Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Biochimie 166 (2019) 4–18

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

Origins of peptidases

Neil D. Rawlings*, Alex Bateman

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

ARTICLE INFO

Article history:
Received 7 March 2019
Accepted 29 July 2019
Available online 1 August 2019

Keywords:
Proteolytic enzyme
Peptidase
Evolution
Last universal common ancestor
Horizontal gene transfer
Asparagine peptide lyase

ABSTRACT

The distribution of all peptidase homologues across all phyla of organisms was analysed to determine within which kingdom each of the 271 families originated. No family was found to be ubiquitous and even peptidases thought to be essential for life, such as signal peptidase and methionyl aminopeptides are missing from some clades. There are 33 peptidase families common to archaea, bacteria and eukaryotes and are assumed to have originated in the last universal common ancestor (LUCA). These include peptidases with different catalytic types, exo- and endopeptidases, peptidases with different tertiary structures and peptidases from different families but with similar structures. This implies that the different catalytic types and structures pre-date LUCA. Other families have had their origins in the ancestors of viruses, archaea, bacteria, fungi, plants and animals, and a number of families have had their origins in the ancestors of particular phyla. The evolution of peptidases is compared to recent hypotheses about the evolution of organisms.

1. Introduction

Although all proteolytic enzymes perform the same basic function, namely the cleavage of a carbon-nitrogen bond between two amino acids in a peptide or protein, they are otherwise remarkably diverse. There are two very different catalytic mechanisms. Most proteolytic enzymes are peptidases that activate a water molecule which results in hydrolysis of the peptide bond (classified as hydrolases in Enzyme Nomenclature, subclass EC 3.4) [1]. However, there are a small number of self-processing peptidases that are not hydrolases but which bring about peptide bond cleavage by the rearrangement (cyclization) of an asparagine residue to a succinimide [2]; these are lyases (EC subclass 4.3). Amongst the peptidases, there are at least six different nucleophiles that bring about hydrolysis. In three of these activated water is bound to either aspartic acid residues (aspartic peptidases), glutamic acid residues (glutamic peptidases) or one or two metal ions (metallopeptidases). In the other three, the nucleophile is the hydroxyl of a serine (serine peptidases) or threonine (threonine peptidases) residue, or the thiol of a cysteine residue (cysteine peptidases). Although most frequently thought to be a triad [3], the catalytic mechanism can be dependent upon only two residues, for example endopeptidase La (family S16) [4], or in the case of metallopeptidases, may include many more. Metallopeptidases can rely upon one or two metal ions, in which case the metal is bound to the protein via three (one metal) or five (cocatalytic metals) residues [5]. The metal, although most frequently zinc, can also be cobalt, manganese, iron or copper. Then there are differences in activity. Many peptidases are endopeptidases, but there are also aminopeptidases, carboxypeptidases, dipeptidases, peptidyl-dipeptidases, dipeptidyl-peptidases and tripeptidyl-peptidases. Endopeptidases can also be restricted in the length of a peptide that can be cleaved (oligopeptidases) or many cleave peptide bonds that form between amino acid sidechains (isopeptidases). In a very few cases, peptide cleavage is associated with the transfer of another moiety to one of the free termini generated by the cleavage (transpeptidases) [6]. The number of different families of proteolytic enzymes, assembled from statistically significant similarities in protein sequence, is large. Even if the similarity in tertiary structures of proteolytic enzymes are taken into account, which reveals homologies not visible when only sequences are examined, there is still considerable variety. Usually, all the members of a single family utilize the same catalytic dyad or triad, or the same metal ligands for a metallopeptidase, but it does not follow that all members of a family are endopeptidases or exopeptidases, or that all members of a metallopeptidase family use the same divalent metal ions [7].

It can be assumed that some of this diversity results from the

* Corresponding author.
E-mail address: ndr@ebi.ac.uk (N.D. Rawlings).

https://doi.org/10.1016/j.biochi.2019.07.026

different environmental conditions under which the proteolytic enzymes have to function. For example, differences in pH in different cellular or extracellular compartments, and differences in extracellular temperature. Both pH and temperature greatly affect protein stability and particular structures would be required for proteolytic enzymes to be stable and functional at the extremes of these.

The MEROPS website (http://www.ebi.ac.uk/merops) and database lists 271 different families of proteolytic enzymes: 261 families of peptidases and ten of asparagine peptide lyases. These can be assembled into 56 clans (also known as superfamilies) based on similarities in tertiary structures, each representing a unique origin for the tertiary fold [8].

The purposes of this paper is to address where each family arose in the kingdoms and phyla of living organisms and to identify those families of peptidases that are thought to have originated in the last universal common ancestor (LUCA) of cellular organisms. Recent discoveries from the genome sequencing of environmental samples has led to new proposals about the origins of the major domains of life, and it is of interest to see if the origins of peptidase families provided support for these new ideas.

2. Methods

2.1. Detection and classification of homologues

A family of peptidases was assembled initially using the protein sequence of a well-characterized peptidase, for example bovine chymotrypsin, known as the “type example”. To find homologues of a type example, sequence searches were conducted against either the UniProt knowledgebase [9] or the non-redundant protein sequence library at NCBI [10] using BlastP [11]. Within a family, other well characterized proteolytic enzymes with different substrate preferences were identified manually and each of these was designated a “holotype” for a particular set of substrate preferences. Any of these holotypes could be used for subsequent BlastP searches to identify more homologues in the family. Additionally, the sequence of the type example could be used to search for homologues in the UniProt knowledgebase using the HMMER3 website [12]. BlastP searches with holotype sequences other than the type example were necessary for any family in which there are more than 20,000 known homologues, because 20,000 is the maximum number of hits returned by an automated, remote BlastP search. In all of these searches, there had to be overlap within a region defined as the “peptidase unit”, the (usually) single domain that carries the active site residues and/or the metal ligands, and the sequence relationship must have been at an E-value of 0.01 or less. If the active site residues (and/or metal ligands for metallopeptidase) were not all conserved, then the sequence was considered a “non-peptidase homologue”. The results presented in this paper were derived from numerous searches conducted regularly since 1992, the last of which were BlastP searches at NCBI using type example sequences in April 2018.

For certain completely sequenced genomes, a different method of analysis was used. This was a version of the MEROPS batch Blast [13]. A complete proteome was downloaded from the NCBI FTP site and submitted to a BlastP search against the merops_scan.lib sequence library, which contains the protein sequence of every holotype defined in MEROPS, which is approximately 5,000 sequences. Hits that were returned with an E-value of e^{-10} or lower were considered to be peptidase homologues. This method was used to analyse the recently determined proteome of *Thecamonas trahens*.

Peptidase families were assembled into a clan when the structural coordinates for a protein from each family were available at the Protein Data Bank [14] and these were compared using the DALI website [15]. A root mean square value of 6 standard deviation units or more was used to indicate that the structures were homologues. In the absence of a structure for a family, that family could be included in a clan if the active site residues (and/or metal ligands for a metallopeptidase) had been identified and were the same and in the same order in the sequence as those in an existing clan.

2.2. False positives and contaminants

One problem with any bioinformatics study is the unfortunate detection of false positives. Matching all active site residues (and metal ligands) considerably reduces the probability that a sequence is a false positive, and, for the purposes of this paper, only a non-peptidase homologue was considered a false positive. For each peptidase family, the number of species or each kingdom was counted and calculated as a percentage of the total number of species with homologues in that peptidase family. Where the number of species in a kingdom was less than 5% of the total species in a peptidase family and all the sequences included were non-peptidase homologues these were considered to be false positives.

If should be noted that many peptidase families are known to have non-peptidase homologues that no longer function as a peptidase and play a different role, so the absence or replacement of an active site residue (or metal ligand) is not in itself proof that the sequence is a false positive.

