by at least three different genes [cathepsin L1 (Roche *et al.*, 1997), cathepsin L2 (this work) and cathepsin L3 (Pannacio *et al.*,1994)]. There is direct evidence that two of them, cathepsin L1 and cathepsin L2 (and probably the corresponding homologues detected in other laboratories), are the secreted forms detected in the gut microvilli, and consequently are the key players in the interaction with the host.

It is possible that the gene isolated by Panaccio *et al.* (cathepsin L3) encoded for the activity detected in the Mehlis gland, and interference with the function of this enzyme could be responsible for the anti-embryonation effect detected in vaccine trials. Michel *et al.* (1995) localized a cathepsin L of 31 kDa from *S. mansoni* to the reproductive tract of female worms. Cysteine proteinases have also been detected in eggs of the trematodes *Schistosoma mansoni* (Asch and Dresden, 1979, Sung and Dresden, 1986) and *Paragonimus westermani* (Kang *et al.*, 1995). However, the proteinases purified from *S.mansoni* eggs have a different molecular mass and seem to be more similar to vertebrate cathepsin Bs. Immunolocalization studies indicate the presence of these enzymes in the miracidial penetration gland. Recent studies indicate that treatment of schistosome infected mice with specific cysteine proteinase inhibitors reduce worm burden and egg production. The same inhibitors block schistosme hemoglobin degradation *in vitro*. Biotin tagged inhibitors allowed the identification *in vitro* of the targets of the inhibition as proteins of molecular masses 29, 31-35 and 45 kDa (Wasilewski *et al.*, 1996). However in this case, as in the vaccination trials with *F. hepatica* cathepsin L-like proteinases, it is not clear if the antiembryonation effect is direct, acting on cysteine proteinases expressed in the reproductive tract, or related to an inhibition of the digestive capabilities of the fluke.

A growing body of evidence in other organisms indicates that cathepsin-like cysteine proteinases might play important roles in early development. In several insects, cysteine proteinases are detected in eggs and a role in yolk degradation has been suggested. A cathepsin L-like enzyme detected in ticks is stored as a precursor, and is activated at acidic pH *in vitro*, and *in vivo* at a time corresponding with the active digestion of yolk (Fagotto, 1990a and b). Similar results were obtained in eggs of the silkmoth (Takahashi *et al.*, 1993), and a cathepsin L-like gene was cloned by immunoscreening of an ovary cDNA library of silkmoth (Yamamoto *et al.*, 1994). Acid activated cathepsin L-like activity capable of degrading vitellin, was purified from eggs of the orthopteran *Blatella germanica* (Liu *et al.*, 1996). Precursors of a cathepsin-B like enzyme activated by acidic treatment are present in eggs of the common fly (Ribolla and De Bianchi, 1995) and the fruit fly (Medina *et al.*, 1988). Digestion of yolk proteins by cysteine proteinases is not restricted to insects, as the same function has been suggested for a major cysteine proteinase present in embryos of the brine shrimp *Artemia franciscana* (Warner *et al.*, 1995). An acid cysteine proteinase activity was purified from *Xenopus* embryos, and the N-terminal sequence obtained showed similarities to cathepsin L (Miyata and Kihara, 1995). Based on these data, it is possible that the cathepsin-like enzymes immunolocalized to the reproductive tract of *F. hepatica* have a vital role to play in the development and life cycle of the parasite.

7.2 Yeast expression of recombinant cathepsin L2

While several cysteine proteinases of the papain superfamily have been expressed in bacterial systems, the resulting recombinant proteins accumulate in inclusion bodies, and have to be solubilized and refolded to obtain active enzyme (Smith and Gottesman, 1989, Eakin *et al.*, 1992, Mc Grath *et al.*, 1995, Groves *et al.*, 1996). This is mainly due to the absence of the mechanisms for post-translational modification of the recombinant enzyme, resulting in an improperly folded product. Insect and mammalian cell systems have been used to circumvent this problem with success (Vernet *et al.*, 1990, Menard *et al.*, 1990, Wiederanders *et al.*, 1991, Kane, 1993, Tao *et al.*, 1994). However, these systems are much more complex, costly and laborious. *Saccharomyces cerevisiae* is particularly useful for the expression of proteins because it is genetically well characterized and it possesses most of the sophisticated post-translational processing mechanisms of eukaryotic cells.

Functional expression of introduced genes in yeast requires the recognition of intracellular trafficking signals within the gene (Moir and Davidow, 1991). Cathepsin L proteinases are synthesized as preproenzymes that are sequentially processed to the mature molecules during their passage through the endoplasmic reticulum, Golgi complexes and lysosomes or secretory vesicles. During these processes the pro-peptide is required for proper enzyme folding, maintenance of stability, intracellular sorting and regulation of enzyme activity (Nishimura *et al.*, 1988, Rowan *et al.*, 1992, Tae *et al.*, 1994). The requirement of the propeptide for correct folding and sorting in the yeast system used was confirmed in our laboratory in experiments with recombinant liver fluke cathepsin L1. When yeast were transformed with a construct lacking the pre- and most of the pro-peptide of the cathepsin L1 no secretion of the enzyme was obtained, and although low levels of intracellular enzyme was observed, no detectable activity was present in cell extracts. This strongly suggests that the intracellular enzyme was incorrectly folded (Roche *et al.*, 1997). Here we report the functional expression of cathepsin L2 in yeast as a secreted enzyme using exclusively its own processing signals.

While secretion of eukaryotic proteins in *S. cerevisiae* using their own signals for intracellular sorting has been reported (Romanos *et al*, 1992), functionally active mammalian cathepsin B (Rowan *et al.*, 1992), cathepsin S (Bromme *et al.*, 1993), papain (Vernet *et al.*, 1993) and *S. mansoni* cathepsin B (Lipps *et al.*, 1996) have only been expressed in yeast as fusion proteins with the a-factor pre or pre-pro signals. In addition, these zymogens required a subsequent activation step to obtain functionally active enzyme. Mouse lysosomal cathepsin L was successfully expressed and targeted to the yeast vacuole using the native gene translation and post-translational signals, however, functionally active enzyme was not recovered (Nishimura and Kato, 1992). The first description of the functional expression of a cathepsin like enzyme in yeast without the need to use a fusion protein was achieved in our laboratory with cathepsin L1 (Roche *et al.*, 1997). Here we demonstrate that the procathepsin L2 gene also contains all the information necessary for processing and secretion in yeast . The extracellular expression of functionally active *F. hepatica* cathepsin L2 in DBY746 yeast suggests that the enzyme is trafficked through the normal secretory pathway (Fig. 4.1).

One problem often encountered with *S. cerevisiae* as a host for protein expression is the tendency for proteins to be over-glycosylated (Vernet *et al.*, 1993). In the case of rat cathepsin B, the yeast expressed products were heterogeneously glycosylated ; however, modification of the N-linked glycosylation motif -Asn-Gly-Ser- circumvented this problem (Hasnain *et al.*, 1992). The same modification was adopted for the yeast expression of cathepsin L (Kane, 1993, Carmona *et al.*, 1996). A similar strategy was used to express *Schistosoma mansoni* cathepsin B in *S. cerevisiae*, which was subsequently activated to a functional form (Lipps *et al.*, 1996). Unlike the mammalian cathepsins, the *F. hepatica* cathepsin like enzymes do not contain N-glycosylation sequences. It is not clear the significance of this difference.

By immunoblot analysis a major band of 24 kDa and a minor band corresponding in size to the proenzyme were detected in the supernatant of yeast cultures, confirming that the proform and the active form of the cathepsin L2 are secreted. N-terminal sequencing of the recombinant mature enzyme revealed that the processing of the cathepsin L2 proteinase in yeast involved cleavage at the precise position as that observed for the native enzyme (Dowd, pers. comm.). The *F.*

hepatica cathepsins L1 and L2 differ from mammalian cathepsin Ls in that the mature enzymes have an additional amino acid, Ala, at the N-terminus (Smith *et al.*, 1993, Dowd *et al.* 1994). Similarly an additional Asp or an Arg-Ala extension at the N-terminal of other members of the *F. hepatica* cathepsin L family have been reported (Tkalcevic *et al.*, 1995, Yamasaki and Aoki, 1993). Since processing in yeast also produced a mature enzyme with an additional Ala, it follows that this mechanism is not only dictated by the specificity of the processing enzymes but also by the structure of the molecule being processed. Processing in yeast involves endoproteinases, such as *KEX2* and *STE13*, that cleave at susceptible peptide bonds within the pro-region (Moir and Davidow, 1991). It is possible that the final processing steps are performed by exopeptidases that clip only as far as the molecular conformation of the mature enzyme will allow. The differences at the N-terminal end of the different fluke cathepsin L-like enzymes supports this idea. Although the processing leading to the mature form is still controversial in mammalian cathepsins, several reports suggest a concerted action of endopeptidases followed by exopeptidases (Rowan *et al.*, 1992, Mach *et al.*, 1993, Ishidoh and Kominami, 1994).

Asparaginyl residues were shown to occur near the cleavage site between the pro-peptide and mature enzymes of cathepsin L, cathepsin B, cathepsin C and cathepsin D proteinases of schistosomes, but were absent from their mammalian homologues (Dalton and Brindley, 1996, 1997). This observation led these authors to propose that a novel cysteine proteinase, an asparaginyl endopeptidase characterized in schistosomes, was the endoproteolytic activity involved in the initial cleavage of the pro-peptide of these proteinases (Dalton *et al.*, 1996). Asparaginyl residues are also present in the vicinity of the cleavage point between the pro-region and mature protein of the cathepsin L2 described in this study, and in cathepsin L1. Furthermore, *F. hepatica* also expresses an asparaginyl endopeptidase (Dowd and Dalton, unpublished). These observations suggest that trematodes possess a common mechanism, involving asparaginyl endopeptidases, for the processing and activation of cathepsin proteinases. A recent entry into the public databases (Genbank accession number U32517, PID g914991) revealed that *S. cerevisiae* possesses a putative asparaginyl endopeptidase. If this enzyme is involved in the maturation of secreted proteins it may explain why the *F. hepatica* cathepsin L is processed in yeast in a similar manner to that of the native protein.

7.3 Biochemical characterization and compared substrate specificity of *F. hepatica* cathepsin L-like proteinases

The production of the cathepsin L-like proteinase in the culture medium correlated well with the growth of the transformed yeast cells (Fig. 4.1). In addition, SDS-PAGE analysis revealed that the recombinant proteinase was a relatively major protein component in the culture medium but a minor component in the whole-cell extract (Fig. 4.2). Different substrate preferences of recombinant cathepsin L2 to recombinant cathepsin L1 (Roche *et al.*, 1994) were detected in preliminary assays with supernatants of yeast cells transformed with both recombinant clones. Moreover, a two to three fold higher efficiency towards the synthetic substrate Tos-Gly-Pro-Arg-NHMec was detected for the recombinant cathepsin L2 enzyme over recombinant cathepsin L1. Similar results with a second substrate with Pro in the P_2 position, Boc-Val-Pro-Arg-NHMec, are a clear indication that the recombinant enzyme has a different substrate preference to cathepsin L1 (Fig. 4.3). This pattern correlates with the previously described catalytic properties of native cathepsin L2 (Dowd *et al.*, 1994), confirming that the cloned gene encodes the secreted cathepsin L2 of *F. hepatica*.

Whether the antibodies raised against cathepsin L1 and cathepsin L2 are recognizing linear or structural epitopes, the differences in immunoreactivity observed between them should lie in the different residues found in both proteinase sequences. As the antibodies were raised against, and recognize, the mature enzymes, only differences in that portion of the sequence should be considered relevant. Modeling of the recombinant enzymes based on the high degree of fold conservation observed in cysteine proteinases of the papain superfamily has been used as a tool to gain understanding of the peculiarities of members of the family (Khouri *et al*, 1991, Klinkert *et al.*, 1994, Butler *et al.*, 1995, Groves *et al.*, 1996, Brinkworth *et al.*, 1996). Comparison of the models generated for *F. hepatica* cathepsin L1 and cathepsin L2 showed that 80 % of the residues that differ between the mature enzymes are exposed to the solvent, representing 28 % of the surface of the molecule (Fig.3.6). The distribution of these differences scattered over the surface of the molecule does not allow one to suggest any particular region as a primary target for the observed differences in antibody recognition. A more detailed analysis of those positions is needed to determine which of these are involved in the antibody recognition. Epitope mapping using short peptides spanning the regions which show differences between the two enzymes could clarify the point.

Pure cathepsin L2 was obtained following concentration of the culture medium and a single gel filtration chromatographic procedure. SDS-PAGE and immunoblot analysis of the purified material revealed that the purified recombinant protein co-migrated with the native cathepsin L2. In addition, the recombinant enzyme was reactive with anti-cathepsin L2 antibodies (Fig. 4.4) and exhibited a similar pH optimum against Z-Phe-Arg-NHMec to the native cathepsin L2 (Dowd, pers. comm). Moreover, a comparison of the enzyme reaction kinetics of the recombinant yeast-expressed and the native *F. hepatica* cathepsin L2 with peptide substrates demonstrated similar parameters for both enzymes (Table 4.1). Collectively, these data suggest that the cDNA isolated in this study encodes the preproenzyme of the major secreted cathepsin L2 proteinase.

Both recombinant enzymes exhibited a preference for the substrate Boc-Val-Leu-Lys-NHMec over Z-Phe-Arg-NHMec. Bromme *et al.*, (1987) demonstrated that an elongation of the substrate peptide chain enhances the hydrolysis rate of cathepsin B. As substrate binding is a cooperative process , the interactions established in one of the several subsites on the active site cleft of the enzyme, diminish the interaction energy for the binding at other subsites (Berti *et al.*, 1991). However, although this effect might exist, it can not explain the level of variation detected. While the differences in crude extracts of yeast culture supernatants are just two-fold, these are much greater in the enzymes purified from liver flukes (Dowd *et al.*, 1994), and in the purified recombinant enzymes (Dowd pers. comm.). This indicates that both *F. hepatica* cathepsin L-like enzymes have a general preference for substrates with Leu in the P_2 subsite over Phe in that position. A similar substrate specificity is detected in vertebrate cathepsins S and K (Bromme *et al.*; 1989, 1993,1996).

