

Proteases: Evolution

Christopher Southan, *Oxford GlycoSciences Ltd, Abingdon, UK*

c1 email:

Human proteases have diverse fold structures, catalytic mechanisms, biochemical functions and evolutionary origins. Most arose by gene duplication, but some by horizontal transfer from prokaryotes. Adaptation and diversification of their biological roles have resulted from shifts in substrate specificities, expression patterns and new domain combinations.

Intermediate article

Article contents

- Diversity
- Numbers and Classification
- Structural Homology
- Gene Duplication and Domain Accretion
- Lineage-specific Distribution
- Inactive Homologs
- The HtrA Proteases
- Conclusions and Prospects

Diversity

1141.1 The challenge of reviewing the evolution of human proteases is the diversity of this enzyme group in function, catalytic mechanism, structure and origin. For example, the aspartic acid–histidine–serine (Asp–His–Ser) catalytic triad is shared by the trypsin-like proteases, subtilisin-like proteases, esterases and $\alpha\beta$ -hydrolases, which include lipases as well as proteases. The catalytic mechanism has converged to the same active site geometry, but evolved independently on different structural scaffolds (Krem and Di Cera, 2001). The catalytic units of three of these ancient protease superfamilies have evolved by duplication and adaptation in the human genome. There are approximately 110 catalytic sequences for the trypsins, nine for the subtilisins and six for the $\alpha\beta$ hydrolase proteases. For the cysteine proteases with a common mechanism based on catalytic dyads of cysteine and histidine, as many as seven different origins have given rise to approximately 120 human sequences (Barrett and Rawlings, 2001). Given the broad diversity of both catalytic and noncatalytic human protease sequences, it is possible to review only selected aspects in this article.

c2

Numbers and Classification

1141.2 In addition to catalytic mechanism and fold architecture, proteases can be classified by extended sequence alignments, biochemical function and the combination of domains ancillary to the catalytic domain. All of these attributes provide insight into the evolutionary history of proteases. By grouping protease sequences at the species and family level, the MEROPS protease database has provided an alignment-based protease classification system that will be used in this article (Rawlings *et al.*, 2002). The mechanistic distribution

within the current MEROPS total of 498 human proteases includes 3% aspartyl, 23% cysteine, 36% metallo and 32% serine. Although the ratio of this distribution is broadly similar for many organisms, there has been a proportional increase in cysteine and metalloproteases in vertebrates that may be at least partially associated with their evolutionarily more recent roles in apoptosis and tissue modeling respectively. Across all organisms so far sequenced, the proportion of proteases against genome size ranges from approximately 1.5% to 2.5%. So, even assuming a conservative human gene number of 35 000, we could expect to find in the order of another 100 sequences in the finished genome. The experimental characterization of new proteases and the application of more sensitive methods of homology detection are likely to further increase this total.

Structural Homology

Similarity in biochemical function or catalytic mechanism in proteases is not a reliable measure of evolutionary relatedness. Although the sequence alignment of catalytic domains is a key method of defining homology, there are many sequence families that have diverged beyond the limits of conventional similarity detection methods. The fact that only 10% of human proteases, or 2% of all proteases, have a solved structure means that fold space is being filled in at a much slower rate than the paralogous and orthologous expansion of sequence families. However, it is increasingly clear that new structures and sensitive methods of comparing them are revealing new evolutionary relationships. Protease sequence families not previously shown to be related by sequence can now be connected at the superfamily level by being shown to have homologous folds. An example from

1141.3

the metalloproteases is the finding that the nicastrin transmembrane glycoprotein, part of the multicomponent alzheimer-linked protein complex, is a new member of the aminopeptidase superfamily, which includes the transferrin receptor (Fagan *et al.*, 2001). A second example from the cysteine proteases is the identification of numerous, previously undetected members of the caspase-hemoglobinase superfamily, which includes a new protease sequence family typified by the HetF protein from the cyanobacterium *Nostoc* (Aravind and Koonin, 2002).

