*Genetics and Molecular Biology, 24 (1-4), 275-283 (2001)*

# Identification, classification and expression pattern analysis of sugarcane cysteine proteinases

*Gustavo Coelho Correa[1,2], Márcia Margis-Pinheiro[2] and Rogério Margis[1,2*]*

*Abstract*

Cysteine proteases are peptidyl hydrolyses dependent on a cysteine residue at the active center. The physical and chemical properties of cysteine proteases have been extensively characterized, but their precise biological functions have not yet been completely understood, although it is known that they are involved in a number of events such as protein turnover, cancer, germination, programmed cell death and senescence. Protein sequences from different cysteine proteinases, classified as members of the E.C.3.4.22 sub-sub-class, were used to perform a T-BLAST-n search on the Brazilian Sugarcane Expressed Sequence Tags project (SUCEST) data bank. Sequence homology was found with 76 cluster sequences that corresponded to possible cysteine proteinases. The alignments of these SUCEST clusters with the sequence of cysteine proteinases of known origins provided important information about the classification and possible function of these sugarcane enzymes. Inferences about the expression pattern of each gene were made by direct correlation with the SUCEST cDNA libraries from which each cluster was derived. Since no previous reports of sugarcane cysteine proteinases genes exists, this study represents a first step in the study of new biochemical, physiological and biotechnological aspects of sugarcane cysteine proteases.

## INTRODUCTION

Proteinases, or endopeptidases, are enzymes that catalyze the hydrolysis of peptide bonds within proteins. Based on their catalytic mechanisms these enzymes are classified as serine, cysteine, aspartic or metallo-proteinases (Barrett, 1980). Cysteine or thiol proteinases (EC 3.4.22) are those that contain a cysteine residue in the active site. These proteases have been identified in phylogenetically diverse organisms, such as bacteria, eukaryotic micro-organisms, plants and animals (Rawlings and Barret, 1994).

More than 30 families of peptidases, grouped in at least six clans (or superfamilies), make up the class of cysteine proteases. Members of the six major clans are defined according to the nature and linear organization of the catalytic residues along the primary sequence as follow: Clan CA has the catalytic residues Cys, His and Asn or Asp ordered in sequence; Clan CD presents two catalytic residues, His and Cys, in sequence; Clan CE has a triad formed by His, Glu or Asp and Cys at the C-terminus; Clan CF also presents a catalytic triad, but ordered as Glu, Cys and His; Clan CG has a dyad of two cysteine residues and Clan CH presents a Cys, Thr and His triad with the catalytic cysteine at the N-terminus (Rawlings and Barret, 2000).

A common feature to all cysteine proteinases with known three-dimensional structure is the existence of a bi-lobed structure, with the catalytic site located in the cleft between the lobes (Rawlings and Barret, 2000). The papain superfamily, or clan CA, corresponds to the best-known cysteine peptidases and has the catalytic residues Cys-25 and His-159 conserved in all of its members. They are synthesized as preproenzymes and are located in lysosomes or analogous organelles. The most studied cysteine proteinase is papain, from *Carica papaya*, that represents the typical member of this superfamily.

Sequence analysis has revealed that other higher plant cysteine proteinases and cathepsins B, H, L and S from mammalian lysosomes are members of the papain C1A family. In addition, bleomycin (family C1B), calpains (family C2), streptopain (family C10) and viral proteases also belong to this superfamily (Rawlings and Barret, 2000). The calpains are cytoplasmic, calcium dependent cysteine proteases, which differ in requiring micro or milimolar concentrations of $Ca^{2+}$ for activity and have a very high conserved molecular structure (Croall and Demartino, 1991).

In mammals, cysteine proteases such as lysosomal cathepsins comprise a group of small proteases having M*r* values of less than 30.000 and are active at acidic pH. They

[1]*Departamento de Bioquímica, Instituto de Química, CCMN, Universidade Federal do Rio de Janeiro, Ilha do Fundão, 21944-970 Rio de Janeiro, RJ, Brazil.*
[2]*Laboratório de Genética Molecular Vegetal, Departamento de Genética, Instituto de Biologia, CCS, Universidade Federal do Rio de Janeiro, Ilha do Fundão, Rio de Janeiro, RJ, Brazil.*
*Send correspondence to Dr. R. Margis. Laboratório de Genética Molecular Vegetal, Departamento de Genética, Instituto de Biologia, CCS, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil. E-mail: margisr@ufrj.br.*

are synthesized as precursor molecules, which contain N-terminal signal peptides that are cleaved off during transport through the membrane of the endoplasmic reticulum. Cathepsin B is one of the well-characterized lysosomal cysteine proteinases. Human cathepsin B was the first lysosomal cysteine proteinases whose crystal structure was elucidated (Mort, 1998)

Cotyledonous legumes present an increasing peptidase activity during germination, corresponding to an atypical cysteine endopeptidase (legumain) with cleavage specificity for asparagine or aspartate residues in the P1 position of the peptide target (Ishii, 1994). Legumain has been shown to have sequence and functional similarity to plant vacuolar processing enzymes (VPE) and also to hemoglobinase from *Schistosoma mansoni* (Chappell and Dresden, 1986). Plant VPEs (Hiraiwa *et al*., 1993) have been proposed to play a role in the degradation of seed storage proteins or, alternatively, in their limited proteolysis, which also occurs during maturation of these proteins (Kembhavi *et al.*, 1993). Thus, it seems that similar enzymes are involved in two opposite processes, *i.e.* storage protein deposition and mobilization. In addition to its location in the vacuoles of developing seeds, this processing enzyme can be detected in other vegetative organs, such as hypocotyls, roots and mature leaves suggesting that VPE could be key enzymes in vacuolar metabolism (Hara-Nishimura *et al*., 1998).

In spite of their very well characterized physicochemical properties, the role of the cysteine proteases *in vivo* is not yet completely understood. The complex spatial and temporal regulation of the expression of some cysteine proteases suggests that they can play diverse functions in cell metabolism. Moreover, in many cases the expression of multiple cysteine proteinases within a single organism is independently regulated (Watanabe *et al.*, 1991; Koehler and Ho, 1990; Linthorst *et al*., 1993).

As part of the ongoing program to characterize sugarcane Expressed Sequence Tags (ESTs) we have identified sugarcane cysteine proteases, sub-sub-class 3.4.22, and correlated their localization and putative functions. Understanding the evolutionary relationship between cysteine protease could help identify the function of individual proteases.