The residues present in the S_2 subsite have been indicated as the main determinant of the substrate specificity. An analysis of the residues lining the S_2 subsite in several cysteine proteinases of the papain superfamily show that these position are generally conserved within families (Tables 5.1 and 5.2). While the hydrophobic character of the residues forming the site is well conserved, subtle differences might account for different interactions with the substrate, generating diverse specificity. Cathepsin B enzymes are the most divergent in these positions, an observation consistent with the phylogenetic analysis (Berti and Storer, 1995 and below). In particular, a Gly is present in position 157 while aliphatic residues are preferred in the remaining members of the superfamily. Also, the residue at position 205, although not very conserved, tends to be Glu or other amphoteric residue in cathepsin Bs. It has been shown that this position does not contribute major interactions with the substrate in papain, and that could be the case in all the enzymes that present short side chains in this position (Storer and Menard, 1997).

As only two residues differ between *F. hepatica* cathepsin L1 and cathepsin L2 in the S_2 subsite, these may account for the observed differences in substrate specificity between the enzymes. As the substitution in position 70 (67 in papain numbering) implies a variation in the geometry and charge of the position, we decided to analyze the effect of mutating this position in cathepsin L1 (Leu) to the residue present in cathepsin L2 (Tyr). Site directed mutagenesis has been comprehensively employed in the analysis of the role of different amino acid residues in the function of enzymes, and is an approach widely adopted in the analysis of cysteine proteinases (Menard *et al.*, 1990, Khouri *et al.*, 1991, Hasnain *et al.*, 1992, Bromme *et al.*, 1994, Fox *et al.*, 1995, Katerelos and Goodenough, 1996). However, these studies have been limited to the papain and caricain, cathepsin B and cathepsin S; this is the first report of using this approach to analyze the biochemical properties of a parasite enzyme.

An overlap extension-derived PCR procedure was chosen for the mutagenesis because it is inexpensive, fast, and sensitive. Furthermore, the method allows the introduction of several mutations by using an appropriate combination of primers. In this study, some other unexpected point mutations were detected in the sequence of the product (Fig. 5.4). These can be explained by errors introduced by the Taq polymerase, and could be diminished by using polymerases with proof-reading activity, after the initial mutagenic cycle. The expected L70Y mutation and the others introduced, did not affect the yeast expression of the protein, nor its processing to the mature form. The presence of mature enzyme as detected by immunoblot in supernatants of yeast cultures transformed with the mutant gene, are clear indication that the protein was properly folded , sorted and processed (Fig. 5.5).

The comparative analysis of supernatants of cells transformed with the L70Y mutant cathepsin L1 showed no major differences to the recombinant non-mutated enzyme towards the substrates including a Pro residue in the P_2 position (Fig. 5.6). Similar preferences between purified cathepsin L1 and L70Y cathepsin L1 was detected in continuous assays with the substrates Z-Phe-Arg-NHMec, Boc-Val-Leu-Lys-NHMec, Tos-Gly-Pro-Arg-NHMec and Boc-Val-Pro-Arg-NHMec (Fig. 5.8). These results, while confirming that the mutation did not affect the ability of the enzyme to cleave synthetic substrates, indicate that the Tyr residue at position 70 is not responsible for cathepsin L2's preference for Pro in the P_2 position.

A Tyr residue is present in the equivalent position in vertebrate cathepsin Bs, cathepsin Ks, in the papaya proteinases, and in actinidin. This residue is in a pivotal position in the active site, forming the right side wall of the S_2 subsite and the left side of the S_3 subsite. It has been proposed that Tyr75 in cathepsin B allows for the binding of a P_2 Phe side chain through aromatic hydrophobic interactions (Klinkert *et al.*, 1994). Taralp *et al.* (1995) indicated that Tyr75 in cathepsin B takes part in the stabilizing of the residue at P_3, by establishing hydrogen bonds or aromatic-aromatic interactions with the substrate residue present at the site. Similar interactions with residues at P3 can be observed in the crystal structures of papain and glycyl endopeptidase complexed with synthetic substrates (Drenth *et al.*, 1976, O'Hara *et al.*, 1995). A comparison of the crystal structures of cathepsin B and L demonstrate that the presence of Tyr in cathepsin B makes the S_3 subsite narrower, while a Leu present in cathepsin L accounts for the ability to accept more bulky hydrophobic residues at P_3 (Fujishima *et al.*, 1997). These residues correspond to the ones present in cathepsin L1 and cathepsin L2.

The variation in position 157, Val in cathepsin L1 and Leu in cathepsin L2, could account for a more shallow S_2 subsite in cathepsin L1. The residues forming the S_2 subsite in cathepsin L2 are the same as human cathepsin K (Table 6.1). Recently the crystal structure of human cathepsin K complexed with two different inhibitors was resolved (McGrath *et al.*, 1997, Zhao *et al.*, 1997). The S_2 subsite of this enzyme is wide, but more shallow than the corresponding one in cathepsin L (Fujishima *et al.*, 1997). This is due to the presence of a Leu in position 205, while a smaller Ala is present in cathepsin L. Leu^{205} creates a pocket that is too shallow for P_2 Phe, while beta branched residues such as Leu or Val can be more easily accommodated in the ample S_2 pocket. As a Leu residue is also present in both fluke cathepsins, this could explain the preference observed for the substrate Boc-Val-Leu-Lys-NHMec over Z-Phe-Arg-NHMec. The inhibitor APC3328 binds the active site cleft of human cathepsin K in the same orientation as substrates, and has a Leu residue in P2 position (McGrath *et al.*, 1997). Two hydrogen bonds are established between the alpha and beta carbons of the substrate Leu and Gly^{66}, while one of the side chain carbons is hydrogen bonded to Leu^{157}, and the other to Tyr^{67} and Leu ^{205}. The presence of a Leu^{67} in cathepsin L1 would not affect substantially the binding of a Leu in P_2, as the hydrogen bond is established with one of the delta carbons of Tyr, a position that has a counterpart in Leu.

The crystal structure of cathepsin K detected an interaction between the oxygen atom of Asn^{158} (papain numbering) and the amino group of the residue occupying the S_1 subsite. The residue at this position is conserved in the different families, being Gly in cathepsin Bs, Asp in cathepsin L, papain and other plant cysteine proteinases, and Asn in the remaining vertebrate cathepsins. The role of the residue at position 158 has been widely discussed. Menard *et al.* (1990) pointed out that this residue was not essential for papain catalytic activity, a view that was contradicted by Katerelos and Goodenough (1996) analysis of caricain. A conserved hydrogen bond between the residue in this position and the amino group of the P_1 residue of the substrate has been detected in the crystal structures of all but cathepsin Bs (Drenth *et al.*, 1976, Katerelos and Goodenough, 1996, Coulombe *et al*, 1996, Fujishima *et al.*, 1996, McGrath *et al.*, 1997, Zhao *et al.*, 1997). Although the high degree of conservation of the residue at position 158 in cathepsin L2 a single mutation A → C has changed it from Asn (AAT) to Thr (ACT). Although this is highly unusual, the position of this residue at the boundaries of the S_1 and S_2 site, and the conserved interaction if any, with the P_1 residue, suggest that this variation is not involved in the particular substrate specificity of cathepsin L2.

7.4 Phylogenetic relationships of cysteine proteinases of the papain superfamily

The analysis of the evolutionary relationships of cysteine proteinases could give useful information for the design of control strategies for parasitic diseases. It is reasonable to assume that the proteases have evolved from a single ancenstral gene to a complex series of paralogous genes acomplishing different functions. Berti and Storer (1995) have established that the most similar eukaryotic cysteine proteinases to the bacterial cysteine endopeptidases C are the bleomycin hydrolases, and also, that the calpains have a distant relative in an enzyme from *Porphyromonas gingivalis*.

Evidence for an early origin of papain-like cysteine proteinases in eukaryotes is provided by the presence of members of the superfamily in diplomonads and trichomonads. These two taxa are considered between the earliest diverging eukaryotes, a feature evidenced by their absence of mitochondria (Sogin, 1991, Knoll, 1992). The three cysteine proteinases purified from *Giardia* are closely related to the vertebrate cathepsin B (Ward *et al*., 1997), while the genes isolated in *Trichomonas vaginalis* and *Tritichomonas foetus* are more closely related to the vertebrate cathepsin Ls (Mallinson *et al*., 1994, 1995). The presence of two different but related families of cysteine proteinases in early eukaryotes is an indication of an ancient duplication in the evolution of the superfamily. It is premature to establish if this duplication preceded the eukaryote / prokaryote divergence, but from the homologies shown by the genes from amitochondriate organisms with genes from other eukaryotes, it is clear that they originated very early in evolution, generating the two main branches of papain-like proteinases present in the mitochondrial eukaryote lineage.

Ward *et al*. (1997) postulate that the *Giardia* genes are the first branch of the cathepsin B family, ancestral to both cathepsin Bs and cathepsin Cs. In our analysis the *Giardia* genes branch before the cathepsin B genes, but after the dipeptidyl peptidases. However, in the previous report, only the region corresponding to the mature enzyme was considered, while we took into account the complete coding sequence. We decided to use this strategy due to the high homology detected in the propeptides within the two main groups of papain-like proteinases. No major differences in the resulting trees topologies were detected when considering the complete coding sequences (Fig. 6.1) and the mature regions (6.2) in our analysis, with variation only in the less defined nodes. Furthermore, the inclusion of the propeptides in the analysis proved to be useful in resolving the relationships of the cysteine proteinase genes from *Dictiostelium discoideum* with their plant and metazoan counterparts (see below). According to the analysis of Ward *et al.* (1997), cathepsin Cs precedes the divergence of *Leishmania* cathepsin B, indicating that is plausible to find homologous to this enzyme in all mitochondriate eukaryotes. Our analysis based on the complete coding sequences and in partial sequences indicate that the *Giardia* genes diverged after the dipeptidyl peptidases, opening the possibility of discovering cathepsin C-like genes in amitochondriate organisms.

The cathepsin B-like genes of *Giardia* do not contain the fragment corresponding to the occluding loop. The loop is also missing in the plant homologues to cathepsin B, while a reduced loop is present in some of the *C. elegans* cathepsin B-like genes. However, in other nematodes (including other genes from the same *C.elegans* family) the insertional loop is present, and is also present in the platyhelminth sequences and in the *Leishmania mexicana* gene (Fig. 6.4). Recently, Illy *et al.* (1997) have demonstrated the role of the occluding loop in the exopeptidase activity of human cathepsin B by deleting this region. The resulting enzyme, has endopeptidase activity similar to the wild type enzyme, but lacks completely the exopeptidase activity, has a slower autoprocessing to the mature form, probably related to a tighter binding of the propeptide than the one observed in the native enzyme. It is reasonable to suspect the lack of the exopeptidase activity in those genes where the occluding loop is missing, shortened, or substituted in the main residues for this type of activity. However, is difficult to assess if this region and the enzymatic activity related to it is ancestral, or if it originated later in evolution. The absence of the loop in diplomonads and plants, that are the first diverging groups within the family, support the idea of the absence of this region in the ancestral gene; conversely, the presence of the loop in *L. mexicana* suggest an origin of the loop at least as early as the first mitochondrial eukaryotes.

The poor resolution observed outside the cathepsin B class is probably due to a period of rapid duplication and divergence. Cysteine proteinases are present in several protist groups, often in multiple copies. The relationships of the cysteine proteinases of protists with metazoan or plant homologues are not clear, since they tend to separate at the deeper branches of the tree. Moreover, the genes from the different protist lineages find their closest relatives within each lineage, indicating that duplication and diversification took place independently in several protist taxa, and suggesting that several of these cysteine proteinases form distinct groups not to be found anywhere else in the evolution of the superfamily. Furthermore, the early evolutionary history of the cysteine proteinases is obscured by sequence sampling bias and unequal rate of evolution. The majority of the cysteine proteinases from protists and invertebrates are from parasitic organisms, and are involved in the interaction with the host , but no information is available on other proteinase genes related to housekeeping functions. This partiality is also present in the plant sequences, as the majority of the sequences are derived from studies of senescence or seed development. Secondly, different substitution rates in different branches can affect the tree topology (Swofford and Olsen, 1990), a problem detected in cysteine proteinase evolution by Berti and Storer (1995).

A well defined clade including protist, metazoan and plant genes, constitutes the cruzipain class, constituting the first diverging class in the non-cathepsin B cysteine proteinases. The presence of cysteine proteinases from kinetoplastids , the heterolobosean amoeba *Naegleria fowleri,* and the slime mold *Dictiostelium discoideum* in the cruzipain group is indicative that this class preexists the divergence of this protist taxa. This has been noticed by Berti and Storer (1995), who suggested that similar genes might exist in all eukaryotes that diverged after these protist groups. A confirmation of this is provided by the presence of enzymes from plants and the inclusion of the "cathepsin L1" genes from *Schistosoma mansoni* and *S. japonicum,* and the neutral proteinase from *Paragonimus westermani* as metazoan examples of the group.

Only the cathepsin B, cruzipain and cathepsin H classes contain genes of metazoan and plant origin, indicating that at least three different genes were present at the time of divergence of viridiplantae and metazoa, and consequently orthologous genes to these should be present in all plants and metazoans.

The situation of the cathepsin H class is ambiguous. While representatives of plant and vertebrate origin are present in this group indicating an early divergence, no enzymes from other groups are included in the class, and no clear correlation with other families or individual genes is obtained. Berti and Storer (1995) positioned this group as branching before the cathepsin L class, but without significant confidence. Ward *et al.* (1997) on the other hand, cluster the cathepsin H class with the papain class, but no confidence values were provided in this study. We detected that this class is poorly resolved, resulting in different topologies of the tree (with low confidence levels), when different regions are considered (i.e. complete coding region, and mature enzyme). However, while the cathepsin H class has plant and metazoan representatives, only enzymes from plants are included in the papain class, while exclusively metazoan genes constitute the cathepsin L class. The absence of gene sequences from other groups in these classes could be due to a sampling bias, or could represent a true picture of the relationships between the existing enzymes. If this was the case, is tempting to speculate that a cathepsin H-like gene gave rise to the different classes present exclusively in plants and metazoans.