Gene Duplication and Domain Accretion

1141.4 The large protease sequence families in the human genome were originally created by gene duplication. The exceptions are endogenous proteases of viral ancestry that have arisen by retrotransposition and those with high sequence similarity to prokaryotic homologs that have entered metazoan genomes by horizontal transfer. Analysis of recent duplications suggests that new paralogs are initially retained because they offer short-term advantages of protein dosage regulation, and that they can divide the multiple functions of the ancestral gene between them. In the longer term, functional diversion and/or changes in the expression context are likely to provide advantages.

1141.5 Although the diversity of biological functions of proteases is partially a result of the variety of substrate specificities, another important factor in their evolution is domain exchange by exon shuffling that occurred during the 'big bang' of metazoan radiation. The domain distribution in proteases is documented by InterPro, a comprehensive protein family and domain database that provides automated annotation for all proteins (Apweiler *et al.*, 2000). This facilitates comparative searches of all the known domains that are combined with protease catalytic modules and can be used to infer evolutionary history (Southan, 2000).

1141.6 A good example is the complex set of domain permutations that have occurred in the S1 trypsins. The current repertoire of domains and subfamilies that have been detected by the InterPro analysis for the S1 trypsin proteases is shown in **Table 1**.

1141.7 The PDZ domain is an ancient example which, like the trypsin family it has been combined with, has been retained over billions of years in both eukaryotic and prokaryotic sequences (Ponting, 1997). Others such as apple, C-lectin, fibronectin type 1, kringle and GLA domains are recent metazoan innovations for cell–cell communication roles that are absent from plants.

Table 1 Domains and subfamilies associated with trypsin-like serine proteases in the human genome

Identifier	Description	?????	C4
IPR000001	Kringle domain	17	
IPR000024	Frizzled CRD region domain	1	
IPR000082	SEA domain	3	
IPR000083	Fibronectin, type I domain	4	
IPR000294	γ -carboxyglutamic (GLA) domain	8	
IPR000436	Sushi domain/SCR repeat/CCP module	16	
IPR000561	EGF-like domain	19	
IPR000562	Type II fibronectin collagen-binding domain	2	
IPR000859	CUB domain	10	
IPR000867	Insulin-like growth factor-binding protein, IGFBP domain	3	
IPR000998	MAM domain	1	
IPR001190	Speract/scavenger receptor Family	15	
IPR001478	PDZ/DHR/GLGF domain	4	
IPR001940	HtrA/DegQ protease	4	
IPR002035	Von Willebrand factor, type A	4	
IPR002172	Low density lipoprotein-receptor, class A domain	14	
IPR002350	Serine protease inhibitor, Kazal type domain	3	
IPR003006	Immunoglobulin/major histocompatibility complex domain	1	
IPR003014	N/apple PAN domain	8	
IPR003884	Factor I membrane attack complex domain	1	

Identifiers and descriptions are taken from the InterPro database.

Lineage-specific Distribution

1141.8 Genomic data now allow us to compare the phylogenetic distribution of substantially complete protease sequence families between yeast, worm, fly, mouse and human. These data sets have now been complemented by an extensive set from *Fugu rubripes*. This fish lineage diverged from humans about 450 million years (Myr) ago. A comparison between these two distant genomes provides insights into the evolutionary changes that have shaped the two vertebrates. The theory that the human genome is derived from ancient octaploidy would predict an incremental expansion of gene families related to organism complexity (Gibson and Spring, 2000). In fact, protease families from the model organisms with completed genomes show uneven distributions that present a challenge for evolutionary interpretation. A selection of these is show in **Table 2**.