## MATERIALS AND METHODS

### Sequence data, alignment and phylogenetic analysis

The cysteine proteinase (E.C.3.4.22 sub-sub-class) amino acid and deduced amino acid sequences, were accessed from the SwissProt (SP) data bank (Bairoch, 2000). The cysteine proteinase official name, E.C. number, organism name and SP data bank accession number of the cysteine proteinases used are shown in Table I. A T-Blast-n search (Altschul *et al*., 1997) was performed using these bait sequence against the full SUCEST cDNA data bank.

**Table I** - Cysteine proteinase official name, E.C. number, scientific name of the organism and accession number of the cysteine proteinases used in this work.

| Cysteine proteinase | E.C. number | Organism | Accession number |
|---|---|---|---|
| cathepsin B | E.C.3.4.22.1 | *Homo sapiens* | SP: P07858 |
| papain | E.C.3.4.22.2 | *Carica papaya* | SP P00784 |
| chymopapain | E.C.3.4.22.6 | *C. papaya* | SP: P14080 |
| clostripain | E.C.3.4.22.8 | *Clostridium histolyticum* | SP: P09870 |
| streptopain | E.C.3.4.22.10 | *Streptococcus pyogenes* | SP: P00788 |
| actinidain | E.C.3.4.22.14 | *Actinidia chinensis* | SP: P00785 |
| cathepsin L | E.C.3.4.22.15 | *H. sapiens* | SP: P07711 |
| cathepsin H | E.C.3.4.22.16 | *H. sapiens* | SP: P09668 |
| calpain | E.C.3.4.22.17 | *H. sapiens* | SP: P17655 |
| cathepsin S | E.C.3.4.22.27 | *H. sapiens* | SP: P25774 |
| caricain | E.C.3.4.22.30 | *C. papaya* | SP: P10056 |
| ananain | E.C.3.4.22.31 | *Ananas comosus* | SP: P80884 |
| bromelain | E.C.3.4.22.32 | *A. comosus* | SP: P14518 |
| legumain | E.C.3.4.22.34 | *Canavalia ensiformis* | SP: P49046 |
| caspase-1 | E.C.3.4.22.36 | *H. sapiens* | SP: P29466 |

The multiple alignment program (MAP) computes a multiple global alignment of sequences using a pairwise method. Its algorithm for aligning two sequences computes the best overlapping alignment between two sequences without penalizing terminal gaps. In addition, long internal gaps in short sequences are not heavily penalized. The MAP produces a consistent alignment notwithstanding some sequences present in long terminal or internal gaps, and the MAP is designed in a space-efficient manner, allowing long sequences to be aligned (Huang, 1994). This method was used to align the different cysteine proteinases standards from the E.C.3.4.22 sub-sub-class and also to align the proteins deduced from the SUCEST clusters with other cysteine proteinases from plants, vertebrates and invertebrates, the acronyms and accession numbers of these sequences being shown in Table II.

Phylogenic analyses were performed using the Molecular Evolutionary Genetics Analysis (MEGA) software, version 2.0 (Kumar *et al*., 2000). The pair-wise deletion option was adopted on the treatment of amino acid gaps on the sugarcane cysteine protease multiple alignment. Trees were obtained from Neighbor-joining analysis derived from the p-distance method. In the phylogenetic tree construction, the confidence levels assigned at various nodes were determined after 5000 replications using the Interior Branch test (Sitnikova *et al*., 1995).

**Table II** - Cysteine proteinases from plants, vertebrates and invertebrates.

| Acronym | Accession number | Organism |
|---|---|---|
| Cathepsins | | |
| TaesB | GB: CAA46811 | *Triticum aestivum* |
| NrusB | GB: CAA57522 | *Nicotiana rustica* |
| Atha1B | GB: AAC24376 | *Arabidopsis thaliana* |
| Atha2B | GB: CAB77732 | *A. thaliana* |
| PsatB | GB: CAB62589 | *Pisum sativum* |
| IbatB | GB: AAF04727 | *Ipomoea batatas* |
| PhybH | GB: AAC49361 | *Petunia x hybrida* |
| OsatH | SP: P25778 | *Oryza sativa* |
| HvulH | GB: KHBH | *Hordeum vulgare* |
| HarmB | GB**:** AAF35867 | *Helicoverpa armigera* |
| GgalB | GB: S58770 | *Gallus gallus* |
| BtauB | GB: KHBOB | *Bos tauros* |
| HsapB | GB: XP_005133 | *Homo sapiens* |
| RnorB | GB**:** NP_072119 | *Rattus norvegicus* |
| MmusB | GB: KHMSB | *Mus musculus* |
| ScroH | GB: O46427 | *Sus scrofa* |
| RnorH | GB: NP_037071 | *R. norvegicus* |
| HsapL | SP: XP005441 | *H. sapiens* |
| CaelL | GB: AAG35605 | *Cercopithecus aethiops* |
| CfamL | GB: CAC08809 | *Canis familiaris* |
| DmelL | GB: S67481 | *Drosophila melanogaster* |
| HglyL | GB: CAA70693 | *Heterodera glycines* |
| Legumains | | |
| Osat | GB: BAA84650 | *Oryza sativa* |
| Zmai | GB: CAB64544 | *Zea mays* |
| Pvul1 | GB: T12043 | *Phaseolus vulgaris* |
| Pvul2 | GB: T12044 | *P. vulgaris* |
| Vmu | GB: BAA76745 | *Vigna mungo* |
| Vsat | SP: P49044 | *Vicia sativa* |
| Atha1 | GB: BAA18924 | *Arabidopsis thaliana* |
| Atha2 | GB: T05302 | *A. thaliana* |
| Vna | GB: CAB42655 | *Vicia narbonensis* |
| Csin | SP: P49043 | *Citrus sinensis* |
| Gmax | SP: P49045 | *Glycine max* |
| Sin | GB: AAF89679 | *Sesamum indicum* |
| Rcom | SP: P49042 | *Ricinus communis* |
| Ntab | GB: CAB42651 | *Nicotiana tabacum* |
| Lesc | GB: CAB51545 | *Lycopersicon esculentum* |
| Hsap | GB: NP_005597 | *Homo sapiens* |
| Mmus | GB: NP_035305 | *Mus musculus* |
| Rnor1 | GB: BAA84750 | *Rattus norvegicus* |
| Rnor2 | NP_071562 | *R. norvegicus* |
| Cele | GB: AAF21773 | *Caenorhabditis elegans* |
| Dmel | GB: T13411 | *Drosophila melanogaster* |
| Sjap | SP: P42665 | *Schistosoma japonicum* |
| Sman | GB: CAB71158 | *Schistosoma mansoni* |