The cysteine proteinases of the trichomonads, *Entamoeba* and *Dictyostelium* tend to cluster together in all the cases, but with low confidence values (Fig. 6.6). This cluster is poorly resolved in relation to the other classes and families, with existing different possible topologies. However, in most of the cases this node appears as branching before the cathepsin H and cathepsin L classes (Fig. 6.1.B and 6.2.B) and in others before the cathepsin H, cathepsin L and papain classes. Although this has no statistical support it is very suggestive; the metazoan and plant relatives of the amitochondriate trichomonads are not clear, but an interesting conjecture is to relate these enzymes with the cathepsin H, cathepsin L and papain classes.

An interesting variation was observed in the position of the genes of the slime mold *D. discoideum*. When only the mature enzyme or conserved regions were analyzed, the *Dictiostelium* enzyme clustered within the cathepsin L class, in accordance to what has been reported previously (Berti and Storer, 1995, Ward *et al.*, 1997) (Fig. 6.2, arrow). However, if the whole coding region was taken into account the slime mold gene separated at a deeper branch closer to *Entamoeba* genes and the diplomonads (Fig. 6.1, arrow, Fig. 6.6). This was not dependent on the long insertion at position 179 (cathepsin L numbering, 164 in papain numbering) previous to the active site asparagine residue, as the analysis of the complete mature enzyme or only conserved regions (excluding the insertion) gave similar results. An insertion in a similar position exists in the *Plasmodium* genes. The addition of the propeptide in the analysis seems to be responsible for this behavior. Although the alignment of the propeptides of several members of the class is possible, the *Dictiostelium* cysteine proteinase 2 propeptide is more similar to the proregion of the members of the papain class. This could be the reason for it positioning in a deeper branch when the whole coding region is considered. Recently two novel cysteine proteinases from the slime mold were cloned (Souza *et al.*, 1995). These enzymes (DICDI_CP4 and DICDI_CP5) have a long serine-rich domain, a target for phospho-glycosylation, in the same position of the DICDI_CP2 gene. These two sequences clustered with very high confidence values with DICDI_CYS2, and the three enzymes together showed the same behaviour to the latter, when the whole coding region , mature only or propeptide regions were considered (Fig. 6.6).

At least four different papain-like cysteine proteinase genes existed by the time the first bilateral metazoans emerged. This is evidenced by the presence of representatives of four different families, belonging to three classes in *Schistosoma*, indicating that orthologous genes are present in all bilateral taxa. Although cathepsin L-like enzyme activities have been detected in nematodes, the majority of the genes isolated correspond to cathepsin B-like enzymes arranged in multigene families. Brinkworth *et al.* (1996) provided an explanation for this contradiction by a structural analysis of the cathepsin B-like enzyme from *Ancylostoma caninum,* detecting residue changes that could account for the observed activity. More recently an enzyme has been cloned from Toxocara canis (TOXCA_CATL) that is more similar to the other classes than to cathepsin Bs. A very similar sequence has been obtained from an expressed sequence tag (EST) of *C.elegans* (CAEEL_R07e). These enzymes show a low level of homology to the characterized families, indicating that possibly they constitute a new group of cysteine proteinases. Furthermore, a search into the EST databanks generated several partial sequences from nematodes with similarities to members of the cathepsin L and cruzipain classes, confirming the presence of members of these classes in nematodes. While other cysteine proteinases might be present, it is clear that parasitic nematodes are cathepsin B-like rich, and at least some of these enzymes play major roles in the interaction with their hosts.

While partial sequences for two cathepsin B-like genes had been obtained in *F. hepatica*, the remaining cloned genes belong to a family that separated before the divergence of the mammalian cathepsin Ls, cathepsin S and cathepsin K. Genes related to these have been isolated from *S. mansoni* and *S. japonicum*, and from the cestodes *Spirometra erinacei* and *Spirometra mansonoides*. Based on the activity towards synthetic substrates these genes were generally regarded as cathepsin Ls, and in fact, we denominated the genes isolated in a previous study (Roche *et al.*, 1997) and in this study as cathepsin L1 and cathepsin L2. However, the distances of both sequences to the vertebrate cathepsin L, cathepsin S and cathepsin K are similar (Table 6.1), consistent with a later divergence of the vertebrate enzymes. Several genes from arthropods branched after the flatworm sequences, but before the three vertebrate families.

Based on comparisons of the distances between cathepsins L and S from rat and human to an outgroup (DICDI_CYS2), Berti and Storer (1995) proposed that cathepsin S has undergone faster evolution, consistent with cathepsin L being the original enzyme and cathepsin S being originated from a more recent duplication. In their analysis no cathepsins K were included and only the mature regions were considered. When cathepsins K are included, they constitute a well defined node with the cathepsins S. This would indicate a common origin by a recent duplication. The cluster S-K is significantly related to the cathepsins L if we consider the whole coding sequence. However, when only the mature portions of the genes are considered, the arthropod cysteine proteinases join in an internal node to the cathepsins L, while the cathepsins S and K remain in a tight second node. As the selective pressures are not homogeneous along the protein, it is obvious that the mature regions would be more conserved. The clustering of the arthropod mature enzymes with vertebrates cathepsins L could indicate that this is in fact the original enzyme, and a duplication lead to the cathepsin S and K cluster. The presence of a cathepsin L from zebrafish (EMBL Y08321), and a cathepsin S like gene from carp (CARP_CPRO) indicate that this duplication took place early in the vertebrate lineage. In this second group, it is difficult to establish which was the original enzyme. The distances measured to an outgroup gave similar results although slightly higher for cathepsin K (see table 6.1). Human cathepsin S and K are encoded by single genes located very closely in chromosome 1 (Gelb *et al.*, 1997, Rood *et al.*,1997). Furthermore, two partial sequences from the trout show very high homology to cathepsin L and cathepsin K respectively. This data taken together indicate that cathepsins L, S and K have diverged at least when the first Osteichtyes appeared. As cathepsin K (also known as cathepsin O2) is produced by the osteoclasts, the duplication that gave origin to this enzyme might have occur just before bone tissue originated, probably at the origin of the bony fishes. Based on this data, we propose that two rapid duplications gave origin to the cathepsin S family first, and almost immediately a second duplication generated the cathepsin K family.

According to the above evidence, the flatworm genes are the first in a line that included the digestive enzymes of decapods and other arthropod sequences, and resulted, as a consequence of recent duplications, in at least three different genes in the vertebrate genome. In this situation the invertebrate genes should not be formally considered as "cathepsin L" as they are as related to them as to the cathepsins K or cathepsins S. The process of gene duplication and diversification is a constant in protein evolution, by which different enzymes emerge and acquire new functions (Neurath, 1984, Creighton and Darby, 1989). Evidence of this process in the cysteine proteinases is everywhere, but probably the clearer examples, to mention just a few, are, the *Entamoeba* genes, the cruzipain multigenic family, the several papaya enzymes, the nematode cathepsin B families, the *F. hepatica* family, the decapod digestive proteases, and the members of the cathepsin L class in vertebrates (namely cathepsins L, S and K). The high degree of relatedness of the cysteine proteinases in some taxa suggest that the corresponding genes arose as a result of duplications after the divergence of these taxa, resulting in paralogous genes. The *F. hepatica* cathepsin L-like genes are just one example of this. While the different cathepsin L-like proteinases of liver flukes are paralogous, they represent a line of genes that is orthologous to the vertebrate cathepsin Ls, cathepsin S and cathepsin K. While at least two different types of papain-like cysteine proteinases were present in early eukaryotes, new proteinases evolved from that basic pattern very early in the eukaryotic lineage, being at least four early in the metazoan evolution, and resulting in more than ten different types in vertebrates. However, while the vertebrate genes are encoded by single copy genes, the existence of multicopy gene families seem to be common in invertebrates and in some protists. This is the case in some of the cysteine proteinases of Trichomonads, *Entamoeba hystolitica,* the cysteine proteinases B of *Leishmania mexicana* (LEIME_LCPB), the cruzipains, some of the cathepsin B-like enzymes of the nematodes *Haemonchus contortus*, *Caenorhabdithis elegans*, and one of the digestive proteinases of the American lobster.

While several of the protist cysteine proteinases occur in lysosome-like organelles, other are released extracellularly (Robertson *et al.*, 1997). This indicates that the non-cytoplasmic cysteine proteinases followed the two main roads in the vesicular pathway since very early in the eukaryotic evolution. The majority of the cysteine proteinases of the papain superfamily from invertebrates are localized in the digestive tract, and there is evidence that several of them are secreted. However, there is a notorious representation bias in the existing sequences. Most of the cloned enzymes are from parasitic organisms where proteolytic activity has been previously demonstrated as an important effector in the interaction with the host, suggesting that other proteins members of the family involved in more general metabolic processes might have been overlooked. In the three invertebrate phyla considered, other patterns of expression are also detected, most significantly in the reproductive tract. In most of the cases the antibodies used to isolate screen cDNA libraries were raised against purified proteinase activities from tissue extracts or excretion/secretion proteins. For this reason, is difficult to establish if the multiple tissue localizations are due to a single proteinase or to multiple, but closely related enzymes.

However, cysteine proteinases seem to have played an important role in nutrition at least in platyhelminths, one of the first organisms with digestive systems. A similar localization might be indicative of similar roles in nematodes, decapods and insects. In *Drosophila melanogaster* embryos, cysteine proteinase expression in the midgut begins when the yolk has been enclosed by the midgut (Matsumoto *et al.*, 1995). Moreover, recently cloned cysteine proteinases from the corn pest *Sitophilius zeamais* localize to the digestive tract, consistent with previous observations that phyto-cystatins can inhibit the hydrolysis of protein substrates of midgut extracts from this coleopteran (Matsumoto *et al.*, 1997).

Although the vertebrate cathepsins are mainly lysosomal, they have been found extracellularly in several normal and pathological processes. However, in the vertebrate examples products of the same genes are targeted to the lysosomes or secreted, as part of a complex transcriptional and post-translational regulatory mechanism (Elliot and Sloane, 1997). The role of glycosylation in the sorting of the vertebrate enzymes to the lysosome is well documented (Kane, 1993, Tao *et al.*, 1994). Interestingly, the secreted cathepsin L-like genes of flatworms and arthropods have no N-glycosylation signals. However , the cruzipain-like genes of *Schistosoma,* and the cathepsin B-like genes from trematodes and nematodes are secreted, regardless of the presence of N-glycosylation signals. The importance of this modification in the proper targeting of the enzymes to their rightful destination still remains to be cleared.

A common feature between invertebrates and vertebrates is that when secreted, the cysteine proteinases seem to be secreted as proenzymes, and the activation step is extracellular. As the propeptides have been indicated as potent specific inhibitors of the mature enzyme, a comparative analysis of the propeptides was performed. A general conservation of residues was detected, and correlated with the secondary structure features described for the three crystallized proenzymes, namely procathepsin B, procathepsin L and procaricain (Cygler *et al.*, 1996, Turk *et al.*, 1996, Coulombe *et al.*, 1996, Groves *et al.*, 1996). This is a strong indication that the general mode of interaction of the propeptide with the mature enzyme is conserved in all the papain superfamily. Residue variation was found in these general conserved patterns, and this variation generally correlated with the diverse families and classes described. However, the level of conservation was lower than the one observed in the mature enzyme, suggesting that the selective pressures on the propeptides are more relaxed than in the catalytic domains. Nonetheless, due to the vital role that the propeptide seems to play in the proper folding and sorting of the proenzyme, the possibilities of variation over a general fold appear to be restricted. Significant mutations would result in loss of the enzymatic activity due to misfolded or ectopically active enzyme, either of which would be lost during evolution. On the contrary, mutations in the catalytic domain might result in enzymes with altered properties, which could still be folded and sorted by the pro-region. The altered mature enzyme, however, would not be inhibited as efficiently by its proregion, until variations in the latter domain reach a new equilibrium with its cognate enzyme. An example of this process has been provided by the analysis of the inhibitory properties of the papaya proteinases; the glycyl endopeptidase propeptide has not lost the ability to inhibit papain, chymopapain or caricain, but it has evolved to be a better inhibitor of its own enzyme (Taylor *et al.*, 1995). This implies that the probabilities of a propeptide inhibiting a different cysteine proteinase are related to the relatedness of the enzymes involved. If this proves to be the case, the propeptides of the parasite cysteine proteinases could be important tools in the design of specific inhibitors, that could block the action of the parasite catalyst without affecting the function of the host enzymes. In this sense, preliminary work in our laboratory indicate that the propeptide of the *F. hepatica* cathepsin L1 is a potent inhibitor of cathepsin L1 and L2, but has little effect on bovine cathepsin B, human cathepsin L and papain. Based on the sequence alignment of the proregions, peptides spanning the suspected region of interaction with the active site has been synthesized and are currently being tested as inhibitors of the parasite enzyme.

An interesting observation when analyzing the propeptides of the diverse cysteine proteinases, is that the carboxy-terminal portion of the long propeptide of the cathepsin Cs has similar conserved blocks more related to the other classes than to the short propeptides of cathepsin Bs. As the dipeptidyl-peptidases are evolutionarily linked to the cathepsin B enzymes, and the whole class seem to have diverged very early in the evolution of the superfamily, this raises the question of how the longer conserved domain, and in fact the whole cathepsin C propeptide has emerged. A more general but intriguing question is the origin of the propeptide in the superfamily, as the closest relatives to the papain superfamily are the cytoplasmic enzymes bleomycin hydrolases and calpains that do not possess the amino extension.

While the majority of the new cysteine proteinase activities detected in parasitic organisms are characterized as "cathepsin B" or "cathepsin L", based on the activity towards Z-Phe-Arg-NHMec and Z-Arg-Arg-NHMec, the enzymes responsible for these activities might not be related to the vertebrate enzymes used as typical examples of the class. Likewise, the cysteine proteinase sequences obtained should be compared against a representative set of sequences from the diverse families before assigning them to any particular group. The different conservation patterns observed in the proregion of the diverse families and classes could be used as a tool for this enzyme characterization process, complemented with sequence comparisons with representatives of the diverse families. A simple set for comparison should include at least one representative of the different classes, namely, cathepsin B, cathepsin H, cathepsin L, papain and cruzipain, and refinement of these could be achieved by a secondary analysis within a family or class, according to the results of the primary analysis. This could probably prevent the existence in the literature of innumerable references to parasitic "cathepsin L"-like enzymes, that are but distantly related to the typical mammalian enzyme.