Identification of a contaminant is much more difficult to determine, because it can be difficult to distinguish a contaminant from a recent horizontal gene transfer, especially in microorganisms. A contaminant can be an active peptidase or a non-peptidase homologue. Most contaminants are derived from genome sequencing projects where shotgun sequencing has been performed but the sequences have not been assembled into chromosomes. However, misidentification of a single gene can occur when a single peptidase is being studied. The most obvious contaminant is apparently present in a proteome from one species but not in that species’ closest relatives. Frequently the same proteome contains more than one contaminant. In eukaryotes, contaminants are likely to be derived from endosymbionts.

Contaminants are likely to be rare, so we have examined all incidents where less than 1% of species in a kingdom has homologues from a particular peptidase family. Where the homologues are present in only one species from a single phylum, these are, for the purposes of this paper, considered contaminants (though if only one species in a phylum has had its genome completely sequenced, then some might be *bone fide* horizontal gene transfers). In addition, where an organism, usually a eukaryote, possesses such unusual homologues from many peptidase families, these are also identified as contaminants. Eukaryotes with homologues from ten or more peptidase families otherwise only found mainly in prokaryotes (99% of species) are the arthropod *Ceratitis capitata* (Mediterranean fruit fly; 18 families), the nematode *Litomosoides sigmodontis* (11), the alga *Micromonas pusilla* (11); and plants *Populus trichocarpa* (black cottonwood; 14), *Cucumis sativus* (cucumber; 13) and *Ricinus communis* (castor bean; 20).

A homologue that is unusual in a phylum or kingdom but not identified as either a false positive or a contaminant is assumed to have arisen as a result of a horizontal gene transfer.

3. Results

From comparison of protein sequences, peptidases can be clustered in 261 different families, comprising 16 families of aspartic peptidases, two families of glutamic peptidases, 96 families

of cysteine peptidases, 76 families of metallopeptidases, 53 families of serine peptidases, six families of threonine peptidases, two families of mixed catalytic type and 10 families of unknown catalytic type. In addition, there are 10 families of asparagine lyases. Family names are taken from the MEROPS website and database [8], but when first mentioned in the text below, the name of the type example peptidase from that family is shown in parenthesis.

Supplementary Table 1 shows an alphabetical list of families of proteolytic enzymes and the number of phyla in each organism kingdom with homologues from that family. For simplicity, we have adopted a five kingdom system for eukaryotes (animals, plants, fungi, chromists and protozoa). However, because presence in a phylum may be the result of horizontal gene transfer, and absence the result of gene loss, it cannot be assumed that the peptidase gene was present or absent in the ancestor of that phylum. On the assumption that gene loss is as likely as horizontal gene transfer, a gene is assumed to have been present in the ancestor of the kingdom only if it is present in half or more of the phyla in that kingdom. Kingdoms in which half or more of the phyla contain a homologue from a peptidase family are highlighted in yellow in Supplementary Table 1. Because the number of candidate phyla for bacteria is so high (77 out of 120 phyla), and there is doubt about the completeness of any genome from a candidate phylum, the criteria for presence in the ancestral bacterium is presence in half or more of non-candidate phyla (more than 22 phyla).

3.1. False positives and contaminants

For presumed false positives, the cell representing the kingdom in Supplementary Table 1 is highlighted in red (with the number of phyla in white text). It should be noted that the method for identifying false positives described above (section 2.2) cannot be applied to the families for which active site residues are not known. In these families all homologues are considered to be peptidases. The families are listed in Table 1.

The families with most non-peptidase homologues are shown in Table 2. In many of these families enzymes other than peptidases are known.

Table 1
Peptidase families where active site residues are not known or are uncertain.

Family	Type example
A5	thermopsin
A37	sso1175 protein
C49	strawberry mottle virus peptidase
C84	prtH peptidase
C95	lysosomal 66.3 kDa protein
C102	GtgE peptidase
M73	camelysin
M75	imelysin
M76	Atp23 peptidase
M77	tryptophanyl aminopeptidase 7-DMATS-type peptidase
M79	RCE1 peptidase
M82	PrsW peptidase
M98	YghJ protein
S46	dipeptidyl-peptidase 7
S62	influenza A PA peptidase
S68	PIDD auto-processing protein unit 1
T8	HopB1 protein
U32	*Porphyromonas* collagenase
U40	protein P5 murein endopeptidase
U49	Lit peptidase
U56	homomultimeric peptidase
U57	yabG protein
U69	AIDA-I self-cleaving autotransporter protein
U72	Dop isopeptidase
U74	neprosin

It should also be noted that some sequences may be identified as non-peptidase homologues in error. A large insert in a sequence can lead to a sequence misalignment, and consequently misidentification of active site residues. Sequencing errors can lead to erroneous frame-shifts, and gene assembly errors can result in exons being omitted or introns being incorporated into the coding sequence, all of which can give rise to apparent inserts or deletions in the protein sequence and apparent absence or replacement of active site residues.

For presumed contaminants, the cell representing the kingdom in Supplementary Table 1 is highlighted in brown (with the number of phyla in white text).

A peptidase family in which homologues are not identified as either false positives or contaminants, but which occurs in a minority of species in a phylum or phyla within a kingdom, is presumed to have arisen through a horizontal gene transfer. The direction of the transfer is usually assumed to have been from an organism in a kingdom which has the widest distribution amongst its phyla to an organism in a phylum or kingdom where the distribution of homologues is more restricted. However, it must be acknowledged that there may be cases where a transfer in the other direction would be equally or more likely. For example, an ancient gene transfer may have been derived from an organism in a phylum or kingdom that has subsequently undergone mass extinction. In a challenging environment with little biodiversity, such as a thermal spring or the deep sea, then horizontal gene transfers will only occur between organisms that thrive there, and it may not be possible to identify the direction of the transfer.

3.2. Unclassified organisms

The MEROPS database currently includes 1,150,089 peptidase sequences from 31,095 different organisms. Peptidase homologues are known from 28,375 species of cellular organisms (plus 2,720 viruses) with representatives from 184 different phyla. In addition, peptidase homologues are known from over 400 organisms for which the classification is not known. These are mostly prokaryotes (45 are archaea and 355 are bacteria). However, because the classification of single-celled eukaryotes is in a state of flux, there are also seven eukaryotes that are not assigned to a phylum. The majority of these organisms are derived from metagenomics studies and environmental samples, where it has been possible to assemble 135 genomes even though the organisms from which these genomes have been derived have never been seen, isolated or cultured.

3.3. Species with completely sequenced genomes

Supplementary Table 2 shows the phyla from which organisms with peptidase homologues have been sequenced. Many of these phyla are described as “candidate”, derived mostly from metagenomics projects and environmental samples in which genomes have been assembled for organisms that have not been observed or otherwise studied. The completeness of such genomes must be questioned, because although it is possible to assemble a single chromosome, this would not represent the entirety of genetic material for most organisms. Eukaryotes and some bacteria contain more than one chromosome, and many organisms have organelles or plasmids that have their own genomes; and it is challenging to combine any of these to obtain the full genome of any organism that has not been isolated and studied. Supplementary Table 2 also shows the number of species per phylum where the genome has been completely sequenced, and the average number of peptidase homologues per genome. It might be assumed that the more complex the organism, the more peptidase genes its genome will

Table 2
Peptidase families with most non-peptidase homologues.