### Description of SUCEST cDNA libraries

All sugarcane sequences used in this work were obtained from the Brazilian SUCEST project (http://sucest.lad.dcc.unicamp.br/en/) and derived from cDNA libraries specific to different sugarcane tissues, organs or growth conditions. The libraries were are as follows: apical meristem from mature (AM1) and (AM2) immature plants; 1 cm (FL1) and 5 cm (FL3) flower base; 50 cm (FL4), 20 cm (FL5) and10 cm (FL8) flower stem; lateral buds (LB1 and LB2); large (LR1) and small (LR2) leaf-root insert libraries; etiolated leaves (LV1); grouped data of two non-redundant libraries (NRn); grouped data of three root libraries (RTn); grouped data of three leaf-root transition zone libraries (RZn); stem bark (SB1); grouped data of two seed libraries of different insert sizes (SDn); grouped data of two stem libraries from the first and fourth internodes (STn); libraries derived from *calli* submitted to a 4-37 °C temperature change and three (CL3), four (CL4), six (CL6) and eight (CL7) hours of a light/dark cycle; plants infected with the bacteria *Gluconacetobacter diazotroficans* (AD1) and *Herbaspirillum rubrisubalbicans* (HR1).

## RESULTS AND DISCUSSION

### Relationships among E.C.3.4.22. cysteine proteinase members

The evolutionary history of protease families may be regarded as the evolution from a single general-purpose ancestral protease to multiple and increasingly specific paralogous enzymes through a process of repeated gene duplication. There is biochemical evidence for this in relation to the papain superfamily in the trichomonads (North, 1991) and for cysteine-dependent proteases in *Giardia* (Parenti, 1989), two groups of protozoa, which were among the earliest diverging eukaryotes (Knoll, 1992). It therefor appears that the papain superfamily originated early during eukaryote evolution, and may, indeed, have occurred before the divergence of prokaryotes and eukaryotes.

The analysis of the phylogenetic tree (Figure 1) clearly shows the clustering of all cysteine proteinases members from clan CA and family C1, except cathepsin B. In fact, this major cluster could be sub-divided in two groups: group I containing the proteins more closely related to papain, the type member of the C1 family, and group II, formed by cathepsins H, K, L and S, which presents a weak internal branch support of 79%. Cathepsin B could be placed in group II by its enzymatic similarities with this group, but corresponds to a member that strongly diverged from the other members of the C1 family. The five other cysteine proteinases included in this analysis form a very heterogeneous cluster composed of two different clans (CA and CD) from five different families (C2, C13, C10, C11 and C14).

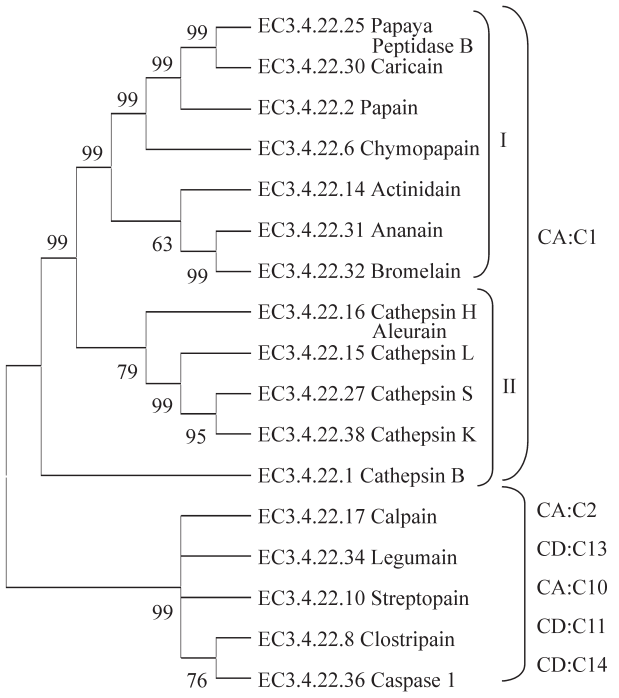


**Figure 1** - A phylogenetic unrooted tree of standard members of the cysteine proteinases E.C.3.4.22 sub-sub-class. The tree was constructed using MEGA software with the following parameters: p-distance, neighbor joining, pair-wise deletion, and internal branch with 5000 replications. The analysis was based on the MAP alignment of the 17 sequences. The values obtained from internal branch replication are indicated at the main branch points.

## Identification of SUCEST cysteine proteinase homologous sequences

All clusters present in the SUCEST data bank were automatically submitted to a general BLAST search against DNA, cDNA and protein data banks worldwide. This type of analysis allowed the identification of correlated sequences with the lowest and most significant e-values. Sometimes, however, these homologous sequences were not yet characterized or their real biological functions still await confirmation, and for this reason we have used the reverse approach to identify sugarcane cDNA clones with a real potential for presenting cysteine proteinase activity. In other words, the sequences of 24 well characterized and defined cysteine proteinases (17 of them presented in Figure 1) were used to find homologous SUCEST clusters having the lowest e-values.

The overall analysis, with a cut-off value of $e^{-5}$, allowed the identification of 76 different clusters separated into 12 distinct groups (Table III). Each group corresponds to one of the standard cysteine proteinases used in the T-Blast-n search. Several clusters were identified by more than one E.C. standard sequence. In Table III, a cluster is placed in a specific E.C. group only when the cluster has produced its lower and more significant e-values with the standard bait sequence of that corresponding group. The values may be considered parsimonious because some clusters present partial sequences as evaluated by the relative size ratio between the amino acids deduced from the sequence of the cluster and those present in the standard E.C. proteases (Table I).

Legumain (22 clusters) and actinidain (15 clusters) were the two most representative groups. Only one homologous cluster was found for calpain and bromelain and three clusters were identified as being common to papaya-protease-B, caricain and ananain. Papain group had four clusters, chymopapain eight clusters and cathepsins B, l and H, five, seven and four clusters respectively.

More than 26 clusters, with e-values lower than $e^{-30}$, were identified by the T-Blast-n algorithm when the search was performed with cathepsin-S (E.C.3.4.22.27) and cathepsin-K (E.C.3.4.22.38) sequences. These data are not presented in Table III because all the clusters detected with these two cathepsins presented lower e-values when other standard E.C.3.4.22 sequences were used for searching.