In summary, we have cloned the second major secreted cathepsin L-like enzyme of the liver fluke *Fasciola hepatica*. The cloned gene was successfully expressed in a yeast system using the trafficking signals contained within the propeptide, resulting in functionally active enzyme. This system of enzyme expression could be employed in the study of other cysteine proteinases which are known to be involved in crucial biological functions of several animal and human pathogens such as schistosomes, hookworms and malaria. The expression of recombinant enzyme is important in clarifying the role of cathepsin L2 as a vaccine candidate against fasciolosis, as sufficient amounts of highly purified recombinant enzyme can be made readily available for further vaccine trials. Furthermore, mutagenesis studies were performed to gain information of the structural and functional peculiarities of the liver fluke enzymes. An understanding of the topology of the active site and the interactions with different substrates is essential in order to design appropriate chemotherapy tools for the control of the disease. However, care should be taken in this latter approach, to target the parasite enzyme without affecting host enzymes that could be related to the target enzyme. To clarify this point a phylogenetic analysis of the papain superfamily was performed, and demonstrated that the liver fluke enzymes

8. References

Aldape,K., H. Huizinga, J. Bouvier, and J.H. McKerrow (1994) *Naegleria fowleri*: characterization of a secreted histolytic cysteine protease. Exp. Parasitol. 78 (2) : 230-241.

Ambrosio, J., A. Landa , M.T. Merchant and J.P. Laclette (1994) Protein uptake by cysticerci of *Taenia crassiceps*. Arch Med Res 25 (3): 325-330.

Andresen, K., T.D.Tom and M. Strand (1991) Characterization of cDNA clones encoding a novel calcium-activated neutral proteinase from *Schistosoma mansoni*. J.Biol.Chem. 266(23):15085-15090.

Asch, H.L. and M.L.Dresden (1979) Acidic thiol proteinase activity of *Schistosoma mansoni* egg extracts. J. Parasitol. 65(4): 543-549.

Barrett, A. J., Kembhavi, A. A., Brown, M. A., Kirschke, H., Knight, C. G., Tamai, M. & Hanada, K. 1982. L-*trans*-epoxysuccinyl-leucylamido(4-guanidino)butane (E-64) and its analogues as inhibitors of cysteine proteinases including cathepsins B, H and L. Biochem. J. 201, 189-198.

Barrett, A.J. and H. Kirschke (1981) Cathepsin B, Cathepsin H and Cathepsin L. Methods in Enzymol. 80: 535-561.

Barrett, A.J. and N. D. Rawlings (1994) Families of cysteine peptidases. Methods in Enzymol. 244 : 461-486.

Barrett, A.J. and N. D. Rawlings (1996) Families and clans of cysteine peptidases. Persp. Drug Disc. and Design , 6: 1-32.

Baylis,H.A., A. Megson, J.C. Mottram and R.Hall (1992) Characterisation of a gene for a cysteine protease from *Theileria annulata*. Mol. Biochem. Parasitol. 54 (1): 105-107.

Becker, M.M., S.A. Harrop, J.P. Dalton, B.H. Kalinna, D.P. McManus and P.J. Brindley (1995) Cloning and characterization fo the *Schistosoma japonicum* aspartic proteinase involved in hemoglobin degradation. J. Biol. Chem. 270:1-6.

Berasain, P., F.Goni, S.McGonigle, A.Dowd, J.P. Dalton, B.Frangione and C. Carmona (1997) Proteinases secreted by *Fasciola hepatica* degrade extracellular matrix and basement membrane components. J. Parasitol. 83 (1): 1-5.

Berti, P. and A.C.Storer (1995) Alignment /Phylogeny of the papain superfamily of cysteine proteinases. J. Mol. Biol. 246:273-283.

Bond, J.S, and E. Butler (1987) Intracellular proteases. Ann. Rev. Biochem. 56:333-364.

Bozas, S.E. and T.W. Spithill (1996) Identification of 3-hydroxyproline residues in several proteins of *Fasciola hepatica*. Exp.Parasitol. 82 (1): 69-72.

Brinkworth, R.I., P.J.Brindley and S.A.Harrop (1996) Structural analysis of the catalytic site of AcCP-1, a cysteine proteinase secreted by the hookworm *Ancylostoma caninum*. Biochem. Biophys. Acta 1298:4-8.

Bromme, D. and K. Okamoto (1995) The baculovirus cystein protease has a cathepsin B-like S2 subsite specificity. Biol. Chem. Hoppe-Seyler 376:611-615.

Bromme, D., K. Bescherer, H. Kirschke and S. Fittkau (1987) Enzyme-substrate interactions in the hydrolisis of peptides by cathepsins B and H from rat liver. Biochem. J. 245 (2):381-385.

Bromme, D., P.R.Bonneau, P. Lachance and A.C. Storer (1994) Engineering the S_2 subsite specificity of human cathepsin S to a cathepsin L and cathepsin B specificity. J. Biol. Chem. 269: 30238-30242.

Bruchhaus, I. , T. Jacobs, M. Leippe and E. Tannich (1996) *Entamoeba hystolitica* and *Entamoeba dispar*: differences in mnumbers and expression of cysteine proteinase genes. Molec. Microbiol. 22 (2): 255-263.

Bryce, S.D., S. Lindsay, A.J. Gladstone, K. Braithwaite, C. Chapman, N.K. Spurr and J. Lunec (1994) A novel family of cathepsin L-like (CTSLL) sequences on human chromosome 10q and related transcripts. Genomics 24 (3): 568-576.

Caffrey, C.R. and M.F. Ryan (1994) characterization of protealytic activity of excretory-secretory products from adult *Strongylus vulgaris*. Vet.Parasitol. 52 : 285-296.

Campetella, O. , J. Henriksson, L. Aslund, A.C. Frasch, U. Patterson and J.J. Cazzulo (1992) The major cysteine proteinase (cruzipain) from *Trypanosoma cruzi* is encoded by multiple polymorphic tandemly organized genes located on different chromosomes. Mol. Biochem. Parasitol. 50 (2): 225-234.

Carmona, C. S. McGonigle, A.J. Dowd, A.M. Smith, S. Coughlan, E. McGowran and J.P. Dalton (1994) A dipeptidylpeptidase secreted by *Fasciola hepatica*. Parasitology 109: 113-118.

Carmona, C., A.J. Dowd , A.M. Smith and J.P. Dalton (1993) Cathepsin L proteinase secreted by *Fasciola hepatica* in vitro prevents antibody-mediated eosinophil attachment to newly excysted juveniles. Mol. Biochem. Parasitol. 62 (1): 9-17.

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

Carter, B.L.A., M. Irani, V.L. Mackay, R.L.Seale, A.V. Sledziewski and R.A. Smith (1987) in DNA cloning (D.M. Glover, ed.) vol. 3, IRL Press, Oxford.

Cazzulo, J.J. (1991) Proteinases of *Trypanosoma cruzi*. In Biochemical Protozoology, G.H.Coombs and M.J.North (Eds.) Taylor &Francis, London - Washington.

Cazzulo, J.J., V. Stoka and V. Turk (1997) Cruzipain, the major cysteine proteinase from the protozoan parasite *Trypanosoma cruzi*. Biol. Chem. 378 (1) : 1-10.

Cejudo, F.J., G. Murphy, C. Chinoy, and D.C. Baulcombe (1992) A giberellin-regulated gene from wheat with sequence homology to cathepsin B of mammalian cells. Plant J. 2 (6):937-948.

Chagas, J.R., M. Ferrer-DiMartino, F. Gauthier and G. Lalmanach (1996) Inhibition of cathepsin B by its propeptide: use of overlapping pepetides to identifiy a critical segment. FEBS Lett. 392 : 233-236.

Chapman, C.B. and G.F. Mitchell (1982) Proteolytic cleavage of immunoglobulin by enzymes released by *Fasciola hepatica*. Vet.Parasitol. 11(2-3): 165-178.

Chapman, R.L., S.E. Kane and A.H. Erickson (1997) Abnormal glycosilation of procathepsin L due to N-terminal point mutations correlates with failure to sort to lysosomes. J. Biol.Chem. 272 (13): 8808-8816.

Chauhan, S.S., N.C. Popescu, D. Ray, R. Fleischmann, M.M. Gottesman and B.R. Troen (1993) Cloning, genomic organization and chromosomal localization of human cathepsin L. J. Biol. Chem. 268:1039-1045.

Chen, Y., C. Plouffe, R. Menard and A.C. Storer (1996) Delineating functionally important regions and residues in the cathepsin B propeptide fcor inhibitory activity. FEBS Lett. 393 :24-26.

Chen,J.-S., A.R. Hays, E.S. Snigirevskaya and Raikhel,A.S. (1995) Mosquito cathepsin B-like thiol protease involved in embryonic degradation of vitellin is produced as a latent extraovarian precursor. Umpublished, GenBank entry L41940.

Chomczynski, P. & Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. 162, 156-159.

Chung, Y.B., Y.Yong, I.J.Joo,S.Y.Cho and S.Y. Kang (1995) Excystment *of Paragonimus westermani* metacercariae by endogenous cysteine protease. J.Parasitol. 81(2):137-142.

Coles,G.C. and D.Rubano (1988) Antigenicity of a proteolytic enzyme of *Fasciola hepatica*. J.Helminthol. 62 (3): 257-260.

Coulombe, R., P. Grochulski, J. Sivaraman, R. Menard, J.S. Mort and M. Cygler (1996) Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. EMBO J. 15 (20): 5492-5503.

Cox, G.N., D. Pratt, R. Hageman and R.J. Boisvenue (1990) Molecular cloning and primary sequence of a cysteine protease expressed by *Haemonchus contortus* adult worms.Mol. Biochem. Parasitol. 41 (1): 25-34.

Cygler, M., J. Sivaraman, P. Grochulski, R. Coulombe, A.C. Storer and J.S. Mort (1996) Structure of rat procathepsin B: model for inhibition of cysteine protease activity by the proregion. Structure 15 (4): 405-416.

Dalton, J.P and M.Heffernan (1989) Thiol proteases released in vitro by *Fasciola hepatica*. Mol.Biochem.Parasitol. 35:161-166.

Dalton, J.P. and P.J. Brindley (1996) Schistosome asparagynil endopeptidase Sm32 in haemoglobin digestion. Parasitol. Today 12 : 125.

Dalton, J.P. and P.J. Brindley (1997) Proteases of Trematodes. In Advances in Trematode Biology (Fried, B. and T.K. Graczyk, eds.) Boca Raton (FL), CRC Press.

Dalton, J.P., L. Hola-Jamriska and P.J. Brindley (1995) Asparaginyl endopeptidase activity in adult *Schistosoma mansoni*. Parasitology 111(5): 575-580.

Dalton, J.P., K.A.Clough, M.K.Jones and P.J.Brindley (1996) Characterization of the cathepsin-like cysteine proteinases of *Schistosoma mansoni*. Infect. Immun. 64 (4): 1328-1334.

Dalton, J.P., K.A.Clough, M.K.Jones and P.J.Brindley (1997) The cysteine porteinases of *Schistosoma mansoni* cercariae. Parasitology 114:105-112.

Dalton, J.P., S.McGonigle, T.P.Rolph and S.J. Andrews (1996) Induction of protective immunity in cattle against infection with *Fasciola hepatica* by vaccination with cathepsin L proteinases and with hemoglobin. Infect. Immun. 64(12): 5066-5074.

Davis, A.H., J. Nanduri and D.C. Watson (1987) Cloning and expression of *Schistosoma mansoni* protease. J.Biol.Chem. 262 (26): 12851-12855.

Day, S.R., J.P.Dalton, K.A.Clough, L.Leonardo, W.U.Tiu and P.J.Brindley (1995) Characterization and cloning of the cathepsin L proteinases of *Schistosoma mansoni*. Biochem. Biophys. Res. Commun. 217(1): 1-9.

Deussing, J., W. Roth, W. Rommerskirch, B. Wiederanders, K. von Figura and C. Peters (1997) The genes of the lysososmal cysteine proteinases cathepsins B, H, L and S map to different mouse chromosomes. Mamm. Genome 8 (4): 241-245.

Dolenc, I. , B. Turk, G. Pungercic, A. Ritonja and V. Turk (1995) Oligomeric structure and substrate induced inhibition of human cathepsin C. J. Biol. Chem. 270 (37):21626-21631.

Dowd, A.J., A.M.Smith, S.McGonigle and J.P.Dalton (1994) Purification and characterization of a second cathepsin L proteinase secreted by the parasitic trematode *Fasciola hepatica*. Eur.J.Biochem. 223:91-98.

Dowd, A.J., J.P. Dalton, A.C. Loukas, P. Prociv and P.J. Brindley (1994) Secretion of cysteine proteinase activity by the zoonotic hookworm *Ancylosotoma caninum*. Am. J. Trop. Med. Hyg. 51: 341-347.

Dowd, A.J., S.McGonigle and J.P.Dalton (1995) *Fasciola hepatica* cathepsin L proteinase cleaves fibrinogen and produces a novel type of fibrin clot. Eur.J.Biochem. 232(1): 241-246.

Drake, L.J. , A.E. Bianco, D.A.P. Bundy and F. Ashall (1994) Characterization of peptidases of adult *Trichuris muris*. Parasitology 109: 623-630.

Drenth, J., K.H.Kalk and H.M.Swen (1976) Binding of chloromethylketone substrate analogues to crystalline papain. Biochemistry 15:3731-3738.

Dresden, M.L., C.K. Sung and A.M. Deelder (1983) A monoclonal antibody from infected mice to a *Schistosoma mansoni* egg proteinase. J. Immunol. 130(1): 1-3.