Family	Type example	Non-peptidase homologues	Percentage of all sequences
C19	ubiquitin-specific peptidase 14	5,745	23.4%
C56	Pfpl peptidase	10,596	76.5%
M16	pitrilysin	7,603	24.5%
M20	glutamate carboxypeptidase	5,894	12.5%
M38	isoaspartyl dipeptidase	17,801	77.0%
S1	chymotrypsin A	8,158	11.0%
S8	subtilisin	8,860	23.0%
S9	prolyl oligopeptidase	6,882	11.2%
S33	prolyl aminopeptidase	7,253	21.8%

encode, but as can be seen from the table, a sponge (Porifera) or an acorn worm (Hemichordata) has more peptidase homologues than a chordate. The numbers of proteins (and therefore peptidases) in a genome will depend upon the automated methods used to identify genes and the mRNAs they encode, and also upon the genome assembly. Chordate chromosomes can be rich in regions of tandem copies of genes; such regions can be difficult to assemble correctly, and individual genes difficult to identify, leading to an underestimation of the number. Some predicted proteomes are apparently rich in differently transcribed isoforms, even though there may be no evidence for their existence, and because the analyses performed here is comparison of protein sequences, isoforms will be seen as unique proteins. This inflates the number of protein genes in a genome. For model organisms, such as human, these peptidase homologues are merged following manual curation so that the number of homologues represents the number of genes, but this is time consuming and so far has not been automated. For organisms where individual strains that had their genomes sequenced, the number of protein-coding genes may again be inflated because of the presence of different plasmids. The organism that apparently has the most peptidase homologues is *Bacillus cereus* with 4,569, but many strains have been sequenced. This is also true of *Escherichia coli* (3,005 peptidase homologues), *Pseudomonas fluorescens* (1,979), and *Streptococcus pneumoniae* (1,726). Polyploidy also affects the number of protein coding genes, particularly in cultivated plants such as rice (2,292), grape (1,746), and soya (1,653). All of these organisms apparently have more peptidase homologues than human (1,599).

The phyla in which organisms have fewest peptidase genes are Nanoarchaeota (15 peptidase homologues) and Tenericutes (18). It can also be seen from Supplementary Table 2 that not all phyla have an example with a completely sequenced genome. For example, none have been completely sequenced from a moss animal (Bryozoa) or a comb jelly (Ctenophora). There are also phyla for which we know of no peptidase homologues, for example the phylum Chaetognatha, which contains the arrow worms.

Supplementary Table 2 also shows the number of peptidase families per phylum. The phylum of organisms with peptidases from most families is Proteobacteria, with examples known from 174 families of peptidases. There are ten other phyla with peptidases from over a hundred peptidase families. These are the bacterial phyla Firmicutes (146 families), Bacteroidetes (143), Actinobacteria (129) and Cyanobacteria (104); the animal phyla Arthropoda (127), Chordata (125) and Nematoda (114); the plant phylum Streptophyta (120); the fungal phylum Ascomycota (106); and the Archaean phylum Euryarchaeota (108). Unsurprisingly, these eleven phyla include some of the most intensively studied organisms.

At the opposite end of the scale, some phyla contain peptidases from only a single family. In the case of the animal phylum Bryozoa, there are no completely sequenced genomes and the only known peptidase homologue is a cysteine peptidase from family C64 (Cezanne peptidase) from *Fredericella sultana*. Only the cyanelle (a primitive plastid-like organelle) genome of the Glaucophyte *Cyanophora paradoxa* has been completely sequenced and contains a Clp-like serine peptidase from family S14. Similarly, no genome has been completely sequenced for any single-celled organism from the phylum Rhombozoa, and the only known peptidase homologue is a cathepsin L-like fragment (peptidase family C1) from *Dicyema japonicum*. There are also several candidate phyla proposed for genomes assembled from metagenomics studies and environmental samples that also apparently contain peptidases from only one family.

3.4. Peptidases families from viruses

Of the 271 families of proteolytic enzymes, 71 are predominantly found in viruses (and the hosts they infect). The mutation rate in viruses is very high, because viruses do not have error-correcting mechanisms [16], which means that detecting distant sequence relationships is difficult. Hence there are many families of peptidases from viruses, and many families are known to share structural relationships indicating that many are related to one another even if this relationship is not detectable from comparison of the peptidase sequences alone. In most cases, a family of viral peptidases frequently contains only one family of viruses. Peptidase family S39 is an exception and contains sequences from four families of viruses. Viruses are not classified into taxa higher than order, but are known to have genomes that are single-stranded RNA (ssRNA), double-stranded RNA (dsRNA) or double-stranded DNA (dsDNA). A virus with ssRNA encodes at least one polyprotein, which may contain peptidases within it required to cleave the polyprotein into individual proteins. A virus with dsDNA encodes several genes for individual proteins, one or more of which may be a peptidase required to activate one or more of the proteins that is synthesized as a precursor.

There are 59 families with homologues mainly from viruses. Families with homologues from viruses with a dsDNA genome are C57 (vaccinia virus I7L processing peptidase), C63 (African swine fever virus processing peptidase), C71 (pseudomurein endoisopeptidase Pei), C76 (viral tegument protein deubiquitinylating peptidase), C120 (mavirus processing peptidase), N7 (reovirus type 1 coat protein) and S80 (prohead peptidase gp175). Omitting possible false positives, contaminants and host species, the following families can also be added to this list: A3 (cauliflower mosaic virus-type peptidase), C5 (adenain), C104 (PlyC phage lysin), M44 (pox virus metallopeptidase: the only family of metallopeptidases found exclusively in viruses), S21 (cytomegalovirus assemblin) and S77 (prohead peptidase gp21). Not all of these families contain protein-processing enzymes. For example, pseudomurein endoisopeptidase Pei from C71 degrades the cross-links within the cell wall glycoprotein of infected archaea, allowing the

release of assembled virions [17] and family C76 contains the only known viral deubiquitinating enzyme, though the role this plays in infection is unknown [18].

Families with homologues from viruses with a dsRNA genome are: C7 (chestnut blight fungus virus p29 peptidase), C8 (chestnut blight fungus virus p48 peptidase), C21 (tymovirus peptidase), N5 (picobirnavirus self-cleaving protein), S69 (Tellina virus 1 VP4 peptidase) and U40 (protein P5 murein endopeptidase). Omitting possible false positives and contaminants, family S50 (infectious pancreatic necrosis birnavirus Vp4 peptidase) can also be added to this list.

The 28 families with homologues from viruses with a ssRNA genome are shown in Table 3. Omitting possible false positives, contaminants and host species, the following families can also be added to this list: A9 (spumapepsin), C3 (picornain 3C), C4 (nuclear-inclusion-a peptidase), C6 (potato virus Y-type helper component peptidase), C9 (sindbis virus-type nsP2 peptidase), C18 (hepatitis C virus peptidase 2), C53 (pestivirus Npro peptidase), C87 (nairovirus deubiquitinylating peptidase), N8 (poliovirus capsid VP0-type self-cleaving protein), S7 (flavivirin) and S62 (influenza A PA peptidase).

Viruses with dsDNA and dsRNA are mainly bacteriophages that infect bacteria (though some infect eukaryotes), whereas viruses with ssRNA are mainly pathogens of eukaryotes. It is noticeable that many more families of cysteine peptidases occur in ssRNA viruses than dsDNA viruses. Structural studies have shown that the picornains from family C3 have a similar Greek key fold to members of the chymotrypsin family (S1), and both families are included in the same clan (PA) in the MEROPS classification [19]. The exchange of an active site residue is not unknown, but the replacement of one nucleophilic residue for another is unusual (in this case a cysteine replaces a serine). It is possible that the eukaryote nucleus is a slightly acidic environment (it does contain nucleic acids), and cysteine peptidases are more stable and active at acidic pH than serine peptidases. An isoform of cathepsin L has been found in the nucleus [20], whereas other isoforms are found in the lysosome, an acidic organelle. However, the pH of the nucleus has been measured in plant cells (pH 7.2) and shown to be similar to that of the cytoplasm (pH 7.3) [21].