No significant match was found when ficin (E.C.3.4.22.03), asclepain (E.C.3.4.22.07), clostripain (E.C.3.4.22.8), streptopain (E.C.3.4.22.10), cathepsin-T (E.C.3.4.22.24), cancer pro-coagulant (E.C.3.4.22.26), fruit bromelain (E.C.3.4.22.33), histolysain (E.C.3.4.22.35), caspase-1 (E.C.3.4.22.26) and gingipain-R (E.C.3.4.22.37) sequences were used as bait in the T-Blast-n search of the SUCEST data bank.

## Classification and relationship of sugarcane cathepsin-like proteinases

The initial search and classification of sugarcane proteinases using T-Blast-n and e-values, allowed the identification of 16 clusters corresponding to cathepsin-like proteinases. To verify that these clusters were members of the cathepsin B, H and L group, their sequences were aligned (using the MAP algorithm) with 23 other cathepsin sequences. Due to the small size and no sequence overlapping of some cathepsin cluster, these small C- or N-terminal clusters were omitted from the file used to generate the cathepsin phylogenetic tree. The phylogenetic tree (Figure 2) obtained clearly shows the existence of three major groups, corresponding to each of the cathepsin classes (B, H and L) and are 100% supported by the internal branch test. Moreover, when cathepsin sequences from other plants were present, the sugarcane clusters showed a closer and statistically significant relationship with monocotyledon sequences. The close relationship of clusters SC B1, SC B2 and SC B3 may suggest that they were derived from a single ancestral gene now present in at least three active copies on the polyploid genome of sugarcane. The distinctive nature of the three previous cluster in relation to clusters SC B4 and SC B5 is also observable in the topology of the tree.

**Table III -** List of sugarcane cysteine proteinases grouped according to e-values in relation to standard EC numbers.

| e value | Cluster | Acronym | Size | e value | Cluster | Acronym | Size |
|---|---|---|---|---|---|---|---|
| EC 3.4.22.01 – Cathepsin B | | | 100 | EC 3.4.22.16 – Cathepsin H | | | 100 |
| 2e-70 | SCCCLB1023F09.g | SC B1 | 73 | 3e-88 | SCEPRT2048G06.g | SC H2 | 100 |
| 2e-57 | SCRULB1060C09.g | SC B2 | 95 | 3e-58 | SCCCRZ2001D03.g | SC H1 | 101 |
| 1e-31 | SCJFRT2057H07.g | SC B3 | 58 | 2e-33 | SCVPRZ2042B11.g | SC H3 * | 31 |
| 4e-27 | SCSGAD1142H04.g | SC B4 | 51 | 1e-29 | SCQGRT1039H08.g | SC H4 * | 67 |
| 5e-25 | SCJLRT1021B02.g | SC B5 | 50 | EC 3.4.22.17 - Calpain | | | 100 |
| EC 3.4.22.02 - Papain | | | 100 | 1e-39 | SCEQRT1030G12.g | SC Cal1 | 62 |
| 2e-38 | SCSGST3118H06.g | SC Pap1 * | 56 | EC 3.4.22.25 – Papaya pep B | | | 100 |
| 6e-28 | SCRLAD1140E01.g | SC Pap2 * | 51 | 5e-67 | SCSFRT2069E03.g | SC PPB1 * | 115 |
| 8e-28 | SCEQRT2026D05.g | SC Pap3 * | 42 | 2e-61 | SCAGLR1021H06.g | SC PPB2 * | 70 |
| 3e-22 | SCSBFL4067C01.g | SC Pap4 * | 57 | 1e-33 | SCEZSB1091H06.g | SC PPB3 * | 45 |
| EC 3.4.22.06 - Chymopapain | | | 100 | EC 3.4.22.30 - Caricain | | | 100 |
| 3e-95 | SCEQRT2028D09.g | SC Chy1 * | 109 | 8e-45 | SCRLLV1050G02.g | SC Car1 * | 50 |
| 1e-72 | SCVPRT2082G12.g | SC Chy2 * | 84 | 1e-37 | SCSGFL1078D08.g | SC Car2 * | 57 |
| 7e-71 | SCJFRT2054B11.g | SC Chy3 * | 89 | 2e-34 | SCQSRT3052E06.g | SC Car3 * | 31 |
| 2e-52 | SCEZFL4045D12.g | SC Chy4 * | 64 | EC 3.4.22.31 - Ananain | | | 100 |
| 3e-51 | SCCCFL1001E01.g | SC Chy5 * | 53 | 9e-31 | SCQGAM2110C03.g | SC Ana1 * | 120 |
| 7e-50 | SCAGFL3025D12.g | SC Chy6 * | 62 | 9e-27 | SCQSRT2032C02.g | SC Ana2 * | 56 |
| 1e-47 | SCSFFL4081E11.b | SC Chy7 * | 50 | 2e-22 | SCCCRT2001F03.g | SC Ana3 * | 48 |
| 1e-44 | SCEQRT2027C12.g | SC Chy8 * | 60 | EC 3.4.22.32 - Bromelain | | | 100 |
| EC 3.4.22.14 - Actinidain | | | 100 | 1e-16 | SCEQFL5052E10.g | SC Bro 1 * | 45 |
| 1e-103 | SCCCLR1022B11.g | SC Act01 * | 125 | EC 3.4.22.34 - Legumain | | | 100 |
| 1e-102 | SCEQLB1063D01.g | SC Act02 * | 118 | 1e-157 | SCJLLR1054F03.g | SC Leg07 | 109 |
| 5e-96 | SCQSRT2031E12.g | SC Act03 * | 127 | 1e-129 | SCJLST1020G04.g | SC Leg06 | 86 |
| 2e-79 | SCAGCL6012C06.g | SC Act04 * | 134 | 1e-115 | SCCCST1001F05.g | SC Leg03 | 60 |
| 1e-71 | SCVPRZ2041B10.g | SC Act05 * | 69 | 1e-68 | SCBFST3136H03.g | SC Leg17 * | 53 |
| 1e-65 | SCMCRT2105A06.g | SC Act06 * | 71 | 3e-65 | SCSGAM1095D04.g | SC Leg02 | 53 |
| 2e-50 | SCJLRT1023B01.g | SC Act07 * | 54 | 4e-65 | SCVPRT2081F05.g | SC Leg18 * | 45 |
| 2e-49 | SCEQRT1026D09.g | SC Act08 * | 58 | 3e-59 | SCEQAM1040D11.g | SC Leg08 | 46 |
| 7e-49 | SCEPCL6021B07.g | SC Act09 * | 56 | 5e-57 | SCRLAD1101F08.g | SC Leg09 | 44 |
| 6e-41 | SCJFRT1009G07.g | SC Act10 * | 71 | 5e-57 | SCJLRT1018A05.g | SC Leg14 | 46 |
| 1e-40 | SCEZHR1048C09.g | SC Act11 * | 43 | 7e-55 | SCJFRZ2034G04.g | SC Leg19 * | 45 |
| 7e-38 | SCRURT2012H10.g | SC Act12 * | 48 | 2e-54 | SCJFRT2053G04.g | SC Leg20 * | 40 |
| 1e-35 | SCQSHR1023F11.g | SC Act13 * | 29 | 3e-52 | SCAGLB1070C11.g | SC Leg12 | 43 |
| 1e-33 | SCJFRT1009E03.g | SC Act14 * | 31 | 1e-42 | SCQGLR1086C02.g | SC Leg11 | 38 |
| 6e-31 | SCQSRT2034F06.g | SC Act15 * | 42 | 3e-40 | SCBFST3134A09.g | SC Leg10 | 34 |
| EC 3.4.22.15 – Cathepsin L | | | 100 | 2e-32 | SCUTST3127D12.g | SC Leg13 | 35 |
| 7e-91 | SCUTSD1024F12.g | SC L1 | 101 | 1e-31 | SCSBRZ3126G06.g | SC Leg16 | 28 |
| 9e-51 | SCSGRT2066G02.g | SC L2 * | 95 | 3e-29 | SCJLST1020G05.g | SC Leg15 | 20 |
| 2e-50 | SCQSLR1061H11.g | SC L3 * | 93 | 3e-29 | SCEPAM2057E10.g | SC Leg21 * | 20 |
| 2e-44 | SCMCRT2085E04.g | SC L4 * | 68 | 1e-24 | SCPILB2021F10.g | SC Leg01 | 30 |
| 3e-44 | SCCCRT2002G01.g | SC L5 * | 81 | 8e-18 | SCRFHR1005D12.g | SC Leg05 | 22 |
| 1e-38 | SCEPSD2005A02.g | SC L6 * | 55 | 1e-12 | SCJLRZ3077C10.g | SC Leg22 * | 16 |
| 2e-32 | SCQGFL4073H01.g | SC L7 * | 76 | 7e-08 | SCVPLR1049F02.g | SC Leg04 | 22 |