1141.T002 **Table 2** Distribution of selected protease families

	<i>Saccharomyces cerevisiae</i> 112	<i>Caenorhabditis elegans</i> 360	<i>Drosophila melanogaster</i> 529	<i>Fugu rubripes</i>	<i>Mus musculus</i> 431	<i>Homo sapiens</i> 493
S1 trypsin	1(1)	9(3)	220(42)	99	133(9)	108(16)
S1 kallikrein	–	–	–	(na*)	22	15
S1C HtrA	1	–	1	4	4	4
M13 neprylysin	–	22(6)	23(7)	6	6	7
M2 ACE	–	1(1)	6(2)	2	2	2
S12 lactamase	–	4(4)	–	2(2)	2(2)	2(2)

C5 *Not applicable

Total and individual protease numbers are taken from MEROPS, except for those of *Fugu*. These provisional numbers were determined by BLAST (Basic Local Alignment Search Tool) searches on the Ensembl *Fugu* genome assembly release 7.1.3, with 31 095 genes. Numbers in parentheses are likely to be inactive members of that sequence family, as judged by the absence of critical active site residues.

1141.9 The S1 chymotrypsin-like serine proteases comprise two homologous domains containing six-stranded β -barrels, with the active site between the domains. The constraints on binding site integrity, active site geometry and the necessity to maintain satisfactory packing of the β -barrels has maintained this common fold over an evolutionary span of more than 1 000 000 000 years. The yeast contains a single gene product that has duplicated domains homologous to trypsin, plus a tetrad of PDZ domains (Pallen and Wren, 1997). This classifies the sequence as an S1C, or HtrA-like protease, but the sequence similarity of the trypsin domains is so low they are likely to be inactive. The worm has nine S1 sequences. The fly shows a substantially larger number of S1 proteases, exceeding any organism so far sequenced. It is surprising that these sequences have been maintained rather than degenerating into pseudogenes, but the biological roles for most of them, particularly the 20% inactive members, remains unclear. Intriguingly, the first assemblies of mosquito sequence suggest a preliminary count of 278 S1 proteases. This suggests that the large S1 complement is not just a clade-specific expansion restricted to *Drosophila*, but may be general to the insects. In humans and mice, we see a smaller number of overall S1 homologs, but a greater proportion of potentially active sequences. Even across a relatively short divergence time of approximately 40 Myr, the mouse has selectively expanded just one subfamily within S1, the kallikreins; however, since the duplication event, some of the 30 loci have become pseudogenes. A recent detailed comparison of human chromosome 19, which includes all 15 human kallikrein sequences on 19q13, with the syntenic loci in mice suggest that this selective local duplication may be specific to the rodent lineage against what, from the preliminary analysis of the fish genomes, are beginning to look like similar total gene numbers across all vertebrates (Dehal *et al.*, 2001). The biological selective

advantage of maintaining the duplicated sequences in rodents is not yet clear.

The pattern for the M13 proteases looks opposite to that of the S1 family. Here the numbers in mouse and human are only 30% of those in the fly and worm. One evolutionary explanation that can be offered is non-paralogous functional substitution – that is, that other non-M13 proteases in vertebrates may have taken over the roles of M13 members in fly and worm (Coates *et al.*, 2000). The absence of M13s in yeast suggests that M13s may be associated with multicellularity. Another metalloprotease family, the M2, or angiotensin converting, enzymes, may also be associated with multicellularity. However, these enzymes show what appears to be anomalous expansion of inactive paralogs in the fly, as well as two active members, Ance and ACER. The latter is a *Drosophila* ACE2 homolog involved in heart morphogenesis. Recent genetic data for mouse ACE2 show that it is an essential regulator of heart function *in vivo* (Crackower *et al.*, 2002). Mammalian ACE is responsible for the synthesis of angiotensin II and the inactivation of bradykinin, but the absence of similar peptide hormones in insects suggests other peptide-processing functions for Ance. Clearly, both ACE paralogs diverged early in metazoan evolution, with ACE1 adapting to a quite different function in mammals while ACE2 retained a role in heart development.