There are families of peptidases, though not predominantly only from viruses, that contain homologues from dsDNA and ssRNA viruses. Family A2 (retropepsin) contains homologues from eleven species of dsDNA viruses where the whole of the *pol* polyprotein gene appears to have been acquired from ssRNA viruses by horizontal gene transfer. Family S1 contains 78 homologues from dsDNA viruses (mostly from Phycodnaviridae although none have been characterized) and 61 ssRNA viruses, so the direction of horizontal gene transfer is debatable. In families S14 (peptidase Clp), S21 (cytomegalovirus assemblin) and S24 (repressor LexA), homologues are mostly from dsDNA viruses but each family contains a single homologue from an ssRNA virus, again presumably the result of horizontal gene transfers. The homologue from S21 is allegedly from rabies virus, but the sequence, originally in the PIR database, is not in UniProt.

3.5. Peptidases families from archaea

The peptidase family with most homologues from archaea is T1 (proteasome; 1,484); this is also the family with homologues from most species of archaea (1,484) and the greatest percentage of archaean species (73.4%). No other family has homologues from more than half the archaean species.

3.6. Peptidases families from bacteria

Because more genomes of species have been sequenced from bacteria than any other kingdom of organisms, there are many more homologues in peptidase families than for other kingdoms. The peptidase family with most homologues from bacteria is M20 (40,159 sequences). This is also the family with homologues from the most bacterial species (10,173) and the greatest percentage of bacterial species with peptidase homologues (51.8%). Other families with more than 10,000 homologues are shown in Table 4.

Table 3
Families with homologues only from viruses with a ssRNA genome.

Family	Type example
C16	murine hepatitis coronavirus papain-like peptidase 1
C23	carlavirus peptidase
C24	rabbit hemorrhagic disease virus 3C-like peptidase
C27	rubella virus peptidase
C28	foot-and-mouth disease virus L-peptidase
C30	porcine transmissible gastroenteritis virus-type main peptidase
C31	porcine reproductive and respiratory syndrome arterivirus-type cysteine peptidase alpha
C32	equine arteritis virus-type cysteine peptidase
C33	equine arteritis virus Nsp2-type cysteine peptidase
C36	beet necrotic yellow vein furovirus-type papain-like peptidase
C37	calicivirin
C42	beet yellows virus-type papain-like peptidase
C49	strawberry mottle virus peptidase
C62	gill-associated virus 3C-like peptidase
C74	pestivirus NS2 peptidase
C99	iflavirus processing peptidase
C105	papain-like peptidase 1 alpha
C107	alphamesonivirus 3C-like peptidase
N1	nodavirus peptide lyase
N2	tetravirus coat protein
S3	togavirin
S29	hepacivirin
S30	potyvirus P1 peptidase
S31	pestivirus NS3 polyprotein peptidase
S32	equine arteritis virus serine peptidase
S39	sobemovirus peptidase
S65	picornain-like serine peptidase
S75	white bream virus serine peptidase

Table 4
Peptidase families with more than 10,000 sequence homologues from bacteria.

Family	Type example	Sequences
C26	gamma-glutamyl hydrolase	14,901
C40	dipeptidyl-peptidase VI	18,171
C56	PfpI peptidase	10,673
M3	thimet oligopeptidase	13,098
M15	D-Ala-D-Ala metallocarboxypeptidase	10,568
M16	pitrilysin	19,725
M20	glutamate carboxypeptidase	40,159
M23	beta-lytic metallopeptidase	10,458
M24	methionyl aminopeptidase 1	17,942
M28	aminopeptidase S	13,397
M38	isoaspartyl dipeptidase	17,310
S1	chymotrypsin	25,814
S8	subtilisin	20,534
S9	prolyl oligopeptidase	33,824
S11	D-Ala-D-Ala carboxypeptidase A	15,745
S12	D-Ala-D-Ala carboxypeptidase B	25,355
S14	peptidase Clp	14,057
S16	endopeptidase La	18,997
S24	LexA repressor	14,209
S26	signal peptidase 1	14,976
S33	prolyl aminopeptidase	19,490
S41	C-terminal processing peptidase-1	12,557
S49	signal peptide peptidase A	14,374

3.7. Peptidase families from eukaryotes

3.7.1. Peptidases families from protozoa

The peptidase family with most sequences from protozoa is C1 (1,744 sequences); this is also the family with homologues from most species of protozoa (154) and from the greatest percentage of protozoan species (68.8%). No other family contains more than 50% of protozoan species with peptidases: families M16 and T1 have homologues from 49.6%.

3.7.2. Peptidase families from fungi

The peptidase family with most sequences from fungi is A1 (6,461 sequences); followed by S8 (4,664), T1 (4,632), C19 (4,360). The family with homologues from most species of fungi is S8 (717); this is also the family with the highest percentage of fungal species (75.4%). Other peptidase families with homologues from more than half of fungal species are listed in Table 5. No other kingdom of organisms has as many families with representatives with more than half of the species with peptidase homologues.

3.7.3. Peptidases families from chromists

The peptidase family with most sequences from chromists is S1 (733 sequences). The family with homologues from most species of chromists is M41 (141); this is also the family with the greatest percentage of chromist species (54.5%).

3.7.4. Peptidases families from plants

The peptidase family with most sequences from plants is A1 (6,498 sequences); followed by S8 (5,370), C1 (5,362), S10 (4,220) and S9 (4,007). The family with homologues from most species of plants is S14 (1,120), which is also the family with homologues from most plant species (60.1%); no other family has homologues from more than half the plant species.

3.7.5. Peptidases families from animals

The peptidase family with most sequences from animals is S1 (43,933 sequences); other families with more than 5,000 sequences from animals are shown in Table 6. The peptidase family with homologues from most species of animals is S41 (1,712), followed by C26 (1,706), S1 (1,182). Family S41 is also the one with the greatest percentage of animal species (30.1%).

4. Discussion

Fig. 1 shows the proposed origins of all peptidase families. Peptidase families with an origin in the same clade (which includes kingdom or phylum) are clustered together.

4.1. Peptidase families present in all phyla

One of the questions that we hope to answer in this paper is: what peptidases were present in the last universal common ancestor (LUCA)? This surely represents the minimum number of peptidase families required for an independent, functional organism (parasitic organisms may have a reduced peptidase spectrum because they are able to utilize host peptidases, leading to loss of peptidase genes in the parasite because they have become redundant).

Given that the phylum with most examples of peptidase families is the Proteobacteria with only 66% (174 of the 271 peptidase families so far known), there is no phylum which has representatives of all peptidase families. This means that (1) some peptidase families have arisen in organisms since LUCA; and (2) some organism clades have lost genes for entire peptidase families. In addition, some families of proteolytic enzymes will have been acquired in a phylum by horizontal (or lateral) gene transfer, in which the gene is transferred from a plasmid or an organelle to a chromosome; plasmids are shared between distantly related prokaryotes; or an intermediary, such as a virus, transfers a gene from one organism to another accidently [22].

Table 5
Peptidase families with homologues known from most species of fungi.

Family	Type example	Percentage of species
A1	pepsin	65.5%
C12	ubiquitinyl hydrolase-L1	61.7%
C13	legumain	61.8%
M1	membrane alanyl aminopeptidase	63.4%
M3	thimet oligopeptidase	62.3%
M14	carboxypeptidase A	54.7%
M16	pitrilysin	66.3%
M18	aminopeptidase I	63.7%
M19	membrane dipeptidase	51.6%
M28	aminopeptidase S	60.8%
S8	subtilisin	75.4%
S9	prolyl oligopeptidase	65.2%
S26	signal peptidase 1	62.3%
T1	Proteasome	65.6%

Table 6
Peptidase families with more than 5,000 sequence homologues from animals.