*Clusters not used in the phylogenetic analysis, with acronym derived exclusively from the high e-value grouping. Size corresponds to the SUCEST cluster and the standard EC protease amino acid ratio.

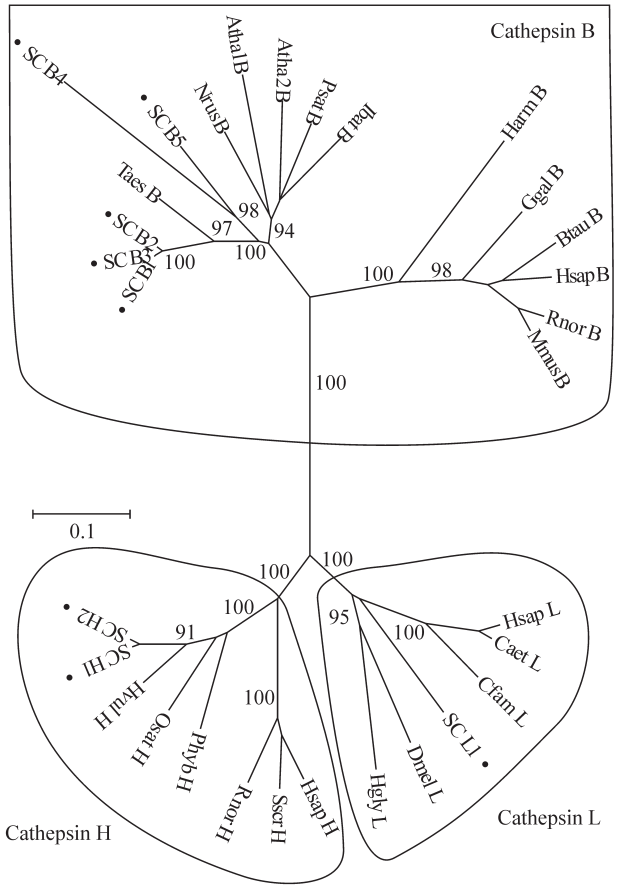


**Figure 2** - Phylogenetic unrooted tree of cathepsins B, H and L. The tree was constructed using Neighbor-joining and pair-wise deletion parameters, with an internal branch test with 5000 replications. The analysis was performed based on the MAP alignment of eight Sugarcane clusters, 12 cathepsin-B, six cathepsins-H and five cathepsin-L. The three major clusters, corresponding to each cathepsin class, are demarked. Black dots identify all sugarcane clusters. The values obtained from internal branch replication (90%) are indicated at the main branch points.

## Classification and relationship of SUCEST legumain-like proteinases

The sugarcane legumain clusters (Table III) were aligned with 16 plant legumain and another eight animal legumains in a similar way to that which was done with the cathepsins. Six clusters were left out from the phylogenetic analysis due to their small size and non-overlapping with the other sequences. Only regions encompassing the three N-terminal conserved domains were used to construct the legumain phylogenetic tree shown in Figure 3, where five major groups can be seen: group-I containing monocotyledons and 13 of the sugarcane clusters; group-II made up of dicotyledons and three sugarcane clusters contains a mixture of plant groups; group III containing only dicotyledons; group IV consisting of only vertebrates and group V comprised of invertebrates only. The existence of different legumains in a same species, possessing different protein targets and biological roles, have been described (Okamoto and Minamikawa, 1999). The presence of legumain clusters in group-I and group-II, allied to the heterogeneous pattern of legumains inside group-I, suggest that at least some of these clusters (SC Leg02, SC Leg03 and SC Leg05) may have differentiated cellular functions.

## Analysis of sugarcane cathepsin and legumain expression pattern

A preliminary analysis of the sugarcane cysteine proteinase expression pattern was made by the direct correlation of the reading frequency of each cluster in the different SUCEST cDNA libraries. Of the 76 cysteine proteinase clusters identified in the SUCEST bank, we focused the analysis on the cathepsin and legumain groups.