Inactive Homologs

Evolutionary considerations must also account for an important observation that is often neglected because of the experimental focus on the biochemistry of proteolysis. This is the fact that a significant proportion of all protease sequence families now revealed by metazoan genomic data are, as judged by the absence of critical active site residues, unlikely to be active

proteases. These need to be considered in any evolutionary assessment of proteases. The MEROPS database has assigned 12% of all human protease sequences in this category. All major protease families in the human genome include at least one inactive paralog, and the ADAM (M12) family has the largest proportional representation (34%). These are on different chromosomal locations, suggesting the drift into inactivity has been stochastic rather than restricted to a few major duplication events. Some of these proteases have mouse orthologs that are also inactive. Few inactive protease homologs have an established function. Inactive homologs revealed in genomic data are worthy of biochemical investigation, and some of them have the same functional set of ancillary domains as their catalytically active paralogs. Haptoglobin is one example of a trypsin homolog that acts as a transport glycoprotein to remove free hemoglobin from the circulation of vertebrates.

1141.12

An interesting example of functional drift has been reported for the recently discovered vertebrate β -lactamases (Smith *et al.*, 2001). Until recently, all members of the S12 sequence family, which includes penicillin-binding proteins and carboxypeptidases, were of bacterial origin. These enzymes use a catalytic dyad, where serine acts as a nucleophile and lysine as a general base. The fold architecture consists of an amino (*N*)-terminal α/β cluster with five β strands, and a carboxy (*C*)-terminal region that contains five helical segments. A recent publication has described homologs from human, mouse and *Caenorhabditis elegans*. Any protein similarity matches from *Saccharomyces cerevisiae* and *Drosophila melanogaster* were notably absent (Table 2). Alignment of the LACT-1 protein with selected matches from microbial phyla showed that outside the S-X-X-K catalytic serine motif, the metazoan sequences did not conform to any known catalytic β -lactamases. In addition, the murein-containing cell wall components that are substrates for microbial β -lactamases are not endogenous to metazoans. An evolutionary explanation has been provided by experiments that suggest this sequence has been recruited into the large subunit (39 S) complex of the mammalian mitochondrial ribosome. This shift to a structural role could explain the loss of catalytic activity by evolutionary drift (Koc *et al.*, 2001). However, this interpretation is complicated by the finding of multiple, and probably inactive, homologs, not only in *C. elegans* but also in humans and *Fugu* (Table 2).

The HtrA Proteases

1141.13

Considered a paralogous sequence family, human HtrA proteases illustrate a number of key evolutionary principles (Table 2). They are defined by the combina-

tion of a trypsin catalytic unit with a *C*-terminal PDZ domain and are now recognized as a new class of molecular chaperones. Recent structures show a funnel-shaped trimer cage with the trypsin domains at the top and the PDZs protruding to the outside (Krojer *et al.*, 2002). The role of the PDZ domain has been adapted during evolution to determining substrate specificity and/or to regulate protease activity. Thus, ancillary domains can coevolve with the catalytic units with which they are combined. Their sequences show 35% amino acid identity with *Synechocystis* and other α -proteobacterial homologs. The HtrAs thus join the increasing number of human proteases with a prokaryotic ancestry suggestive of horizontal transfer. The phylogram in Figure 1a indicates the possible evolutionary relationships between the *Synechocystis* sequence, the single *Drosophila* HtrA, and the four paralogs found in both the human and *Fugu* genomes. The vertebrate homologs can be divided into two groups. The human mitochondrial HtrA2 possesses a transmembrane anchor, and a large section of the *N*-terminus is removed by processing (Figure 1b). Together with the *Synechocystis* sequence, the *Drosophila* sequence and one *Fugu* homolog, these form an outgroup of mitochondrial enzymes. The human HtrA1, 3 and 4 proteins, with their probable *Fugu* orthologs, form a second outgroup. These all contain predicted signal peptides as well as *N*-terminal sections that are recognized as IGF-binding and protease inhibitor domains. The simplest explanation is that the human HtrA2-like sequence in the fly represents the first metazoan ancestors from which the other three vertebrate paralogs arose by duplication. An exception to metazoan distribution is the absence of any detectable *C. elegans* homologs. This suggests the biochemical roles of the mitochondrial HtrA may have been substituted in worm phyla, possibly by another chaperone protease system. The fish sequences suggest that domain shuffling has introduced the 'new' *N*-terminal domains early in the vertebrate lineage. Inspection of the human gene structure (Figure 1) for the HtrAs indicates the possible evolutionary traces of this event. In human HtrA2, at 2p13.1, the transmembrane domain and mitochondrial targeting sequence are in exon 1, the trypsin domain spans exons 2, 3, 4 and 5, and the PDZ domain is located in the 3'-most exons 6 and 7. The human HtrA1 gene on 10q26.2 spans a much larger genomic region, but both 'new' domains are within exon 1. The rest of the gene shows similar organization to that encoding human HtrA2. The highest sequence similarity of the *N*-terminal domains is with the Q8WX77 Mac protein on 9p12, suggesting the 5' section of this gene, or at least its ancestral homolog, is a candidate for the 'shuffled-in' first exon in HtrA1, 3 and 4.