Dunn, B.M (1989) Determination of protease mechanism. In Proteolytic enzymes: a practical approach. R.J. Beynon & J.S. Bond (Eds.) IRL Press, Oxford.

Elliot, E. and B.F.Sloane (1996) The cysteine protease cathepsin B in cancer. Persp. Drug Disc. and Design , 6: 1-32.

Etges, R. and J. Bouvier (1991) The promastigote surface protein of *Leishmania*. In Biochemical Protozoology, G.H.Coombs and M.J.North (Eds.) Taylor &Francis, London - Washington.

Fagbemi, B.O. and E.E. Guaobadia (1995) Immunodiagnosis of fasciolosis in ruminants using a 28 kDa cysteine protease of *Fasciola gigantica* adult worms. Vet. Parasitol. 57(4): 309-318.

Fagbemi, B.O. and G.V.Hyller (1991) Partial purification and characterization of the proteolytic enzymes of *Fasciola gigantica* adult worms. Vet.Parasitol. 40:217-226.

Fagbemi, B.O. and G.V.Hyller (1992) The purification and characterization of a cysteine protease of *Fasciola gigiantica* adult worms. Vet.Parasitol. 43(3-4):223-232.

Fagotto, F. (1990) Yolk degradation in tick eggs: I. Occurrence of a cathepsin L-like acid proteinase in yolk spheres. Arch. Insect Biochem. Physiol. 14 (4):217-235.

Fagotto, F. (1990b) Yolk degradation in tick eggs : II. Evidence that the cathepsin L-like proteinase is stored as a latent, acid activable proenzyme. Arch. Insect Biochem. Physiol. 14 (4):237-252.

Felsenstein, J, (1989) PHYLIP Phylogeny Inference Package (version 3.2). Cladistics 5: 164-166.

Fong, D., M.M. Chan and W.T. Hsieh (1991) Gene mapping of human cathepsins and cystatins. Biomed. Biochim. Acta 50: 595-598.

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

Fox, T., P. Mason, A.C. Storer, J.S. Mort (1995) Modification of the S_1 subsite specificty in the cysteine protease cathepsin B. Protein Eng. 8:53

Francis, S.E., I.Y. Gluzman, A. Oksman, D. Banerjee and D.E. Goldberg (1996) Characterization of native falcipain, an enzyme involved in *Plasmodium falciparum* hemoglobin degradation.

Gelb, B.D., G.P. Shi, M. Heller, S. Weremowicz, C. Morton, R.J. Desnick and H.A. Chapman (1997) Structure and chromosomal assignment of the human cathepsin K gene. Genomics 41(2): 258-262.

Gonheim, H. and M.Q. Klinkert (1995) Biochemical properties of purified cathepsin B from *Schistosoma mansoni*. Int. J. Parasitol. 25(12):1515-1519.

Goose, J. (1978) Possible role of excretory/secretory products in evasion of host defences by *Fasciola hepatica*. Nature 275(5677):216-217.

Gorbalenya, A.E, and E.J. Snider (1996) Viral cysteine proteinases. Persp. Drug Disc. and Design, 6: 64-86.

Gotz, B. and M.Q.Klinkert (1993) Expression and partial characterization of a cathepsin B-like enzyme (Sm31) and a proposed 'haemoglobinase' (Sm32) from *Schistosoma mansoni*. Biochem. J. 290 (Pt 3): 801-806.

Groves, M.R., M.A.J. Taylor, M. Scott, N.J. Cummings, R.W. Pickersgill and J.A. Jenkins (1996) The prosequence of procaricain forms an a-helical domain that prevents access to the substrate binding cleft. Structure 15 (4): 1193-1203.

Halton, D.W. (1967) Observations on the nutrition of digenetic trematodes. Parasitology 57(4): 639-660.

Hamajima, F., M.Yamamoto, S.Tsuru, K.Yamakami, T.Fujino, H.Hamajima and Y.Katsura (1994) Immunosuppression by a neutral thiol protease from parasitic helminth larvae in mice. Parasite immunol. 16:261-273.

Harrop, S.A., N. Sawangjaroen, P. Prociv and P.J. Brindley (1995) Characterization and localization of cathepsin B proteinases expressed by adult *Ancylostoma caninum* hookworms. Mol. Biochem. Parasitol. 71 : 163-171.

Harrop, S.A., Prociv, P. and P.J.Brindley (1995) Amplification and characterization of cysteine proteinase genes from nematodes. Trop. Med. Parasitol. 46 (2) :119-122.

Hasnain, S., T. Hirama, C.P. Huber, P. Mason and J.S. Mort (1993) Characterization of cathepsin B specificity by site-directed mutagenesis: importance of Glu^{245} in the S_2-P_2 specificity for arginine and its role in transition state stabilization. J. Biol. Chem. 268:235-240.

Hawthorne, S.J., D.W. Halton and B. Walker (1994) Identification and characterization of the cysteine and serine proteinases of the trematode *Haplometra cylindracea* and determination of their hemoglobinase activity. Parasitology 108:595-601.

Heussler, V.T. and D.A.E.Dobbelaere (1994) Cloning of a protease gene family *of Fasciola hepatica* by the polymerase chain reaction. Mol.Biochem.Parasitol. 64:11-23.

Hill, D.E. and J.A. Sakanaki (1997) *Trichuris suis*: Thiol protease activity from adult worms. Exp. Parasitol. 85(1): 55-62.

Hola-Jamriska, L., J.F. Tort, J.P. Dalton, S.R. Day, J. Fan, J. Aaskov and P.J.Brindley (1997) Schistosoma japonicum cathepsin C. cDNA encoding the preproenzyme, biochemical analysis and phylogenetic relationships. J. Biol. Chem. Submitted.

Homma, K. , S. Kurata and S. Natori (1994) Purification, characterization and cDNA cloning of procathepsin L from the culture medium of NIH-Spae -4, an embrionic cell line of *Sarcophaga peregrina* (flesh fly), and its involvement in the differentiation of imaginal discs. J. Biol. Chem. 269 (21): 15258-15264.

Howell, R.M. (1966) Collagenase activity of immature *Fasciola hepatica*. Nature 209(24):713-714.

Howell, R.M. (1973) Localization of proteolytic activity in *Fasciola hepatica*. J. Parasitol. 59(3): 454-456.

Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. B 256, 119-124

Hyde, J.E. (1990) Molecular Parasitology. Open University Press, UK.

Ishidoh, K. and E. Kominami (1994) Multi-step processing of procathepsin L in vitro. FEBS Lett. 352 : 281-284.

James , M.N.G. (1980) An X-ray crystallographic approach to enzyme structure and function. Can. J. Biochem. Can J Biochem 58(4): 252-271cited by Neurath, H. (1984) Evolution of proteolytic enzymes, Science, 224, 350-357.

Jia, Z., S. Hasnain, T. Hirama, X. Lee, J.S. Mort, R. To and C.P. Huber (1995) Crystal structures of recombinant rat cathepsin B and a Cathepsin B-Inhiitor complex. J. Biol. Chem 270: 5527-5533.

Kamata, I., M. Yamada, R. Uchinkawa, S. Matsuda and N. Arizono (1995) Cysteine protease of the nematode *Nippostrongylus brasiliensis* preferentially evokes an IgE/IgG1 antibody response in rats. Clin. Exp. Immunol. 102(1): 71-77.

Kane, S.E. (1993) Mouse procathepsin L lacking a functional glycosilation site is properly folded, stable and secreted by NIH 3T3 cells. J. Biol. Chem. 268 (15): 11456-11462.

Kang, S.Y., M.S. Cho, Y.B. Chung, Y.Kong and S.Y.Cho (1995) A cysteine protease of *Paragonimus westermani* eggs. Korean J.Parasitol. 33(4):323-330.

Karanu, F.N., F.R. Rurangirwa, T.C. McGuire and D.P. Jasmer (1993) *Haemonchus contortus*: identification of proteases with diverse characteristics in adult worm excretory-secretory products.

Karrer, K.M., S.L. Peiffer and M.E. DiTomas (1993) Two distinct gene subfamilies within the family of cysteine protease genes. Proc. Natl. Acad.Sci. USA 90 (7): 3063-3067.

Khouri, H.E, T. Vernet, R. Menard, F. Parlati, P. Laflamme, D.C. Tessier, B. Gour-Salin, D.Y. Thomas and A.C. Storer (1991) Engineering of papain: Selective alteration of substrate specificity by site-directed mutagenesis. Biochemistry 30: 8929-8936 .

Kimura, M. (1983) The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, England.

Klinkert, M.Q., D.Cioli, E. Shaw, V.Turk, W. Bode and R. Butler (1994) Sequence and structure similarities of cathepsin B from the parasite *Schistosoma mansoni* and human liver. FEBS Lett. 351: 397-400.

Klinkert, M.Q., R. Felleisen, G. Link, A. Ruppel and E. Beck (1989) Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. Mol.Biochem. Parasitol. 33 :113-122.

Knapp, B. E. Hundt, U. Nau and H.A. Kupper (1989) Molecular cloning, genomic structure and localization in a blood stage antigen of *Plasmodium falciparum* characterized by a serine stretch. Mol. Biochem.Parasitol. 32 (1): 73-83.

Knoll, A.H. (1992) The early evolution of eukaryotes: a geological perspective. Science 256:622-627.

Koga, H., H.Yamada, Y.Nishimura, K.Kato and T.Imoto (1990) Comparative study on specificities of rat cathepsin L and papain: amino acid differences at substrate-binding sites are involved in their specificities. J.Biochem. 108:976-982.

Koizumi, M.; K. Yamaguchi-Shinozaki, H. Tsuji,and K. Shinozaki (1993) Structure and expression of two genes that encode distinct drought-inducible cysteine proteinases in *Arabidopsis thaliana*. Gene 129 (2): 175-182.

Kong, Y., Y.B. Chung, S.Y.Cho and S.Y. Yang (1994) Cleavage of immunoglobulin G by excretory-secretory cathepsin S-like protease of *Spirometra mansoni* plerocercoid. Parasitology 109 : 611-621.

Kopitar, G., M. Dolinar, B. Strukelj, J. Pungercar and V. Turk (1996) Folding and activation of human procathepsin S from inclusion bodies produced in *Escherichia coli*. Eur. J. Biochem. 236(2): 558-562.

Kuroda, M. , M. Ishimoto, K. Suzuki, H. Kondo, K. Abe, K. Kitamura and S. Arai (1996) Oryzacystatins exhibit groth-inhibitory and lethal effects on different species of bean insect pests, *Callosobruchus chinensis* (Coleoptera) and *Riptortus clavatus* (Hemiptera). Biosci. Biotech. Biochem. 60: 209-212.

Laemmli, U. K. 1970. Cleavage of structural proteins during the assembly of the head of the bacteriophage T4. Nature. 227, 680-685.

Larminie, C.G.C. and I.L.Johnstone (1996) Isolation and characterization of four developmentally regulated cathepsin B-like cysteine protease genes from the nematode *Caenorhabditis elegans*. DNA Cell Biol. 15(1): 75-82.

Laycock, M.V., R.M. MacKay, M. Di Fruscio and J.W. Gallant (1991) Molecular cloning of three cDNAs that encode cysteine proteinases in the digestive gland of the american lobster (*Homarus americanus*). FEBS Lett. 292 (1-2): 115-120.

Leatherbarrow, R. J. 1987, Enzfitter, Elsevier Biosoft, Cambridge.

LeBoulay, C., A. van Wormhoudt and D. Sellos (1995) Molecular cloning and sequencing of two cDNAs encoding cathepsin L-related cysteine proteinases in the nervous system and in the stomach of t he norway lobster (*Nephrops norvegicus*). Comp. Biochem. Physiol. 111B: 353-359.

LeBoulay, C., A. van Wormhoudt and D. Sellos (1996) Cloning and expression of cathepsin L-like proteinases in the hepatopancreas of the shrimp *Penaeus vannamei* during the intermolt cycle. J.Comp Physiol 166: 310-318.

Li, W.H. and Graur, D. (1991) Fundamentals of molecular evolution, Sinauer Associates Inc. Publishers, Sunderland, MA, USA.

Lilley , C.J., P.E. Urwin, M.J. McPherson and H.J. Atkinson (1996) Characterization of intestinally active proteinases of cyst-nematodes. Parasitology 113:415-424.

Linnevers, C., S.P. Smeekens and D. Bromme Human cathepsin W, a putative cysteine protease predominantly expressed in CD8+ T-lymphocytes. FEBS Lett 405 (3): 253-259.

Lipps, G., R. Fullkrug and E.Beck (1996) Cathepsin B of *Schistosoma mansoni*. Purification and activiation of the recombinant proenzyme secreted by *Saccharomyces cerevisiae*. J.Biol.Chem. 271(3):1717-1725.

Liu, D.W., H. Kato, T. Nakamura and K. Sugane (1996) Molecular cloning and expression of the gene encoding a cysteine proteinase of *Spirometra erinacei*. Mol. Biochem.Parasitol. 76 (1-2) : 11-21.

Liu, X., R.C. McCarron and J.H. Nordin (1996) A cysteine protease that processes insect vitellin. Purification and partial characterization of the enzyme and the proenzyme. J. Biol. Chem. 271 (52): 33344-33351.

Londsdale-Eccles, J.D. (1991) Proteinases of African trypanosomes. In Biochemical Protozoology, G.H.Coombs and M.J.North (Eds.) Taylor &Francis, London - Washington.

Loukas, A. and R.M.Maizels (1997) Cathepsin L-like cysteine proteinase from *Toxocara canis* infective larvae.Genebank entry accession U53172.

Lustigman, S., J.H. McKerrow, K.Shah, J. Lui, T.Huima, M.Hough and B. Brotman (1996) Cloning of a cysteine protease required for the molting of *Onchocerca volvulus* third stage larvae. J.Biol.Chem. 271 (47):30181-30189.

Mach, L., H. Schwihla, K. Stuwe, A.D. Rowan, J.S. Mort and J. Glossl (1993) Activation of procathepsin B in human hepatoma cells: The consversion into the mature enzyme relies on the action of cathepsin B itself. Biochem. J. 293 (2): 437-442.