Family	Type example	Sequences
A1	pepsin	6,407
C1	papain	9,169
C14	caspase	7,377
C19	ubiquitin-specific peptidase 14	15,034
M1	membrane alanyl aminopeptidase	7,740
M10	matrix metallopeptidase-1	8,971
M12	astacin	25,555
M13	neprilysin	7,058
M14	carboxypeptidase A	12,095
M16	pitrilysin	5,652
S1	chymotrypsin A	43,933
S8	subtilisin	5,910
S9	prolyl oligopeptidase	18,550
S33	prolyl aminopeptidase	8,064
T1	proteasome	7,740

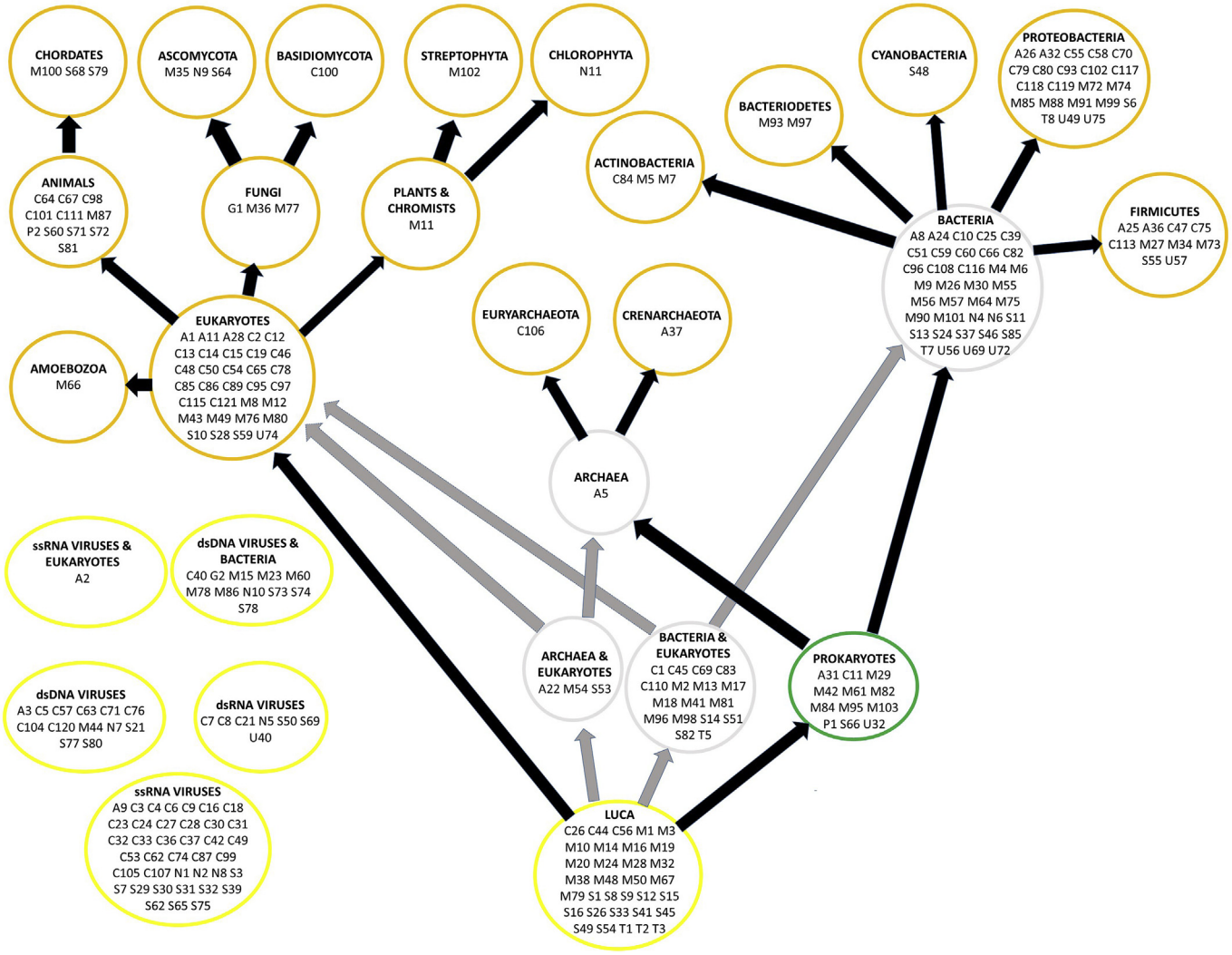

Fig. 1. Shows the predicted origins of families of proteolytic enzymes in the last universal common ancestor (LUCA), viruses, organism kingdoms and phyla. LUCA, a kingdom, a phylum, and viruses with a common nucleic acid type are shown as a circle within which are listed the families of proteolytic enzymes thought to have originated within that grouping. There are also circles that include enzyme families common to two or more kingdoms or other clades. Black arrows indicate evolutionary relationships between kingdoms and kingdoms and phyla, and grey arrows indicate alternative relationships.

Table 7
Peptidase families present in all seven organism kingdoms.

Family	Type example
C56	PfpI peptidase
M1	membrane alanyl aminopeptidase
M3	thimet oligopeptidase
M20	glutamate carboxypeptidase
M48	STE24
M67	isopeptidase Rpn11
S1	chymotrypsin
S8	subtilisin
S9	prolyl oligopeptidase
S12	D-Ala-D-Ala carboxypeptidase B
S16	endopeptidase La
S26	signal peptidase 1
S54	rhomboid
T1	proteasome

Peptidase families widely distributed amongst phyla in each kingdom are assumed to have been present in the last universal common ancestor (LUCA).

Families of proteolytic enzymes that were present in LUCA are most likely to be present in eukaryotes, archaea and bacteria. There are 14 families present in all seven kingdoms (see Table 7).

It is possible that the criteria outlined above are too stringent for some kingdoms with few phyla, and it is possible that the ten families listed in Table 8 were also present in LUCA where the number of phyla with representatives for some kingdoms is low.

Peptidases from these families are active mainly at neutral or basic pH. These families include endopeptidases, oligopeptidases, aminopeptidases, carboxypeptidases, and isopeptidases, which implies that these various categories existed prior to LUCA. Some of these families are in the same clan: C26 and C56 are both in clan PC; M1, M3 and M48 are in clan MA; M20 and M28 are in clan MH; S9 and S33 are in clan SC; T1, T2 and T3 are in clan PB. This implies that these different structures and the gene duplications that gave rise to different families within the same clan predate LUCA. LUCA also contained metallopeptidases with monocatalytic metal ions from families M1, M3, M14, M16, M48, M50 and M67, as well as enzymes with cocatalytic metal ions from families M20, M28 and M38.

Family M38 is unusual in that the majority of characterized proteins are not peptidases. The known peptidases are membrane

Table 8
Other peptidase families possibly present in LUCA. The following peptidase families are present in all kingdoms but not found within most phyla in some kingdoms. These may have been present in LUCA and subsequently lost in some phyla.

Family	Type example	Kingdoms with few represented phyla
C26	gamma-glutamyl carboxypeptidase	Chromista
C44	amidophosphoribosyltransferase precursor	Fungi, Chromista
M14	carboxypeptidase A	Archaea
M16	pitrilysin	Archaea
M28	aminopeptidase Y	Chromista
M38	beta-aspartyl dipeptidase	Fungi
M50	S2P endopeptidase	Fungi
M79	RCE1 peptidase	Fungi
S15	Xaa-Pro dipeptidyl peptidase	Animalia, Viridiplantae (absent from Streptophyta), Fungi, Chromista, Protozoa
S33	prolyl aminopeptidase	Chromista
T2	asparaginase	Protozoa
T3	gamma glutamyl transpeptidase	Protozoa, Archaea

dipeptidases such as beta-aspartyl dipeptidase from *Escherichia coli*, and appear to have developed from enzymes with different catalytic activities, such as urease and dihydroorotase. Urease and dihydroorotase are both divalent cation metallopeptidases, but beta-aspartyl dipeptidase is a monozinc metallopeptidase and its gene appears to have been the product of half a dihydroorotase gene [23]. The nature of dihydroorotase, which in some eukaryotes is derived from a fusion of genes encoding four different enzymatic components of the pyrimidine pathway (glutamine amidotransferase, carbamoyl phosphate synthase, aspartate carbamoyltransferase and dihydroorotase) [24], has been used as a criterion for classifying eukaryotes into Bikonta (plants and chromists) and Unikonta (animals, fungi and some single-celled organisms) [25]. The glutamine amidotransferase domain is classified separately in MEROPS as a homologue in peptidase family C26.