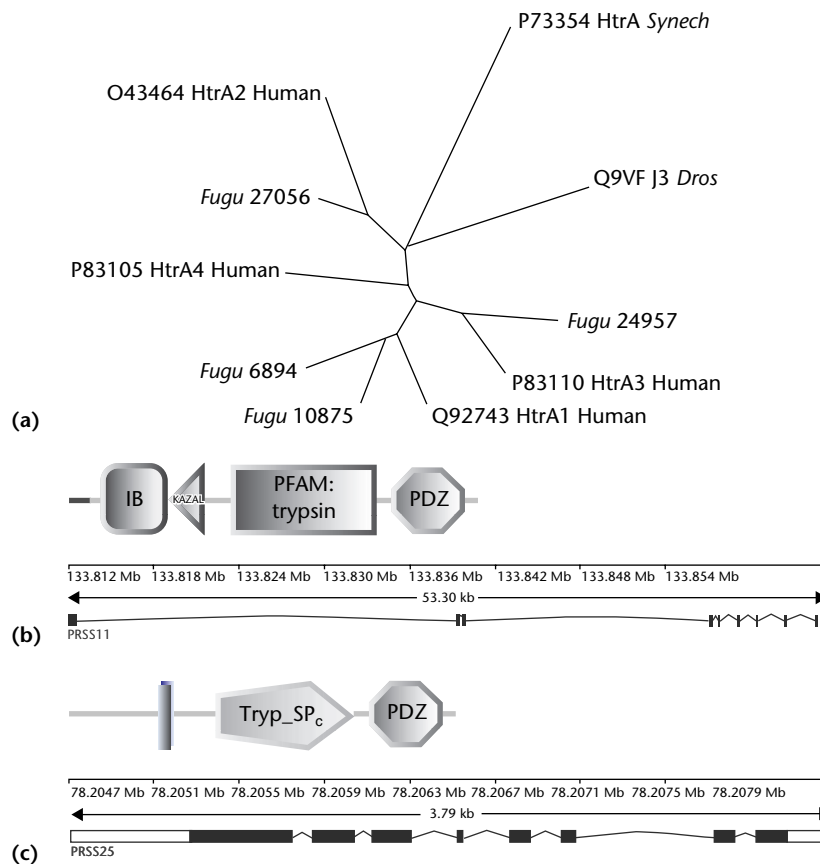


Figure 1 (a) Phylogram of HtrA proteases. (b) Human HtrA1 Q92743 domain and gene structure. (c) Human HtrA2 O43464 domain and gene structure. kb: kilobases; Mb: megabases.