Mach, L., J.S. Mort and J. Glossl (1994a) Maturation of human procathepsin B. Proenzyme activation and proteolytic processing of the precursor to the mature proteinase, in vitro, are primarrily unimolecular processes. J.Bio.Chem. 269 (17): 13030-13035.

Mach, L., J.S. Mort and J. Glossl (1994b) Noncovalent complexes between the lysosomal proteinase cathepsin B and its propeptide account for stable, extracellular, high molecular mass forms of the enzyme. J. Biol. Chem. 269 (17):13036-13040.

Maki, J. and T. Yanagisawa (1986) Demosntration of carboxyl and thiol protease activities in adult *Schistosoma mansoni, Dirofilaria immitis, Angiostrongylus cantonensis*. J. Helminthol 60(1):31-37.

Mallison, D.J., B.C. Lockwood, G.H.Coombs and M.J. North (1994) Identification and molecular cloning of four cysteine proteinase genes from the pathogenic protozoon *Trichomonas vaginalis*. Microbiology 140: 2725-2735.

Mallison, D.J., J. Livingstone, K.M. Appleton, S.J. Lees, G.H. Coombs and M.J. North (1995) Multiple cysteine proteinases of the pathogenic protozoon *Tritrichomonas foetus*: identification of seven diverse and differentially expressed genes. Microbiology 141: 3077-3085.

Martinez, J., J. Henriksson, M. Rydaker, J.J. Cazulo and U. Petterson (1995) Genes for cysteine proteinases from *Trypanosoma rangeli*. FEMS Microbiol. Lett. 129 (2-3): 135-141.

Matsumoto, I., H. Watanabe, K. Abe, S. Arai and Y. Emori (1995) A putative digestive cysteine proteinase from *Drosophila melanogaster* is predominantly expressed in the embryonic and larval midgut. Eur. J. Biochem. 227: 582-587.

Matsumoto, I., Y. Emori, K. Abe, and S. Arai (1997) Characterization of a family encoding cysteine proteinases of *Sitophilus zeamais* (maize weevil), and analysis of the protein distribution in various tissues including alimentary tract and germ cells. J. Biochem. 121: 464-476.

Mc Kerrow, J.H., M.E. Mc Grath, and J.C. Engel (1995) The cysteine proteinase of *Trypanosoma cruzi* as a model for antiparasite drug design. Trends Parasitol. 11 (8) 279-281.

McDonald, J.K. and J.M. Emerick (1995) Purification and characterization of procathepsin L, a self-processing zymogen of guinea pig spermatozoa that acts on a cathepsin D assay substrate. Arch. Biochem.Biophys. 323 (2): 409-422.

McIntyre, G.F. and A.H. Erickson (1991) Procathepsins L and D are membrane-bound in acidic microsomal vesicules. J.Biol. Chem. 266 (23):15438-15445.

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

McIntyre,G.F., G.D. Godbold and A.H. Erickson (1994) the pH dependent membrane association of procathepsin L is mediated by a 9 residue sequence within the propeptide. J. Biol. Chem. 269 (1): 567-572.

Medina , M. P. Leon and C.G. Vallejo (1988) *Drosophila* cathepsin B-like proteinase: a suggested role in yolk degradation. Arch. Bniochem. Biophys. 263 (2) : 353-363.

Menard, R., E. Carmona, C. Plouffe, D. Bromme, Y. Konishi, J. Lefebre and A.C. Storer (1993) The specificity of the S_1' subsite of cysteine proteases. FEBS Lett 328 (1-2): 107-110.

Merckelbach, A., S.Hasse, R. Dell, A. Eschlbeck and A. Ruppel (1994) cDNA sequences of *Schistosoma japonicum* coding for two cathepsin B-like proteins and Sj32. Trop. Med. Parasitol. 45 (3): 193-198.

Michel, A., H.Goneim, M.Resto, M.Q.Klinkert and W.Kuntz (1995) Sequence, characterization and localization of a cysteine proteinase cathepsin L in *Schistosoma mansoni*. Mol. Biochem. Parasitol. 73:7-18.

Mikkonen, A. , I. Porali, M. Cercos, and T.H. Ho (1996) A major cysteine proteinase, EPB, in germinating barley seeds: structure of two intronless genes and regulation of expression. Plant Mol Biol. 31 (2): 239-254.

Mitro, K., A. Bhagavathiammai, O.M. Bobbett, J.H. McKerrow, R.Chokshi, B. Chokshi and E.R. James (1994) Partial characterization of the proteilytic secretions of *Acanthamoeba polyphaga*. Exp. Parasitol. 78 (4) : 377-385.

Miyata, S. and H.K. Kihara (1995) Cathepsin L-like protease from *Xenopus* embryos that is stimulated by nucleoside phosphates and nucleic acids. Zoolog. Sci. 12 (6): 771-774.

Moczon, T. (1994a) A cysteine proteinase in the cercariae of *Diplostomum psuedopathaceum* (Trematoda, Diplostomatidae). Parasitol.Res. 80(8): 680-683.

Moczon, T. (1994b) Histochemistry of proteinases in the cercariae of *Diplostomium pseudopathaceum* (Trematoda, Diplosotomidae). Parasitol. Res. 80: 684-686.

Moir, D.T. and L.S. Davidow (1991) Production of proteins by secretion from yeast. Methods. Enzymol. 194: 491-507.

Morris, J.R. and J.A. Sakanaki(1994) Characterizations of the serine protease and the serine protease inhibitor from the tissue penetrating nematode *Anisakis simplex*. J. Biol. Cxhem. 44: 27650-27656.

Mottram, J.C., M.J. Frame, D.R. Brooks, L. Tetley, J.E. Hutchinson, A.E. Souza and G.E. Coombs (1997) The multiple cpb cysteine proteinase genes of *Leishmania mexicana* encode isoenzymes that differ in their stage regulation and substrate preferences. J. Biol. Chem. 272 (22): 14285-14293.

Muller-Ladner, U., R.E.Gay and S. Gay (1996) Cysteine proteinases in arthritis and inflamation. Persp. Drug Disc. and Design , 6: 1-32.

Musil, D., D.Zucic, D.Turk, R.A.Engh, I. Mayr, R.Huber, T.Popovic, V.Turk, T.Towatari, N.Katunuma and W.Bode (1991) The refined 2.15 A X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity. EMBO J. 10:2321-2330.

Nene,V., E. Gobright, A.J. Musoke and J.D. Lonsdale-Eccles (1990) A single exon codes for the enzyme domain of a protozoan cysteine protease. J. Biol. Chem. 265 (30): 18047-18050.

Nene,V., K.P. Iams, E. Gobright A.J. Musoke (1992) Characterisation of the gene encoding a candidate vaccine antigen of *Theileria parva* sporozoites. Mol. Biochem. Parasitol. 51 (1): 17-27.

Neurath, H. (1984) Evolution of proteolytic enzymes, Science, 224, 350-357.

Nicholas, K.B, H.B. Nicholas and D.W. Deerfield (1997) GeneDoc: Analysis and Visualization of Genetic Variation . Embnet.news 4 : 2 (http://www2.ebi.ac.uk/embnet.news/).

Nishimura, Y. and K. Kato (1987a) Intracellular transport and processing of lysosomal cathepsin B. Biochem. Biophys. Res Commun. 148 (1):254-259.

Nishimura, Y. and K. Kato (1987b) Intracellular transport and processing of lysosomal cathepsin H. Biochem. Biophys. Res Commun. 148 (1):329-334.

Nishimura, Y. and K. Kato (1988) Identification of latent procathepsin H in microsomal lumen:characterization of proteolytic processing and enzyme activation. Arch. Biochem. Biophys. 260 (2):712-718.

Nishimura, Y. T. Kawabata and K. Kato (1988) Identification of latent procathepsins B and L in microsomal lumen: characterization of enzymatic activation and proteolytic processing in vitro. Arch. Biochem. Biophys. 261 (1):64-71.

Nishimura, Y., K. Kato, K. Furuno and M. Himeno (1995) Inhibitory effect of leupeptin on the intracellular maturation of lysosomal cathepsin L in primary cultures of rat hepatocytes. Biol. Pharm. Bull. 18 (7): 945-950.

North, M.J.(1991) Proteinases of trichomonads and *Giardia.* In Biochemical Protozoology, G.H.Coombs and M.J.North (Eds.) Taylor & Francis, London - Washington.

North, M.J., K.I. scott and B.C. Lockwood (1988) Multiple cysteine proteinase forms during the life cycle of *Dictyostelium discoideum* revealed by electrophoretic analysis. Biochem.J. 254 (1):261-268.

O'Hora, B., A.M. Hemmings, D.J. Buttle and L.H. Pearl (1995) Crystal structure of glycil endopeptidase from *Carica papaya*: A cysteine endopeptidase of unusual substrate specificity. Biochemistry 34:13190-13195.

O'Neill, S. M., M. Parkinson, W. Strauss, R. Angles and J.P. Dalton (1997) Immunodiagnosis of *Fasciola hepatica* infection (Fasciolosis) in a human population in the bolivian altiplano using purified cathepsin L cysteine proteinase. Amm. J. Trop. Med. Hygiene. In press.

Okamura, N., M. Tamba, Y. Uchiyama, Y. Sugita, F. Dacheux, P. Syntin and J.L. Dacheux (1995) Direct evidence for the elevated synthesis and secretion of procathepsin L in the distal caput epididymis of boar. Biochim. Biophys. Acta 1245 (2):221-226.

Ord, T., C. Adessi, L. Wang and H.H. Freeze (1997) the cysteine proteinase gene cprG in *Dictyostelium discoideum* has a serine-rich domain that contains GlcNAc-1-P. Arch.Biochem.Biophys. 339 (1): 64-72.

Page. R. D. M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12: 357-358.

Panaccio,M., C.Hollywell, S.Mailer,, G.L. Wijffels and T.W. Spithill (1994) Identification of a new L-like cathepsin protease of *Fasciola hepatica*. Localisation of the sites of expression of the L-like cathepsin proteases to both the gut microvilli and the Mehlis' gland. Unpublished , GeneBank entry L33772.

Park, H., M.Y.Ko, M.K. Paik, C.T. Soh, J.H. seo and K.I.Im (1995) Cytotoxicity of a cysteine proteinase of adult *Clonorchis sinensis*. Korean J. Parasitol. 33(3) : 211-218.

Pears, C.J., H.M.Mahbubani and J.G. Williams (1985) Characterization of two highly diverged but developmentally co-regulated cysteine proteinase genes in *Dictyostelium discoideum*. Nucl.Acids Res. 13 (24): 8853-8866.

Peitsch, M.C. (1996) PROMOD and SWISS_MODEL: Internet-based tools for automated comparative protein modelling. Biochem.Soc. Trans. 24:274

Phares, K (1996) An unusual host-parasite relationship: the growth hormone-like factor from plerocercoids of spirometrid tapeworms. Int.J.Parasitol. 26 (6) : 575-588.

Phares,K. and J.Kubik (1996) The groth factor from plerocercoids of *Spirometra mansonoides* is both a growth hormone agonist and a cysteine proteinase. J.Parasitol. 82(2):210-215.

Poltzer, M., R.M. Overstreet, and H. Taraschewski (1994) Proteinase activity in the plerocercoid of *Protocephalus ambliplitis*. Parasitol. 109B:209-213.

Polzer, M. and U. Conradt (1994)Identification and partial characterization of the protease from different developmental stages of *Schistocephalus solidus*. Int. J. Parasitol 24:967-973.

Pratt, D. G.N. Cox, M.J. Milhausen and R.J. Boisvenue (1990) A developmentally regulated cysteine protease gene family in *Haemonchus contortus*. Mol. Biochem. Parasitol. 43 (2): 181-191.

Pratt, D. L.G. Armes, R. Hagerman, V. Reynolds, R.J. Boisvenue and G.N. Cox (1992) Cloning and sequence comparisons of four distinct cysteine proteases expressed by *Haemonchus contortus* adult worms. Mol. Biochem. Parasitol. 51(2) : 209-218.

Pratt, D., R.J. Boisvenue and G.N. Cox (1992) Isolation of putative cysteine protease genes of *Ostertagia ostertagi*. Mol. Biochem. Parasitol. 56(1) : 39-48.

Rao N.V., G.V. Rao, and J.R. Hoidal (1997) Human dipeptidyl-peptidase I. Gene characterization, localization, and expression. J Biol Chem 272 (15): 10260-10265.

Ray, C. and J.H. McKerrow (1992) Gut-specific and developmental expression of a *Caenorhabditis elegans* cysteine protease gene. Mol. Biochem. Parasitol. 51 : 239-250.

Redinbaugh, M. G. and Turley, R. B. 1986. Adaption of the bicinchoninic acid protein assay for use with microtitre plates and sucrose gradient fractions. *Anal. Biochem.* 153, 267-271.

Rege, A.A., C.Y. Song, H.J. Bos and M.H. Dresden (1989) Isolation and partial characterization of a potentially pathogenic cysteine proteinase from adult *Dictyocaulus viviparus*. Vet. Parasitol. 34 (1-2) : 95-102 .

Rege, A.A., P.R. Herrera, M.Lopez and M.H. Dresden (1989) Isolation and characterization of a cysteine proteinase from *Fasciola hepatica* adult worms. Mol.Biochem.Parasitol. 35(1):89-95.

Rhoads, M.L. and R.H. Fetterer (1996) Extracellular matrix degradation by *Haemonchus contortus*. J. Parasitol 82(3) : 379-383.

Ribolla P.E. and A.G. De Bianchi (1995) Processing of procathepsin from *Musca domestica* eggs. Insect. Biochem. Mol. Biol. 25 (9): 1011-1017.

Richer, J.K., J.A. Sakanaki, G.R. Frank and R.B. Grieve (1992) *Dirofilaria immitis*: proteases produced by third and fourth stage larvae. Exp. Parasitol. 75 (2): 213-222.

Richer, J.K., W.G. Hunt, J.A. Sakanaki and R.B. Grieve (1993) *Dirofilaria immitis*: effects of fluoromethyl cysteine protease inhibitors on the third to fourth stage molt. Exp.Parasitol. 76(3):221-231.