Analysis of cathepsin expression (Table IV) revealed that the most well represented cluster was SC B2, corre-

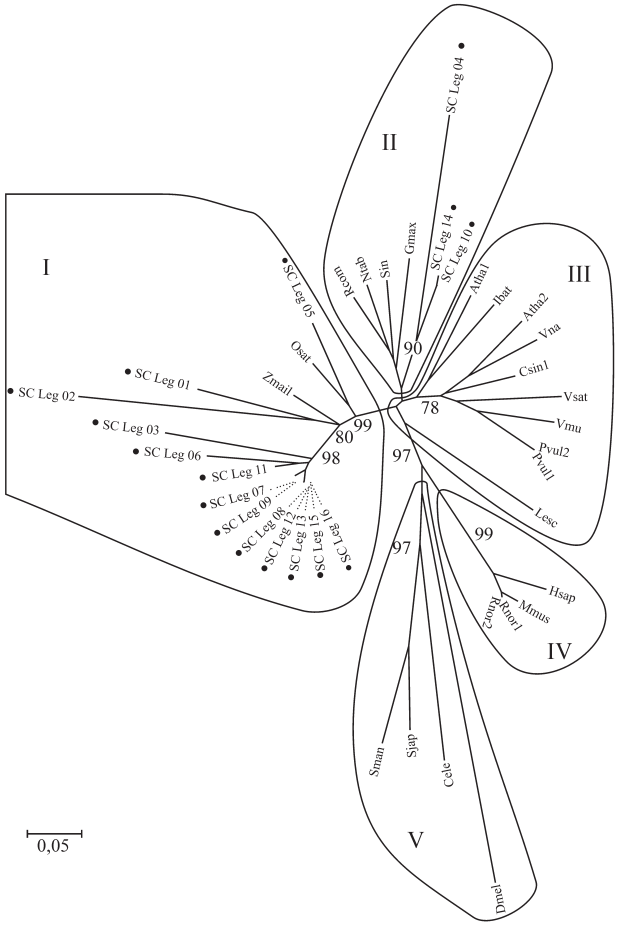


**Figure 3** - Legumain phylogenetic tree, constructed using MEGA software with the following parameters: p-distance, neighbor joining, pair-wise deletion and internal branch test with 5000 replications. The analysis was performed based on the MAP alignment of 16 sugarcane clusters, 16 plant legumains and clusters from another eight sequences from animals. The five major clusters represent (I) mono -cotyledonous plants, (II) dicotyledonous plants, (III) a mixed plant group, (IV) mammals and (V) invertebrates All sugarcane clusters are indicated by black dots. The values obtained from the internal branch replication (75%) are indicated at the main branch points.

sponding to a cathepsin B, which was present in almost all the libraries except for the callus and etiolated-leaf libraries. In general, cathepsin B and H reads were represented in libraries produced from vegetative organs and developing seeds but were rare in callus and flower libraries. Two cathepsin L clusters, SC L1 and SC L6, were found only in the seed library. Cathepsin H occurred less frequently, with a total of 39 reads, against 79 for cathepsin B and 70 for cathepsin L.

The most frequent and ubiquitous legumain cluster was SC Leg07, which was present in libraries constructed from apical meristem, flowers, lateral buds, etiolated leaves, stem and *Herbaspirillum* infected plants. On the other hand, a total of 11 clusters were derived from reads found in one type of library only, this being the case for cluster SC Leg09, which was present only in the library of *Gluconacetobacter*- inoculated plants and cluster SC Leg05, which occurred only in the library of *Herbaspirillum*-infected plants. Other clusters have been derived from stem (SC Leg13, SC Leg15 and SC Leg17), root (SC Leg14 and SC Leg20), leaf-root transition zone (SC Leg19), lateral bud (SC Leg01), and lateral root (SC Leg 04) libraries. However, more detailed analysis concerning the expression pattern of these genes will be necessary in order to confirm the tissue-specificity of these clusters/genes.

Apart from these differences, the presence of legumain in the vegetative organs of sugarcane agree with previously published work describing legumain expression in organs such as roots, leaves, flowers and hypocotyls (Kinoshita *et al*., 1995; Hara-Nishimura *et al*., 1998; Okamoto and Minamikawa, 1999). However, surprisingly, clusters sequenced from libraries constructed from germinating/developing seeds were very rare, and no clusters were derived from seeds only. Some clusters were found in other plant organs as well as seeds e.g. cluster SC Leg06 (20 cm flower stems, lateral buds and stems), cluster SC Leg03 (1 cm flowers and steams) and cluster SC Leg22 (leave-root transition zone). No direct correlation was observed between

**Table IV -** Frequency of sugarcane cathepsins B, L and H related reads and clusters on sugarcane cDNA libraries.

| Libraries * | AM | AM | FL | FL | FL | FL | FL | LB | LB | LR | LR | LV | NR | RT | RZ | SB | SD | ST | CC | CL | CL | CL | AD | HR | To- |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster | 1 | 2 | 1 | 3 | 4 | 5 | 8 | 1 | 2 | 1 | 2 | 1 | n | n | n | 1 | n | n | L3 | 4 | 6 | 7 | 1 | 1 | tal |
| Cathepsin B | | | | | | | | | | | | | | | | | | | | | | | | | |
| SC B1 | 2 | - | 1 | - | - | - | - | 1 | - | - | - | - | - | 2 | 1 | 1 | 1 | - | - | - | - | - | 1 | - | 10 |
| SC B2 | 1 | 1 | 1 | 3 | 1 | 3 | | 2 | 2 | 1 | - | - | - | 13 | 2 | 3 | 3 | 2 | - | - | - | - | 2 | 3 | 43 |
| SC B3 | 2 | 3 | 2 | - | - | - | - | - | - | - | - | - | - | 2 | 2 | 3 | - | - | - | - | - | 2 | - | 1 | 17 |
| SC B4 | - | - | - | - | - | - | - | - | 1 | - | - | 2 | - | - | - | - | - | - | - | - | - | - | 1 | - | 4 |
| SC B5 | - | - | - | - | - | - | 1 | - | - | - | - | - | - | 2 | 1 | - | - | 1 | - | - | - | - | - | - | 5 |
| Total | *5* | 4 | 4 | 3 | 1 | 3 | 1 | 3 | 3 | 1 | - | 2 | - | 19 | 6 | 7 | 4 | 3 | - | - | - | 2 | 4 | 4 | 79 |
| Cathepsin H | | | | | | | | | | | | | | | | | | | | | | | | | |
| SC H1 | - | - | 1 | - | - | - | - | - | - | 2 | - | - | - | 2 | 3 | - | 1 | 3 | - | - | 1 | - | 2 | 2 | 17 |
| SC H2 | - | - | - | - | - | - | - | - | - | - | 1 | 2 | - | - | - | 1 | 1 | - | - | - | - | - | - | - | 6 |
| SC H3 | - | - | - | - | - | - | 1 | - | - | - | - | - | - | 1 | 1 | - | - | - | - | - | - | - | - | - | 3 |
| SC H4 | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 5 | 3 | 2 | - | - | - | - | - | - | 1 | 13 |
| Total | - | - | 1 | - | - | - | 1 | - | - | 2 | 1 | 2 | - | 5 | 9 | 4 | 4 | 3 | - | - | 1 | - | 2 | 3 | 39 |
| Cathepsin L | | | | | | | | | | | | | | | | | | | | | | | | | |
| SC L1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 3 | - | - | - | - | - | - | - | 3 |
| SC L2 | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 2 | - | 4 | - | - | - | 1 | - | 1 | 1 | 10 |
| SC L3 | 2 | - | - | - | - | - | 1 | - | 2 | 2 | - | 2 | - | 2 | 5 | 1 | 4 | 3 | - | - | - | - | 1 | 1 | 26 |
| SC L4 | - | - | - | - | 2 | - | - | - | 2 | - | - | 1 | - | 3 | 2 | - | 3 | - | - | - | - | - | - | - | 13 |
| SC L5 | - | - | - | - | - | - | - | - | - | - | - | - | - | 4 | 2 | - | - | 2 | - | - | - | - | 1 | - | 9 |
| SC L6 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 3 | - | - | - | - | - | - | - | 3 |
| SC L7 | - | - | - | 1 | 1 | - | - | - | - | - | - | 1 | - | - | - | - | 1 | 1 | - | - | - | - | 1 | - | 6 |
| Total | 2 | - | - | 1 | 3 | - | 1 | - | 4 | 2 | - | 4 | - | 10 | 11 | 1 | 18 | 6 | - | - | 1 | - | 4 | 2 | 70 |