Conclusions and Prospects

Although much has already been revealed about human protease evolution from comparative genomics, more information needs to be filled in. Of the 500 or so human sequences, a large proportion remain uncharacterized biochemically *in vitro*. Even less have had physiological substrates identified *in vivo*. Structural coverage is another experimental shortfall, but structural genomics initiatives are expected to narrow this gap and increase the discovery of hitherto unsuspected homologies. Comparative sequence space will be expanded by having two insects (when *Anopheles* is complete), two fish (*Fugu* and *Danio*) and two mammals when the mouse assembly approaches completion. The other advantage of genomic comparisons is facilitating the identification of tissue and developmental stage-specific transcriptional control regions. These are key postduplication adaptations as paralogous proteases evolve patterns of differential expression.

References

- Apweiler R, Attwood TK, Bairoch A, *et al.* (2000) InterPro – an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16**: 1145–1150.
- Aravind L and Koonin EV (2002) Classification of the caspase-hemoglobinase fold: detection of new families and implications for the origin of the eukaryotic separins. *Proteins* **46**: 355–367.
- Barrett AJ and Rawlings ND (2001) Evolutionary lines of cysteine peptidases. *Biological Chemistry* **382**: 727–733.
- Coates D, Siviter R and Isaac RE (2000) Exploring the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes to understand neuropeptide and peptidase function. *Biochemical Society Transactions* **28**: 464–469.
- Crackower MA, Sarao R, Oudit GY, *et al.* (2002) Angiotensin-converting enzyme 2 is an essential regulator of heart function. *Nature* **417**: 822–828.
- Dehal P, Predki P, Olsen AS, *et al.* (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science* **293**: 104–111.
- Fagan R, Swindells M, Overington J and Weir M (2001) Nicastrin, a presenilin-interacting protein, contains an aminopeptidase/transferrin receptor superfamily domain. *Trends in Biochemical Sciences* **26**: 213–214.
- Gibson TJ and Spring J (2000) Evidence in favour of ancient octaploidy in the vertebrate genome. *Biochemical Society Transactions* **28**: 259–264.

- Koc EC, Burkhardt W, Blackburn K, *et al.* (2001) The large subunit of the mammalian mitochondrial ribosome. Analysis of the complement of ribosomal proteins present. *Journal of Biological Chemistry* **276**: 43 958–43 969.
- Krem MM and Di Cera E (2001) Molecular markers of serine protease evolution. *The EMBO Journal* **20**: 3036–3045.
- Krojer T, Garrido-Franco M, Huber R, Ehrmann M and Clausen T (2002) Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine. *Nature* **416**: 455–459.
- Pallen MJ and Wren BW (1997) The HtrA family of serine proteases. *Molecular Microbiology* **26**: 209–221.
- Ponting CP (1997) Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Science* **6**: 464–468.
- Rawlings ND, O'Brien E and Barrett AJ (2002) MEROPS: the protease database. *Nucleic Acids Research* **30**: 343–346.
- Smith TS, Southan C, Ellington K, *et al.* (2001) Identification, genomic organization, and mRNA expression of *LACTB*, encoding a serine β -lactamase-like protein with an amino-terminal transmembrane domain. *Genomics* **78**: 12–14.
- Southan C (2000) Website review: InterPro (the integrated resource of protein domains and functional sites). *Yeast* **17**: 327–334.
- Koonin EV and Aravind L (2002) Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death and Differentiation* **9**: 394–404.
- Krem MM and Cera ED (2002) Evolution of enzyme cascades from embryonic development to blood coagulation. *Trends in Biochemical Sciences* **27**: 67–74.
- Massova I, Kotra LP, Fridman R and Mobashery S (1998) Matrix metalloproteinases: structures, evolution, and diversification. *FASEB Journal* **12**: 1075–1095.
- Patthy L (1999) Genome evolution and the evolution of exon-shuffling – a review. *Gene* **238**: 103–114.
- Ponting CP and Russell RR (2002) The natural history of protein domains. *Annual Review of Biophysics and Biomolecular Structure* **31**: 45–71.
- Southan C (2001) A genomic perspective on human proteases. *FEBS Letters* **498**: 214–218.
- Volker C and Lupas AN (2002) Molecular evolution of proteasomes. *Current Topics in Microbiology and Immunology* **268**: 1–22.
- Yousef GM and Diamandis EP (2001) The new human tissue kallikrein gene family: structure, function, and association to disease. *Endocrine Reviews* **22**: 184–204.