Roads, M.L. and R.H. Fetterer (1995) Developmentally regulated secretion of cathepsin L-like cysteine proteases by *Haemonchus contortus*. J. Parasitol. 81 (4): 505-512.

Robertson, C.D., G.H. Coombs, M.J. North and J.C. Mottram (1996) Parasite cysteine proteinases. Persp. Drug Discov. Design 6: 33-46.

Roche, L., A.J.Dowd, J.Tort, S.McGonigle, A.McSweeney, G.P.Curley, T.Ryan and J.P.Dalton (1997) Functional expression of *Fasciola hepatica* cathepsin L1 in *Saccharomyces cerevisiae*. Eur.J.Biochem. 245:373-380.

Romanos, M.A., C.A. Scorer and J.J. Clare (1992) Foreign gene expression in yeast: a review. Yeast 8 (6): 423-488.

Rood, J.A., S. van Horn, F.H. Drake, M. Gowen and C. Debouck (1997) Genomic organization and chromosome localization of the human cathepsin K gene (CTSK). Genomics 41(2): 169-176.

Rosenthal, P.J. (1991) Proteinases of malaria parasites. In Biochemical Protozoology, G.H.Coombs and M.J.North (Eds.) Taylor &Francis, London - Washington.

Rosenthal,P.J. (1996) Conservation of key amino acids among the cysteine proteinases of multiple malarial species. Mol. Biochem. Parasitol. 75 : 255-260.

Rosenthal,P.J., C.S. Ring, X. Chen, and F.E. Cohen (1993) Characterization of a *Plasmodium vivax* cysteine proteinase gene identifies uniquely conserved amino acids that may mediate the substrate specificity of malarial hemoglobinases. J. Mol. Biol. 241 (2) : 312-316.

Rowan A.D., P. Mason, L. Mach and J.S. Mort (1992) Rat procathepsin B. Proteolytic processing to the mature form in vitro. J Biol Chem 267(22):15993-15999.

Rowan, A.D., P. Mason, L. Mach and J.S. Mort (1992) Rat procathepsin B. Proteolytic processing to the mature form in vitro. J.Biol.Chem. 267 (22):15993-15999.

Ruppel, A., U. Rother, H. Vongerichten, R.Lucius and H.J. Diesfeld (1985) *Schistosoma mansoni,*: immunoblot analysis of adult worm proteins. Exp. Parasitol. 60 : 195-206.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol.Biol.Evol. 4,

Sakanaki, J.A. S. A. Nadler, V.J. Chan, J. C. Engel, C. Leptak and J. Bouvier (1997*)* *Leishmania major*: comparison of the cathepsin L- and B-like cysteine protease genes with those of the other trypanosomatids. Exp. Parasitol 85 (1): 63-76.

Salminen, A. and M.M. Gottesman (1990) Inhibitor studies indicate that active cathepsin L is probably essential to its own processing in cultured fibroblasts. Biochem.J. 272 (1): 39-44.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) Molecular Cloning : A laboratory manual 2nd. edition (Maniatis, T., Fritsch, E. F. & Sambrook, J., Eds.). Cold Spring Harbor Laboratory Press, NY.

Sander, C. & Schneider, R. (1991) Database of homology-derived protein structures. Proteins 9: 56-68.

Shaw, E., S. Mohanty, A. Colic, V. Stoka and V. Turk (1993) The affinity-labelling of cathepsin S with peptidyl diazomethyl ketones.Comparison with the inhibition of cathepsin L and calpain. FEBS Lett 334 (3): 340-342

Shi, G.P., A.C. Webb, K.E. Foster, J.H.M. Knoll, C.A. Lemere, J.S. Munger and H. A. Chapman (1997) Human cathepsin S: chromosomal localization, gene structure and tissue distribution. J. Biol. Chem. 269 (15): 11530-11536.

Siddiqui, A.A., Y.Zhou, R.B.Podesta, S.R.Karcz, C.E.Tognon, G.H. Dekaban and M.W. Clarke (1993) Characterization of Ca^{++} dependent neutral protease (calpain) from human blood flukes *Schistosoma mansoni*. Biochi.Biophys. Acta 1181:37-44.

Simpkin, K.G., C.R. Chapman and G.C.Coles (1980) *Fasciola hepatica*: a proteolytic digestive enzyme. Exp. Parasitol. 49 (2): 281-287.

Smith, A.M. (1994) *Fasciola hepatica*: Isolation and characterization of a cathepsin L proteinase. PhD Thesis, Dublin City University.

Smith, A.M., A.J.Dowd ,S. McGonigle, P.S. Keegan, G.Brennan, A.Trudgett and J.P.Dalton (1993) Purification of a cathepsin L-like proteinase secreted by adult *Fasciola hepatica*. Mol. Biochem. Parasitol. 62 (1): 1-8.

Smith, A.M., A.J.Dowd, M.Heffernan, C.D. Robertson and J.P.Dalton (1993) *Fasciola hepatica*: a secreted cathepsin L-like proteinase cleaves host immunoglobulin. Int. J.Parasitol. 23 (8): 977-983.

Smith, A.M., J.P.Dalton, K.A.Clough, C.L.Kibane, S.A.Harrop, N.Hole and P.J. Brindley (1994) Adult *Schistosoma mansoni* express cathepsin L proteinase activity.Mol.Biochem. Parasitol. 67:11-19.

Sogin, M. L. (1991) Early evolution and the origin of eukaryotes. Curr. Opinion Genet. Develop. 1: 457-463.

Song, C.Y. and A.A. Rege (1991) Cysteine proteinase activity in various developmental stages of *Clonorchis sinensis*: a comparative analysis. Comp. Biochem. Physiol. 99(1): 137-140.

Song, C.Y. and C.L. Chappell (1993) Purification and partial characterization of cysteine proteinase from *Spirometra mansoni* plerocercoids. J.Parasitol. 79(4):517-524.

Song, C.Y. and M.H.Dresden (1990) Partial purification and characterization of cysteine proteinases from various developmental stages of *Paragonimus westermani*. Comp.Biochem.Physiol. 95 (3):473-476.

Song, C.Y. and T.S. Kim (1994) Characterization of a cysteine proteinase from adult worms of *Paragonimus westermani*. Korean J. Parasitol. 32(4):231-241.

Song, C.Y., Dresden, M.H., and A.A. Rege (1990) *Clonorchis sinensis*: purification and characterization of a cysteine proteinase from adult worms. Comp. Biochem. Physiol. 97(4): 825-829.

Song,C.Y., D.H.Choi, T.S.Kim and S.H.Lee (1992) Isolation and partial characterization of cysteine proteinase from sparganum. Kisaengchunghak Chapchi 30(3):191-199.

Souza, A.E., S. Waugh, G. H. Coombs and J.C. Mottram (1992) Characterization of a multy-copy gene for a major stage-specific cysteine proteinase of *Leishmania mexicana*. FEBS Lett. 311 (2): 124-127.

Souza, G.M, J. Hirai, D.P. Mehta and H.H. Freeze (1995) Identification of two novel *Dictyostelium discoideum* cysteine proteinases that carry N-acetylglucosamine-1-P-modification. J.Biol.Chem. 270 (48):28938-28945.

Storer, A. C. and R. Menard (1996) Recent insights onto cysteine protease specificity: lessons for drug design. Persp. Drug Discov. Design 6: 33-46.

Sung, C.K. and M.L. Dresden (1986) Cysteinyl proteinases of *Schistosoma mansoni* eggs:purification and partial characterization. J.Parasitol. 72 (6):891-900.

Swofford, D.L. and G.J. Olsen (1990) Phylogeny reconstruction. In Molecular Systematics (Hillis,D.M and C. Moritz, eds). Sinauer Associates, Massachusetts, USA.

Tagami, K. H. Kakegawa, H. Kamikoka, K. Sumitani, T. Kawata, B. Lenarcic, V. Turk and N. Katunuma (1994) The mechanisms and regulation of procathepsin L secretion from osteoclasts in bone resorption. FEBS Lett. 342 (3): 308-312.

Takahashi, S.Y.,Y. Yamamoto, Y. Shionoya and T. Kageyama (1993) Cysteine proteinase from the eggs of the silkmoth *Bombyx mori*: identification of a latent enzyme and characterization of activation and proteolytic processing *in vivo* and *in vitro*. J. Biochem. 114 (2): 267-272.

Tanaka, T., Y. Kaneda, A. Iida and M. Tanaka (1994) Homologous cysteine proteinases genes located on two different chromosomes from *Trypanosoma rangeli*. Int. J. Parasitol. 24 (2): 179-188.

Tao, K., N.A. Stearns, J. Dong, Q. Wu and G. Sahagian (1994) The proregion of cathepsin L is required for proper folding, stability and ER exit. Arch. Biochem. Biophys. 311(1): 19-27.

Taralp, A., H. Kaplan, I. Sytwu, I. Vlattas, R. Bohacek, A.K. Knap, T. Hirama, C.P. Huber and S. Hasnain (1995) Characterization of the S_3 subsite specificity of cathepsin B. J. Biol. Chem. 276:18036-18043.

Taylor, M.A.K., G.S. Briggs, K.C. Baker, N.J. Cummings, K.A. Pratt, R.B. Freedman and P.W. Goodenough (1995) Expression of the pro-regions of papapin and papaya proteinase IV in *Escherichia coli* and their inhibition of mature cysteine proteinases. Biochem. Soc. Trans. 23(1):80S.

Taylor, M.A.K., K.C. Baker, G.S. Briggs, I.F. Connerton, N.J. Cummings, K.A. Pratt, D.F. Revell, R.B. Freedman and P.W. Goodenough (1995) Recombinant pro-regions of papapin and papaya proteinase IV are selective high affinity inhibitors of the mature papaya enzymes. Protein Eng. 8 (1):59-62.

Tezuka, K.I., Y. Tezuka, A. Maejima, T. Sato, K. Nomoto, H. Kamioka, Y. Hakeda and M. Kumenawa (1994) Molecular cloning of a possible cysteine proteinase predominantly expressed in osteoclasts. J. Biol. Chem 269 (2): 1106-1109.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22:4673-4680.

Tkalcevic, J., K.Ashman and E.Meeusen (1995) *Fasciola hepatica*: rapid identification of newly excysted juvenile proteins. Biochem. Biophys. Res. Commun. 213(1): 169-174.

Traub-Cseko, Y.M., M. Duboise, L.K. Boukai and D. McMahon-Pratt (1993) Identification of the two distinct cysteine proteinase genes *of Leishmania pifanoi* axenic amastigotes using the polymerase chain reaction. Mol. biochem. Parasitol. 57 (1): 101-115.

Turk, D. M. Podobnick, T. Popovic, N. Katunuma, W. Bode, R. Huber and V. Turk (1995) Crystal structure of cathepsin B inhibited with CA030 at 2.0-A resolution: A basis for the design of specific epoxysuccinyl inhibitors. Biochemistry 34 (14): 4791-4797.

Turk, D. M. Podobnik, R. Kuhelj, M. Dolinar and V. Turk (1996) Crystal structures of human procathepsin B at 3.2 and 3.3 A resolution reveal an interaction motif between a papain-like cysteine protease and its propeptide. FEBS Lett. 384:211-214.

Van der Stappen, J.W., A.C. Williams, R.A. Maciewicz and C. Paraskeva (1996) Activation of cathepsin B secreted by a colorectal cancer cell line requires low pH and is mediated by cathepsin D. Int. J. Cancer 67 (4):547-554.

Velazco, G., A.A. Ferrando, X. S. Puente, L.M. Sanchez and C. Lopez-Otin (1994) Human cathepsin O: Molecular cloning from a breast carcinoma, production of the active enzyme in *Escherichia coli* and expression analysis in human tissues. J. Biol. Chem 269 (43) 27136-27142.

Vernet , T., J. Chatellier, D. C. Tessier and D. Y. Thomas (1993) Expression of functional papain precursor in *Saccharomyces cerevisiae*: rapid screening of mutants Protein Eng 6 (2): 213-219.

Vernet, T., H.E. Khouri, P. Laflamme, D.C. Tessier, R. Musil, B.J. Gour-Salin, A.C. Storer and D.Y. Thomas (1991) Processing of the papain precursor. Purification of the zymogen and characterization of its mechanism of processing. J.Biol.Chem. 266 (32): 21451-21457.

Vernet, T., P.J. Berti, C. de Montigny, R. Musil, D.C. Tessier, R. Menard, M.C. Magny, A.C. Storer and D.Y. Thomas (1995) Processing of the papain precursor. The ionization state of a conserved amino acid motif within the pro region participates in the regulation of the intramolecular processing. J. Biol. Chem. 270 (18): 10838-10846.

Volkel, H., U. Kurz, J. Linder, S. Klumpp, V. Gnau, G. Jung and J.S. Schultz (1996) Cathepsin L is an intracellular and extracellular protease in *Paramecium tetraurelia*. Purification, cloning, sequencing and specific inhibition by its expressed propeptide. Eur. J. Biochem. 238: 198-206.

Wang, X., S.J. Chan, R.L. Eddy,M.G. Byers, Y. Fukushima, W.M. Henry, L.L.Haley, D.F. Steiner and T.B.Shows (1987) Chromosome assignment of cathepsin B (CTSB) to 8p22 and cathepsin H (CTSH) to 15q24-q25. Cytogenet. Cell Genet. 46: 710-711.

Ward, W., L. Alvarado, N.D. Rawlins, J.C. Engel, C. Franklin and J.H. McKerrow (1997) A primitive enzyme for a primitive cell: the protease required for excystation of *Giardia*. Cell 89 :437-444.

Warner, A.H., M.J. Perz, J.K. Osahan and B.S. Zielinski (1995) Potential role in development of the major cysteine protease in larvae of the brine shrimp *Artemia franciscana*. Cell Tissue Res. 282 (1): 21-31.

Wasilewski, M.M., K.C. Lim, J. Philips and J.H.McKerrow (1996) Cysteine protease inhibitors block schistosome hemoglobin degradation in vitro and decrease worm burden and egg production in vivo. Mol.Biochem. Parasitol. 81: 179-189.