*The libraries were are as follows: apical meristem from mature (AM1) and (AM2) immature plants; 1 cm (FL1) and 5 cm (FL3) flower base; 50 cm (FL4), 20 cm (FL5) and10 cm (FL8) flower stem; lateral buds (LB1 and LB2); large (LR1) and small (LR2) leaf-root insert libraries; etiolated leaves (LV1); grouped data of two non-redundant libraries (NRn); grouped data of three root libraries (RTn); grouped data of three leaf-root transition zone libraries (RZn); stem bark (SB1); grouped data of two seed libraries of different insert sizes (SDn); grouped data of two stem libraries from the first and fourth internodes (STn); libraries derived from *calli* submitted to a 4-37 °C temperature change and three (CL3), four (CL4), six (CL6) and eight (CL7) hours of a light/dark cycle; plants infected with the bacteria *Acetobacter diazotroficans* (AD1) and *Herbaspirillum rubrisubalbicans* (HR1).

**Table V -** Frequency of sugarcane legumain related reads and clusters on sugarcane cDNA libraries.

| Libraries* | AM | AM | CL | CL | CL | CL | FL | FL | FL | FL | FL | LB | LB | LR | LR | LV | NR | RT | RZ | SB | SD | ST | AD | HR | To- |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 6 | 7 | 1 | 3 | 4 | 5 | 8 | 1 | 2 | 1 | 2 | 1 | n | n | n | 1 | n | 1 | 1 | 1 | tal |
| *Legumain* | | | | | | | | | | | | | | | | | | | | | | | | | |
| SC Leg01 | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | - | - | - | - | - | 1 |
| SC Leg02 | 1 | - | - | - | - | - | 1 | - | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | 3 |
| SC Leg03 | - | - | - | - | - | - | 2 | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 1 | - | - | 4 |
| SC Leg04 | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | - | - | - | - | 1 |
| SC Leg05 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 1 |
| SC Leg06 | - | 1 | - | - | - | - | - | - | - | - | 1 | - | 2 | - | - | - | - | - | - | - | 2 | 2 | - | - | 8 |
| SC Leg07 | 1 | - | - | - | - | - | - | 1 | - | - | - | - | 2 | 2 | - | 1 | - | - | - | - | - | 4 | - | 2 | 13 |
| SC Leg08 | 1 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | 1 | - | 2 | - | - | 1 | 1 | 8 |
| SC Leg09 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | - | 2 |
| SC Leg10 | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | - | - | 2 | 1 | - | 4 |
| SC Leg11 | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 1 | - | - | - | - | - | - | - | 1 | - | - | 3 |
| SC Leg12 | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | - | - | - | - | 1 | 1 | 3 |
| SC Leg13 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | 1 |
| SC Leg14 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | - | - | - | - | - | - | 2 |
| SC Leg15 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | 1 |
| SC Leg16 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 1 | - | - | - | - | 2 |
| SC Leg17 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | - | - | 2 |
| SC Leg18 | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | - | - | - | - | - | 2 |
| SC Leg19 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | - | - | - | - | - | 2 |
| SC Leg20 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 2 | - | - | - | - | - | - | 2 |
| SC Leg21 | - | 1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 |
| SC Leg22 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 1 | - | 1 | - | - | - | 2 |
| Total | 4 | 3 | - | - | - | - | 3 | 1 | - | - | 1 | 1 | 7 | 4 | - | 2 | - | 7 | 4 | 3 | 4 | 15 | 5 | 5 | 68 |

*The libraries were are as follows: apical meristem from mature (AM1) and (AM2) immature plants; 1 cm (FL1) and 5 cm (FL3) flower base; 50 cm (FL4), 20 cm (FL5) and10 cm (FL8) flower stem; lateral buds (LB1 and LB2); large (LR1) and small (LR2) leaf-root insert libraries; etiolated leaves (LV1); grouped data of two non-redundant libraries (NRn); grouped data of three root libraries (RTn); grouped data of three leaf-root transition zone libraries (RZn); stem bark (SB1); grouped data of two seed libraries of different insert sizes (SDn); grouped data of two stem libraries from the first and fourth internodes (STn); libraries derived from *calli* submitted to a 4-37 °C temperature change and three (CL3), four (CL4), six (CL6) and eight (CL7) hours of a light/dark cycle; plants infected with the bacteria *Acetobacter diazotroficans* (AD1) and *Herbaspirillum rubrisubalbicans* (HR1).

sugarcane sequence homology and the expression pattern of different clusters. Thus clusters SC Leg07 and SC Leg09, very close in the tree (Figure 3), differ significantly in their expression pattern (Table V), although clusters SC Leg01 and SC Leg12, which belong to different groups (Figure 3), present a relatively similar expression pattern.