It might be expected that peptidases required for protein synthesis, such as signal peptidase 1, which is required for protein export, and methionyl aminopeptidase, which removes the initiating methionine from some cytoplasmic proteins, would be present in all organisms. However, surprisingly, signal peptidase is absent from most species of the classes Mollicutes (which includes *Mycoplasma*) and Pinopsida (which includes conifers). Homologues of family M24 (which includes methionyl aminopeptidase) are present in phyla from all kingdoms of life, but only in a fraction of chromists, and is absent from species of Cryptophyta. On balance, though, this might best be interpreted as a loss from Chromista, and that LUCA included both S26 and M24 homologues.

There are a number of other families that most probably had an origin in LUCA, but which are lacking in organisms from some kingdoms. Family M10 (matrix metallopeptidase-1) is not present in the majority of phyla of fungi, chromists, archaea or bacteria (most cyanobacteria, *Lactobacillus*, *Streptococcus*, *Yersinia* and *Pseudomonas* species have a homologue). This probably represents gene losses in these kingdoms and that the ancestral gene for the family was present in the genome of LUCA. Additionally, there are homologues from dsDNA viruses, which are presumably the result of horizontal gene transfers. Family M19 (membrane dipeptidase) present in animals, fungi, protozoa, archaea and most bacterial phyla, is absent from chromists and plants. Presumably, a homologue was also present in LUCA but subsequently lost from bikonts. Family M32 (carboxypeptidase Taq) is present in archaea, bacteria, chromists, plants and some protozoa, but is absent from animals and fungi. Presumably a homologue was also present in LUCA but subsequently lost from unikonts. Family S41 (C-terminal processing peptidase-1) is present in archaea, bacteria and some eukaryotes, but is notably absent from most protozoa and fungi. Homologues are present in the alveolates *Chromera velia* and *Vitrella brassicaformis* and the ascomycete *Cordyceps bassiana*. Family S45 (penicillin G acylase precursor) is found in archaea, bacteria and protozoa, but not animals, fungi or plants. The protozoan homologues are from all three major divisions of eukaryotes (the bikonts *Tetrahymena* and *Eimeria*, the excavate *Naegleria* and the unikont *Acanthamoeba*) suggesting either horizontal gene transfers from a prokaryote to protozoa or gene losses in many ancestors of major eukaryote divisions.

There is no family of aspartic or glutamic peptidases or asparagine lysases found in organisms from all kingdoms. This implies that there were no homologues of these in LUCA.

It has been suggested that LUCA was something less than a fully-functioning organism [26]. On the assumption that homologues in peptidase families present in LUCA performed similar functions to modern proteins, then it is clear that LUCA was able to synthesize intracellular (processed by methionyl aminopeptidase) and secreted proteins (processed by signal peptidase). LUCA was able to obtain the amino acid raw materials for protein synthesis by both intracellular (peptidases from families C56, M3, M16, M48, S1, S8, S9, S16, S54 and T1) and extracellular proteolysis (peptidases from M1 and M20). LUCA had membrane-associated proteins, including multipass proteins (such as signal peptidase and peptidases from M48 and S54), type II membrane proteins (family M1) and proteins attached to the membrane via C-terminal lipid groups (as shown by the presence of peptidases from M48 and M79). LUCA also had a cell wall, because homologues from S12 process the precursors of the cell-wall cross-linking peptide. The presence of an S41 homologue suggests that LUCA was able to degrade abnormal proteins [27]. Bacterial peptidases in S1 and M48 are stimulated by heat-shock and degrade abnormal proteins [28,29], suggesting that LUCA may have had a heat-shock response.

4.2. Peptidase families with an origin in viruses

Peptidase families found only in viruses are discussed above (section 3.4). The origin of viruses is much disputed. There are three current theories: (1) that viruses, like parasites, are degenerate derived from single-celled organisms that have lost many genes that are unrequired for a viral life style; (2) that viruses are derived from retrotransposons and plasmids; (3) that virus-like organisms containing nucleic acids and proteins represent the oldest forms of life [30]. The existence of giant viruses such as *Mimivirus*, pathogens of *Amoeba*, give some credence to the first hypothesis [31] and *Acanthamoeba polyphaga* mimivirus includes peptidases from families C1 (papain), C19 (ubiquitin-specific peptidase 14), C48 (Ulp1 peptidase), C89 (acid ceramidase precursor), M13 (neprilysin), M16 (pitrilysin), N10 (intein-containing replicative DNA helicase precursor) and S16 (Lon-A peptidase). Other mimiviruses and megaviruses also have homologues in C44

(amidophosphoribosyltransferase precursor), M43 (cytophagalysin), N11 (intein-containing chloroplast ATP-dependent peptide lyase), S9 (prolyl oligopeptidase) and S54 (rhomboid-1). The similarity of polyproteins from ssRNA viruses and retrotransposons from eukaryotes (including the polyprotein processing peptidase from family A2) and lantiobiotic ABC transporter proteins (family C39) from viruses (*e.g.* bacteriophage SPBc2) to plasmids (*e.g.* plasmid p03 from *Bacillus cereus*) give credence to the second hypothesis. However, it is equally possible that retrotransposons are derived from ssRNA viruses and that plasmids are derived from dsDNA viruses.

There are a number of peptidases families that contain homologues from dsDNA viruses as well as other organisms, and it not possible to tell if these families originated in a virus or a cellular organism. Families found in species from all cellular organism kingdoms, and thus with a potential origin in LUCA, as well as dsDNA viruses, include C1, C44, M10, M41 (FtsH peptidase; absent from Archaea), M67 and S49 (signal peptide peptidase A; absent from animals). There are eleven families found in viruses and bacteria (Table 9). In family G2, the bacterial homologues are mostly from Firmicutes, and in families S73 and S74, from Proteobacteria. Families C40, M23, M60 and N10 also have eukaryote and/or archaean homologues, presumably the result of horizontal gene transfers, but only the C40 peptidase LytFM from the house dust mite (*Dermatophagoides pteronyssinus*) has been characterized [32]. The only peptidase family that contain homologues from dsDNA virus and eukaryotes (in addition to those listed above for mimiviruses) is C19.

4.3. Peptidase families with an origin in prokaryotes

There are six families that are present in most archaean and bacterial phyla: A31 (hydrogenase-processing endopeptidase HybD), C11 (clostripain), M29 (aminopeptidase T), M42 (glutamyl aminopeptidase), M103 (TldD peptidase) and S66 (murein tetrapeptidase LD-carboxypeptidase). Family M55 (D-aminopeptidase DppA) is found in some archaea (particularly order Thermoproteales and class Thermococci), probably the result of ancient gene transfers, and in seventeen bacterial phyla, but most homologues are from the Firmicutes *Saccharomonospora*, *Streptomyces* and *Bacillus*, and Proteobacteria *Bordatella* and *Burkholderia*. Family M61 (glycyl aminopeptidase) occurs in three archaean phyla (and most bacterial phyla) but is only common to members of the crenarchaeote order Sulfolobales, perhaps implying an ancient horizontal gene transfer from a bacterium. Other peptidase families have narrower distributions and it is difficult to determine their origins. Family M82 (PrsW peptidase) is found in euryarchaeotes and species from several bacterial phyla. Family M84 (MpriBi peptidase) is found in three species from the euryarchaeote class Methanomicrobia and the firmicute *Bacillus*. Although most M95 (selecase) homologues are from Proteobacteria species, these are not from the majority of species, and homologues are also known from other bacterial and archaean phyla.