Further Reading

- Gerlt JA and Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Review of Biochemistry* **70**: 209–246.
- Kageyama T (2002) Pepsinogens, progastricsins, and prochymosins: structure, function, evolution, and development. *Cellular and Molecular Life Sciences* **59**: 288–306.

Keywords

protease, evolution, gene duplication, domains, structure

Web links

- European Bioinformatics Institute InterPro database. A resource for whole-genome analysis
<http://www.ebi.ac.uk/interpro/>
- MEROPS Protease Database. A catalog and structure-based classification of proteases, with additional information about them
<http://merops.sanger.ac.uk>

Comments

- C1 Author: if willing, please supply your email address.
- C2 Author: OK as edited?
- C3 Author: Correct as edited?
- C4 Author: Please complete the headings.
- C5 Author: Correct?

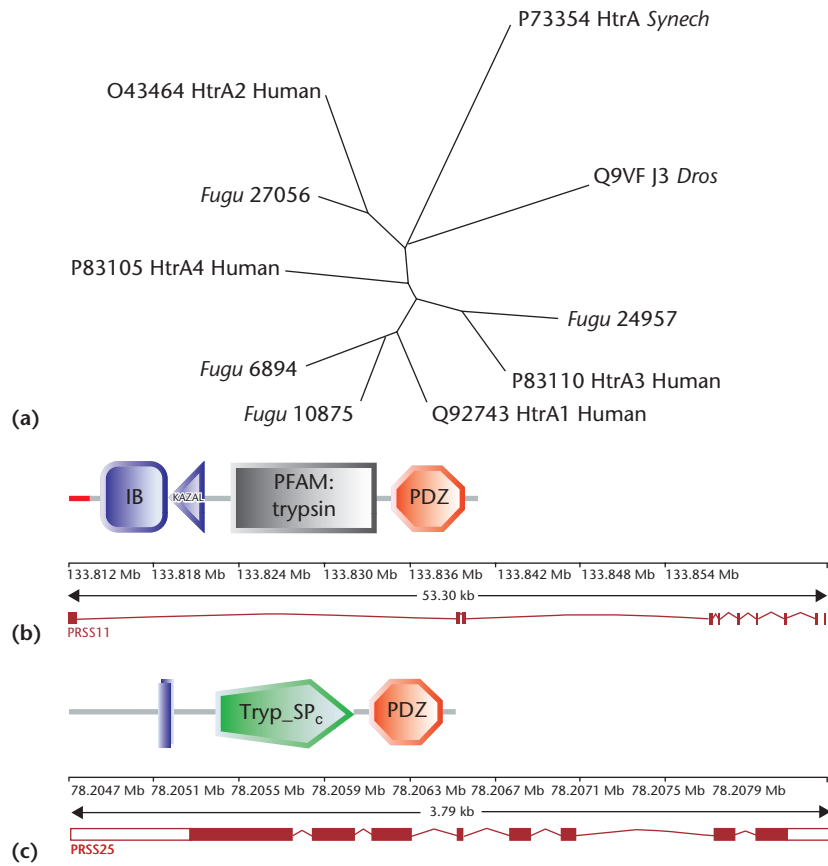


Figure 1 (a) Phylogram of HtrA proteases. (b) Human HtrA1 Q92743 domain and gene structure. (c) Human HtrA2 O43464 domain and gene structure. kb: kilobases; Mb: megabases.