At this stage speculations about a role for different sugarcane proteases may be premature because the analysis of the SUCEST database is still producing new data, but it is hoped that the results presented here will contribute to the understand of the role of cysteine proteases in sugarcane. Much work remains to be done in order to confirm the observed expression patterns and to assess the real biological diversity of each of the cysteine proteinases revealed in this study.

## ACKNOWLEDGMENTS

G.C. Corrêa was supported by a CNPq fellowship. The authors wish to thank Dr. G. Domont and Dr. G. Sachetto-Martins for the critical reading of the manuscript and FAPESP for support and development of the SUCEST project.

## RESUMO

Proteinases cisteínicas são peptidil-hidrolases dependentes de um resíduo de cisteína em seu sítio ativo. As propriedades físico-químicas destas proteinases têm sido amplamente caracterizadas, entretanto suas funções biológicas ainda não foram completamente elucidadas. Elas estão envolvidas em um grande número de eventos, tais como: processamento e degradação protéica, câncer, ger-

minação, morte celular programada e processos de senescência. Diferentes proteinases cisteínicas, classificadas pelo Comitê de Nomenclatura da União Internacional de Bioquímica e Biologia Molecular (IUBMB) como pertencentes à sub-sub-classe E.C.3.4.22, foram usadas na busca de clusters no banco de dados do SUCEST (SUgarCane EST project), utilizando-s o programa T-BLAST-n. Homologia de seqüências foram encontradas com 76 clusters que correspondem a prováveis proteinases cisteínicas. O alinhamento destas seqüências com a de outras proteases cisteínicas, de diversas origens, forneceu informação quanto à classificação e possível função das proteinases de cana-de-açúcar. Além disso, o padrão de expressão de cada gene foi postulado a partir da correlação direta com as bibliotecas de cDNA do SUCEST dos quais os clusters foram derivados. Uma vez que nenhum gene de protease cisteínica foi anteriormente evidenciado em cana-de-açúcar, este estudo representa uma etapa inicial para o estudo de novos aspectos bioquímicos, fisiológicos e biotecnológicos destas enzimas.

## REFERENCES


**Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W.** and **Lipman, D.J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. *25*: 3389-3402.

**Bairoch, A.** (2000). The ENZYME database in 2000. *Nucleic Acids Res*. *28*: 304-305.

**Barrett, A.J.** (1980). Introduction: the classification of proteinases. Ciba Found. Symp. *75*: 1-13.

**Chappell, C.L.** and **Dresden, M.H.** (1986). Schistosoma mansoni: proteinase activity of "hemoglobinase" from the digestive tract of adult worms. *Exp. Parasitol 61*: 160-7.

**Croall, D.E.** and **Demartino, G.N.** (1991). Calcium activated neutral protease (calpain) system: structure, function, and regulation. *Physiol. Rev*. *71*: 813-847.

**Hara-Nishimura, I., Kinoshita, T., Hiraiwa, N.** and **Nishimura, M.** (1998). Vacuolar processing enzymes in protein storage vacuoles and lytic vacuoles. *J.Plant Physiol. 152*: 668-674.

**Hiraiwa, N., Takeuchi, Y., Nishimura, M.** and **Hara-Nishimura I.** (1993). A vacuolar processing enzyme in maturing and germinating seeds: its distribution and associated changes during development. *Plant Cell Physiol*. *34*: 1197-1204.

**Huang, X.** (1994). On Global Sequence Alignment. Computer Applications in the *Biosciences 10*: 227-235.

**Ishii, S.I.** (1994). Legumain: asparaginyl endopeptidases. *Methods Enzymol*. *244*: 604-615.

**Kembhavi, A.A., Buttle, D.J., Knight, C.G.** and **Barret, A.** (1993). The two cysteine endopeptidase of legume seeds: purification and characterization by use of specific fluorometric assays. *Arch. Biochem. Biophys*. *303*: 208-213.

**Kinoshita, T., Nishimura, M.** and **Hara-Nishimura, I.** (1995). Homologues of a vacuolar processing enzyme that are expressed in different organs in Arabidopsis thaliana. *Plant Mol.Biol*. *29*: 81-89.

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

**Koehler, S.M.** and **Ho T.H.D.** (1990). Hormonal regulation, processing, and secretion of cysteine proteinases in barley aleurone layers. *Plant Cell 2*: 769-783.

**Kumar, S., Tamura, K., Jacobsen, I.** and **Nei, M.** (2000). MEGA2: Molecular Evolutionary Genetics Analysis, version 2.0. Pennsylvania and Arizona State Universities, University Park, Pennsylvania and Tempe, Arizona.

**Linthorst, H.J.M., van der Does, C., van Kan, J.A.L.** and **Bol, J.F.** (1993). Nucleotide sequence of a cDNA clone encoding tomato (*Lycopersicon esculentum*) cysteine proteinase. *Plant Physiol*. *101*: 705-706.

**Mort, J.S.** (1998). Cathepsin B. In *Handbook of proteolytic enzymes* (Barrett, A.J., Rawlings, N.D. and Woessner, J.F., eds.), pp. 609-617, Academic Press, London.

**North, M.J.** (1991). Proteinases of trichomonads and Giardia. In: *Biochemical Protozoology* (Combs, G.H. and North, M.J., eds.) pp. 234-244, Taylor and Francis, New York.

**Okamoto, T.** and **Minamikawa, T.** (1999). Molecular cloning and characterization of Vigna mungo processing enzyme 1 (VmPE-1), an asparaginyl endopeptidase involved in post-translational processing of a vacuolar cysteine endopeptidase (SH-EP). *Plant. Mol. Biol. 39*: 63-73.

**Parenti, D.M.** (1989). Characterization of a thiol proteinases in Giardia lamblia. *J.Infect.Dis. 160*: 1076-1080.

**Rawlings, N.D.** and **A.J. Barrett, A.J.** (2000). MEROPS: the peptidase database. *Nucleic Acids Res*. 28: 323-325.

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

**Sitnikova, T., A. Rzhetsky,** and **M. Nei.** (1995). Interior-branch and bootstrap tests of phylogenetic trees. *Mol. Biol. Evolution 12*: 319-333.

**Watanabe H., Abe, K., Emori, Y., Hosoyama, H.** and **Arai, S.** (1991). Molecular cloning and gibberellin-induced expression of multiple cysteine proteinases of rice seeds (Oryzains) *J.Biol.Chem*. *266*: 16897-16902.

**Yeh, E.T., Gong, L.** and **Kamitani, T.** (2000). Ubiquitin-like proteins: new wines in new bottles. *Gene 248*: 1-14.