Family P1 (DmpA aminopeptidase) is found in species from several archaean and bacterial phyla, but also in ascomycete fungi and a few protozoa, presumably the result of horizontal gene transfers from an ancient prokaryote to eukaryotes. Family U32 (*Porphyromonas* collagenase) is found in archaea, bacteria and chlorophytes such as *Micromonas*, *Ostreococcus* and *Volvox*. Presence in chlorophytes is presumably the result of horizontal gene transfers.

Table 9
Peptidase families found in viruses and bacteria.

Family	Type example
C40	dipeptidyl-peptidase VI
G2	pre-neck appendage protein
M15	zinc D-Ala-D-Ala carboxypeptidase
M23	beta-lytic metallopeptidase
M60	enhancin
M78	ImmA peptidase
M86	PghP gamma-polyglutamate hydrolase
N10	intein-containing replicative DNA helicase precursor
S73	gpO peptidase
S74	*Escherichia coli* phage K1F endosialidase CIMCD self-cleaving protein
S78	prohead peptidase

4.4. Peptidase families with an origin in archaea

Only one family, family A5 (thermopsin), has homologues found in members of most archaean phyla and presumably had its origin in the ancestral archaean. Homologues are known only from organisms that are thermophiles and acidiophiles [33]. Peptidases from family A37 (sso1175 protein) are also active at high temperature and low pH, but homologues are found only in species from the phylum Crenarchaeota. Family C106 (archaeosortase A) homologues are found only in species from the phylum Euryarchaeota; archaeosortase processes proteins and links them covalently to the cell surface [34]; these peptidases are analogous to bacterial sortases (family C60).

4.5. Peptidase families with an origin in bacteria

Families C82 (L,D-transpeptidase) and C96 (McjB peptidase) are found only in bacteria and presumably has their origin in the ancestral bacterium. Omitting possible false positives and contaminants, the following families can also be added to this list: C66 (IdeS peptidase), M30 (hyicolysin), M90 (MtfA peptidase), N6 (YscU protein), U56 (homomultimeric peptidase) and U69. Families M9 (bacterial collagenase V), M56 (BlaR1 peptidase), M64 (IgA peptidase) and M74 (murein endopeptidase) are also present in many, but not most, bacterial phyla, which may indicate an origin in the ancestral bacterium and subsequent loss from the ancestors of some phyla.

The following families predominate in bacteria, but are also found in phyla of other organisms. M15 (D-Ala-D-Ala metallocarboxypeptidase), M26 (IgA-specific serine endopeptidase), S11 (D-Ala-D-Ala carboxypeptidase), S13 (D-Ala-D-Ala carboxypeptidase) and S24 (LexA repressor) were presumably present in the ancestral bacterium, but with horizontal gene transfers to protozoa (M15); archaea (M26); protozoa and plants (S11); archaea, protozoa, fungi and plants (S13); and viruses, archaea, plants and animals (S24). Those that are presumably the result of an ancient horizontal gene transfer from a bacterium to an archaean are: C10 (streptopain), A24 (type 4 prepilin peptidase), C25 (gingipain) and M4 (thermolysin). Presence in protozoa of homologues of A8 (signal peptidase 2) at are presumably the result of an ancient horizontal gene transfer from a bacterium to a protozoan. Family M6 (immune inhibitor A peptidase) is also predominantly bacterial but with horizontal gene transfers to archaea, chromists and fungi. The occurrence of M4 homologues in fungi is presumably the result of an ancient horizontal gene transfer from a bacterium to a fungus. Other families with a presumed origin in bacteria, but which are not found in the majority of bacterial phyla and for which a few homologues are known from other organisms include C39 (bacteriocin-processing peptidase; with 15 homologues from euryarchaeotes and 19 from eukaryotes); C51 (D-alanyl-glycyl peptidase; Protozoa and dsDNA viruses), C59 (penicillin V acylase precursor; Fungi, Chromista, Archaea, dsDNA viruses), C60 (sortase; archaea and an ssRNA virus). Homologues from family C108 (Prp

peptidase) are found in species from eight bacterial phyla, with most homologues (765 out of a total 833) from species of Firmicutes. Horizontal gene transfers to dsDNA viruses have also occurred, and the pneumococcal bacteriophage Cp-1 peptidase has been characterized as a maturation enzyme for the major head protein [35].

A number of these families contain peptidases required for the processing of the premurein precursor or the peptidoglycan crosslinks in the cell wall, including A8, C51, C82, M15, M74, S11 and S13. Other families include peptidases that are effective during infection, by degrading antibodies (C66, M26 and M64); the generation of inhibitory peptides such as microcins (C96), and toxins (C10, C25, M30, N6, U69 and some peptidases from C60 and M4); or the degradation of antibiotics (M6, M56). Peptidases from other families are required for processing of proteins: C39 for bacteriocins, C108 for ribosomal protein L27 [36]; M90 for regulatory proteins.

There are a number of other families which predominate in one phylum, but often with a few homologues in others, presumably the result of horizontal gene transfers via plasmids. Of course, any presumed horizontal gene transfer to just one species in a phylum might also be a result of contamination.

Families restricted to members of the phylum Proteobacteria are shown in Table 10. Omitting false positives and contaminants, M72 (peptidyl-Asp metallopeptidase) and M88 (IMPa peptidase) can be added to the list. Other families that predominates in Proteobacteria but with a few homologues from species in other phyla are: A26 (omptin; homologues in species of Fusobacteria and Spirochaetes), A32 (PerP peptidase; homologues from species in four other bacterial phyla), C58 (YopT peptidase; a single homologue from *Schlesneria paludicola*, Planctomycetes), C80 (RTX self-cleaving toxin; homologues in fish and Firmicutes), C93 (LapG peptidase; a single homologue from *Denitrovibrio acetiphilus*, Deferribacteres), S6 (IgA1-specific serine peptidase; some homologues in Fusobacteria and Verrucomicrobia), T8 (a few homologues in species of Chlamydiae), U49 (a homologue in *Leptospira terpstrae*, Spirochaetes), U75 (Ras/Rap1-specific peptidase; a few homologues in species of Actinobacteria). The majority of these families include peptidases required for pathology (A26, C55, C58, C70, C80, C102, C117, C118, M85, M88, M91, S6 and U75). Peptidases in M99 act on cell wall peptidoglycans. Peptidases in family A32 process factors that affect pili formation [37]. The Lit peptidase from U49 brings about bacterial cell death following bacteriophage infection [38].

Family M34 (anthrax lethal factor) is the only one found only in Firmicutes. Family M73 can be added to the list if possible contaminants are omitted, and there are other families that predominate in species of Firmicutes with a few outliers: A25 (gpr peptidase; with homologues in six other bacterial phyla and the archaean *Methanosarcina mazei*), A36 (sporulation factor SpoIIGA; with homologues in Actinobacteria, Bacteroidetes and Chlamydiae), C47 (staphopain A; with homologues in *Prevotella dentalis* from the phylum Bacteroidetes), C75 (AgrB peptidase; *Prevotella dentalis*), C113 (IgdE peptidase; a single homologue from a species of Actinobacteria and another in Tenericutes), S55 (SpoIVB peptidase; some homologues in Acidobacteria, Actinobacteria, Proteobacteria and Synergistetes), U57 (a single homologue from *Streptomyces avermitilis,* Actinobacteria). Family M27 (tentoxilysin) also predominates in members of the Firmicutes, but with a probably horizontal gene transfer to dsDNA viruses. Peptidases from families C47, C113 and M27 are active during pathology. Several peptidases from the families A25, A36, S55 and U57 are involved in sporulation; production of endospores occurs only in firmicute bacteria [39]. Peptidases from family C75 are involved in processing peptides required for quorum-sensing [40].