White A.C. Jr, S.Baig and C.L. Chappell (1997) Characterization of a cysteine proteinase from *Taenia crassiceps* cysts. Mol Biochem Parasitol 85 (2): 243-253.

White, A.C. Jr, J.L. Molinari, A.V. Pillai and A.A. Rege (1992) Detection and preliminary characterization of *Taenia solium* metacestode proteases. J Parasitol 78 (2): 281-287.

White, A.C. Jr., S. Baig and P. Robinson (1996) *Taenia saginata* oncosphere excretory/secretory peptidases. J Parasitol 82 (1): 7-10.

Wijffels, G.L., L.Salvatore, M.Dosen, J.Waddington, L.Wilson, C.Thompson, N.Campbell, J.Sexton, J.Wicker, F.Bowen, T.Friedel and T.W. Spithill (1994) Vaccination of sheep with purified cysteine proteinases of *Fasciola hepatica* decreases worm fecundity. Exp. Parasitol. 78(2): 132-148.

Wijffels,G.L., M.Panaccio, L.Salcatore, L.Wilson, I.D.Walker and T.W.Spithill (1994) The secreted cathepsin L-like proteinases of the trematode *Fasciola hepatica* contain 3-hydroxyporline residues. Biochem. J. 299 :781-790.

Williams, J.G., M.J.North and H. Mahbubani (1985) A developmentally regulated cysteine proteinase in *Dictyostelium discoideum*. EMBO J.4 (4):999-1006.

Wilson, L.R., R. T. Good, M.Panaccio, G.L.Wijffels, J.Creaney, S.E.Bozas, J.C.Parsons, R.M.Sanderman and T.W.Spithill (1997) Characterization and cloning of the major cathepsin B protease secreted by newly excysted juvenile *Fasciola hepatica*. Submitted.

Yamakami, K and F.Hamajima (1987) Purification and properties of a neutral thiol protease from larval trematode parasite *Paragonimus westermani* metacercariae. Comp.Biochem.Physiol. 87(3):643-648.

Yamakami, K and F.Hamajima (1990) A neutral thiol protease secreted from newly excysted metacercariae of trematode parasite *Paragonimus westermani:* purification and characterization. Comp.Biochem.Physiol. 95(4):755-758.

Yamakami, K. (1986) Purification and properties of a thiol protease from lung fluke adult *Paragonimus ohirai*. Comp.Biochem.Physiol. 83(3):501-506.

Yamakami, K., F.Hamagima, S.Akao and T.Tadakuma (1995) Purification and characterization of acid protease from metacercariae of the mammalian trematode *parasite Paragonimus westermani*. Eur.J.Biochem. 233(2):490-497.

Yamamoto, M. K.Yamakami and F.Hamajima (1994) Cloning of a cDNA encoding a neutral thiol protease from *Paragonimus westermani* metacercariae. Mol.Biochem.Parasitol. 64:345-348.

Yamamoto, Y., K. Takimoto, S. Izumi, M. Toriyama-Sakurai, T. Kageyama and S.Y. Takahashi (1994) Molecular cloning and sequencing of cDNA that encodes cysteine proteinase in the eggs of the silkmoth *Bombyx mori*. J. Biochem. 116: 1330-1335.

Yamasaki, H. and T. Aoki (1993) Cloning and sequence analysis of the major cysteine protease expressed in the trematode parasite *Fasciola sp*. Biochem. Mol. Biol. Int. 31(3): 537-542.

Yamasaki,H., E.Kominami and T.Aoki (1992) Immunocytochemical localization of a cysteine protease in adult worms of the liver fluke *Fasciola sp*. Parasitol.Res. 78:574-580.

Yoshino, T.P., M.J.Lodes, A.A.Rege and C.L.Chappell (1993) Proteinase activity in miracidia, transformation excretory-secretory products, and primary sporocysts of *Schistosoma mansoni*. J.Parasitol. 79(1):23-31.

9. Appendix

List of sequences included in the phylogenetic analysis

Name	*sequence ID*	*Data base*	*Accession*
ACTCH_ACTA	226542	prf	226542
ACTCH_ACTD	99595	pir	S12618
ACTCH_ACTN	113285	sp	P00785
AEDEG_CATB	1008858	gb	L41940
ALNGL_CPRO	535454	gb	U13940
ANACO+BROM	115139	sp	P14518
ANCCA*CPRO	496968	gb	U02611
ANCCA__CB1	984960	gb	U18912
ANCCA__CB2	984958	gb	U18911
ARATH_A494	1168251	sp	P43295
ARATH_RD19	1172872	sp	P43296
ARATH_RD21	1172873	sp	P43297
ASCSU_CATB	1777779	embl	U51892
BOMMO_CPRO	1085124	pir	JX0366
BOVIN_CATB	1168789	sp	P07688
BOVIN+CATS	115747	sp	P25326
BRANA_CYS4	118127	sp	P25251
CAEEL*CP1	508264	gb	U11245
CAEEL*CP2	437323	gb	L22447
CAEEL_CPR3	1169083	sp	P43507
CAEEL_CPR4	1169085	sp	P43508
CAEEL_CPR5	1169086	sp	P43509
CAEEL_CPR6	1169087	sp	P43510
CAEEL_CYS1	118116	sp	P25807
CARCN+CC3	926847	gb|bbs	161221
CARP_CPRO	463046	gb	L30111
CARPA_PAP2	1332461	embl	X97789
CARPA_PAP3	1709574	sp	P10056
CARPA_PAP4	1709576	sp	P05994
CARPA_PAPA	129614	sp	P00784
CAT*CATL	115740	sp	P25773
CHICK_CATB	1168790	sp	P43233
CHICK_JTAP	1017831	gb	U37691
CHICK+CATL	359286	prf	359286
CPEA_CPRO	1071886	pir	S49451
DERFA_MMAL	730035	sp	P16311
DERPT_MMAL	730036	sp	P08176
DICAR_CPRO	1361974	pir	S57776
DICDI*PR2G	829172	embl	X03930
DICDI_CP4	1222695	gb	L36204
DICDI_CP5	1222694	gb	L36205
DICDI_CYS1	118117	sp	P04988
DICDI_CYS2	118121	sp	P04989
DROME_CPRO	1093503	prf	1093503
ENTHI*CPRO	881588	gb	U28755
ENTHI_ACP1	1460065	embl	X87214
ENTHI_CP	1246523	embl	X91642
ENTHI_CP5	1246525	embl	X91644
ENTHI_CP6	1246527	embl	X91645
ENTHI_CPP1	544088	sp	Q01957
ENTHI_CPP2	544089	sp	Q01958
ENTHI+CPP3	544090	sp	Q06964
EURMA_EUM1	119656	sp	P25780
FASHE*FCP2	452254	embl	Z22763
FASHE*FCP4	452262	embl	Z22767
FASHE_CL1	1809286	gb	U62288
FASHE_CL2	1809288	gb	U62289
FASHE_CLES	535600	gb	L33772
GIAIN_CPI1	1763659	gb	U83275
GIAIN_CPI2	1763661	gb	U83276
GIAIN_CPI3	1763663	gb	U83277
HAECO*PDM2	1181139	embl	Z69343
HAECO*PDM3	1181141	embl	Z69344
HAECO*PDM4	1181143	embl	Z69345
HAECO*PDM5	1181145	embl	Z69346
HAECO__AC3	478099	pir	D48435
HAECO__AC4	478007	pir	C48435
HAECO__AC5	477808	pir	B48435
HAECO_CYS1	118118	sp	P19092
HAECO_CYS2	118122	sp	P25793
HEMSP_CYSP	1169186	sp	P43156
HEMSP_SEN11	537437	gb	U12637
HOMAM_CYS1	118119	sp	P13277
HOMAM_CYS2	118123	sp	P25782
HOMAM_CYS3	118125	sp	P25784
HORVU_ALEU	113603	sp	P05167
HORVU_CYS1	118120	sp	P25249
HORVU_CYS2	118124	sp	P25250
HUM_CATO	1168795	sp	P43234
HUMAN_CATB	115711	sp	P07858
HUMAN_CATC	1006657	embl	X87212
HUMAN_CATH	115728	sp	P09668
HUMAN_CATK	1168793	sp	P43235
HUMAN_CATL	115741	sp	P07711
HUMAN_CATS	115748	sp	P25774
LEIME_CATB	728602	embl	Z48599
LEIME_LCPA	126021	sp	P25775
LEIME_LCPB	547835	sp	P36400
LEIPI_CYS1	1220383	gb	L29168
LEIPI_CYS2	461905	sp	Q05094
LYCES_CYP2	19195	embl	Z14028
LYCES_CYP3	1235545	embl	Z48736
LYCES+CLOW	118145	sp	P20721
MAIZE_CYS1	1706260	sp	Q10716
MAIZE_CYS2	1706261	sp	Q10717
MECRY_CPRO	944916	gb	U30322
MOUSE_CATB	115712	sp	P10605
MOUSE_CATH	454101	gb	U06119
MOUSE_CATK	1149525	embl	X94444
MOUSE_CATL	115742	sp	P06797
NAEFO+CPRO	1353726	gb	U42758
NEPNO_CLE	630815	pir	S47432
NEPNO_CLS	630816	pir	S47433
NITOB_CATB	1076610	pir	S52212
NITOB_CYP7	419781	pir	S30149
NITOB_CYP8	419782	pir	S30150
NPVAC_CATV	115751	sp	P25783
NPVBM_CATV	1168798	sp	P41721
NPVCF_CATV	1168799	sp	P41715
ONCVU_CATC	1680720	gb	U71150
ORCH_CPRO	1173630	gb	U34747
ORYSA_CPR1	1514953	ddbj	D76415
ORYSA_CPR2	629792	pir	S47434
ORYSA_ORYA	129231	sp	P25776
ORYSA_ORYB	129232	sp	P25777
ORYSA_ORYC	129233	sp	P25778
OSTOS*CYS3	729283	sp	Q06544
OSTOS_CYS1	1345924	sp	P25802
PARTE_CTL1	1403087	embl	X91754
PARTE_CTL2	1403089	embl	X91756
PARWM+NTP	633096	ddbj	D21124
PEA_CP15A	118150	sp	P25804
PEA_CPTPP	100069	pir	S24602
PEA_NACP	1134882	embl	Z68291
PEA_NTH1	1174171	gb	U44947
PENVA_PCP1	1085687	pir	S53027
PENVA_PCP2	1483570	embl	X99730
PHAVU_CEP1	1256830	gb	U52970
PHAVU_CYSP	544129	sp	P25803
PIG_CATL	1468964	ddbj	D37917
PLABR+CYSP	950238	gb	U33420
PLACY+CYSP	950240	gb	U33422
PLAFA_CYSP	118152	sp	P25805
PLAFR+CYSP	950242	gb	U33421
PLAOV+CYSP	950248	gb	U33423
PLAVI_CYSP	1169187	sp	P42666
PLAVN_CYSP	1169188	sp	P46102
PSMEN_PSTZA	1208549	gb	U41902
RABIT_CATK	1168794	sp	P43236
RAT_CATB	1524328	embl	X82396
RAT_CATC	115716	sp	P80067
RAT_CATH	115729	sp	P00786
RAT_CATL	115743	sp	P07154
RAT_CATR12	894096	gb	L14776
RAT_CATS	399190	sp	Q02765
RAT_TEST	1174639	sp	P15242
SARPE_CATB	481614	pir	s38939
SARPE_CATL	1079183	pir	A53810
SCHJP_CB1	1169189	sp	P43157
SCHJP_CB2	345308	pir	S31909
SCHJP+CL1	1185457	gb	U38475
SCHJP+CL2	1185457	gb	U38476
SCHMA_CATB	118153	sp	P25792
SCHMA_CATC	1262412	embl	Z32531
SCHMA_CL1	1094710	prf	1094710
SCHMA+CL2	630486	pir	S44151
SHEEP+CATL	1705639	sp	Q10991
SOYBN_CEND	479060	embl	Z32795
SPIER_CPRO	1834309	ddbj	D63671
SPIMA+CPRO	1272388	gb	U51913
STRRA*CPRO	603044	gb	U09818
TETER_SGC5	476938	pir	A47306
THEAN_CYSP	118154	sp	P25781
THEPA_CYSP	118155	sp	P22497
TOXCA_CATL	1279986	gb	U53172
TRIVG*CP3	542417	pir	S41425
TRIVG_CP1	542419	pir	S41427
TRIVG_CP2	542420	pir	S41428
TROUT*CATK	1698589	embl	U61499
TROUT*CATL	1698586	embl	U61296
TRTAE_CATB	21693	embl	X66012
TRTAE_THBG	21699	embl	X66013
TRYBB*CYPA	162042	gb	M27306
TRYBB_CYSP	118156	sp	P14658
TRYCG_CPRO	408009	gb	L25130
TRYCG_CYSP	480567	pir	S37048
TRYCR*CYPA	419926	pir	A44938
TRYCR_CPNS	577617	gb	M90067
TRYCR_CYPR	1163075	embl	Z25813
TRYCR_CZP2	1136308	gb	U41454
TRYCR_CZPN	118157	sp	P25779
TRYCR_CZPP	323055	pir	A45629
TRYRG*CPRO	538255	gb	M99496
TTMFT*CP3	1364025	pir	S57451
TTMFT*CP4	1364026	pir	S57427
TTMFT*CP5	1364027	pir	S57426
TTMFT*CP6	1364028	pir	S57421
TTMFT*CP7	1364029	pir	S57425
TTMFT*CP8	1364030	pir	S57422
TTMFT*CP9	1364031	pir	S57423
TTMFT_CP1	1141743	gb	U13153
TTMFT+CP2	1141745	gb	U13154
URECA_CATB	945054	gb	U30877
VICFA_CPRO	1401242	gb	U59465
VICSA_CPR1	457756	embl	Z30338
VICSA_CPR2	600111	embl	Z34895
VICSA_CPR3	535473	embl	X75749
VIGMU_CYSP	118158	sp	P12412
ZFISH_CATL	1752664	embl	Y08321
ZINEL_CPRO	641905	gb	U19267

An (*) in the sequence name indicate partial sequences

A (+) in the sequence name, indicate partial sequences, but comprising the complete mature region