Family M7 (snapalysin) is found only in species of Actinobacteria; however family C84 predominates in Actinobacteria (but homologues are found in members of Bacteroidetes, where it is a virulence factor; Cyanobacteria and Proteobacteria), as does family M5 (mycolysin; with homologues in Chloroflexi and Firmicutes). Family M93 (BACCAC_01431 protein) is found only in species of Bacteroidetes, however family M97 (EcxAB peptidase, a virulence toxin) predominates in Bacteroidetes, but homologues are also found in species from Balneolaeota, Gemmatimonadetes and Proteobacteria. Family S48 (HetR peptidase) predominates in Cyanobacteria, with a single homologue in *Edwardsiella ictaluri* (Proteobacteria); HetR is required for formation of heterocyst cells, which are unique to cyanobacteria, during nitrogen starvation [41].

There are a number of families which are found in several bacterial phyla, but for which it is not possible to determine the ancestor. Homologues of family C116 (dermonecrotic toxin) are found in only two species, *Flavobacterium columnare* (Bacteroidetes) and *Pasteurella multocida* (Proteobacteria). Homologues of family M57 (prtB protein) are found in mainly in species of Bacteriodetes and order Myxococcales (phylum Proteobacteria); those from family M75 mostly in Proteobacteria and the spirochaete *Leptospira*; family M101 (flagellinolysin) in Firmicutes (48 homologues) and Proteobacteria (40 homologues); family N4 (Tsh-associated self-cleaving domain) mostly in Firmicutes and Proteobacteria, but also four homologues from Fusobacteria. Homologues of family C118 occur in eight bacterial phyla with two homologues from dsDNA viruses. Family S37 (PS-10 peptidase) occurs in species from seven bacterial phyla, but with some non-peptidase homologues in protozoa and chromists. Homologues of family S85 (small protease) are known from species in eight bacterial phyla, but distribution is patchy in each of them. Family T7 (CwpV protein) is known only from the actinobacterium *Rothia dentocariosa*, the firmicute *Peptoclostridium difficile* and as probable contaminants in the centipede *Strigamia maritima* and Chinese pear (*Pyrus x bretschneideri*). Homologues of family U72 are found in species of Actinobacteria, Nitrospirae, Verrucomicrobia and some Planctomycetes.

There are no peptidase families unique to candidate bacterial (and archaean) phyla. This is not surprising, because few proteins have been characterized from organisms in these phyla and it is not possible to predict that a protein sequence is an example of a novel peptidase family. Only homologues of existing families can be identified. This probably means that with so many candidate phyla there are many families of peptidases waiting to be discovered.

Table 10
Peptidase families found only in Proteobacteria.

Family	Type example
C55	YopJ protein
C70	AvrRpt2 peptidase
C79	ElaD peptidase
C102	GtgE peptidase
C117	SpvD protein
C118	EspL protein
C119	LotA protein
M85	NleC peptidase
M91	NleD peptidase
M99	Csd4 peptidase

4.6. Peptidase families with an origin in archaea and eukaryotes

Archaea and eukaryotes have been considered sister groups [25], a hypothesis given support by the discovery of species identified from metagenomic studies of environmental samples and now classified in the Asgardarchaeota superphylum [42]. The

Asgardarchaeote species from the phylum Lokiarchaeota were shown to be closely related to eukaryotes and to have a repertoire of eukaryote-like proteins. In addition, the number of proteasome components (peptidase family T1) approaches that of eukaryotes, rather than just the two homologues found in species from most other archaean phyla [8]. However, the only peptidase family found in archaea and eukaryotes but not bacteria is A22 (presenilin 1). The function of the archaean homologues is not known, but in animals, A22 peptidases process intramembrane proteins [43]. Family M54 (archaelysin) is also found in archaea and eukaryotes (but absent in plants, arthropods and nematodes), but with some homologues in bacteria, mostly deltaproteobacteria, presumably derived from horizontal gene transfers. Homologues of family S53 (sedolisin) are found in species of archaea (but largely absent from Euryachaeota except for class Thermoplasmata), some bacteria, and eukaryotes. Although found in species from 13 bacterial phyla, homologues are always in a minority of species (except Acidobacteria and most species of *Streptomyces*, *Burkholderia*, *Pectobacterium* and *Xanthomonas*), indicative of horizontal gene transfers. Distribution in eukaryotes is unusual: homologues are found in chordates (but not other animals), fungi (but not yeasts like *Saccharomyces*), the heterokont *Phytophthora* and some protozoa (Filozoa, Mycetozoa, *Naegleria* and *Acanthamoeba*). The functions of M54 and S53 homologues in archaea are unknown.

4.7. Peptidase families with an origin in bacteria and eukaryotes

Although there are peptidase families found in bacteria and eukaryotes, but not archaea, this does not mean that the hypothesis that archaea and eukaryotes are more closely related in false. We have suggested previously that some peptidase families may have appeared in eukaryotes as a result of gene transfers from the protomitochondrion or the protoplastid [44], both derived from endosymbiont bacteria [45]. It is now thought that the mitochondrion was derived from a *Rickettsia*-like bacterium [46], and the plastid from a cyanobacterium [47]. Previously, it had been thought that some protozoa, known as Archezoa, were derived from a eukaryote ancestor from before the origin or these organelles [48], however determination of the genome sequences has shown that this is not the case and archezoans have genes for mitochondrial proteins but have lost their mitochondria [49]. Therefore, the only evidence that a family of proteins has originated from a gene transfer from a proto-organelle to a eukaryote nucleus is the presence of homologues in bacteria and eukaryotes and absence from archaea. However, this hypothesis is undermined if no homologues are known from rickettsias or cyanobacteria.

There are few families found in bacteria and eukaryotes with few or no homologues in archaea: C1, M17 (leucine aminopeptidase 3), M18 (aminopeptidase I) and M41. The following two families also probably had an origin in bacteria and eukaryotes but are reduced in one kingdom (kingdom in brackets): M13 (Protozoa) and S14 (Protozoa). It is surprising that homologues of C1 are not more widely spread in archaeans, because many archaeans live at an acidic pH at which C1 peptidases would be active, but homologues are found only in Euryarchaeota (with a single homologue in the nanoarchaeote Nst1), strongly suggestive of a horizontal gene transfer from either a bacterium or a eukaryote to an ancestral euryarchaeote. Family M17 is also found in Archaea, but most homologues are from species in the phylum Crenarchaeota, with only single homologues from Euryarchaeota, Nanoarchaeota and Thaumarchaeota, again indicative of horizontal gene transfers. There are many homologues from cyanobacteria and rickettsias for families M17, M41 and S14 implying an origin in eukaryotes via an endosymbiont; in addition, the eukaryote homologues of M41 include mitochondrial peptidases [50] and the plant homologues in M17 and S14 are in the chloroplast [51,52]. The only C1 homologue from a rickettsia is a non-peptidase homologue from *Candidatus Caedibacter acanthamoebae*; there are a number of homologues from cyanobacteria, however. There are no homologues from cyanobacteria or rickettsias for family M18, and very few for family M13, so perhaps the genes originated in LUCA and were subsequently lost in an ancestral archaean.

Other families that are predominantly found in bacteria, but not in a majority of phyla, and eukaryotes, and where the eukaryote origin may have been transfer of a gene from a proto-organelle to the host nucleus include the following. Family C83 (gamma-glutamylcysteine dipeptidyltranspeptidase) is found in several bacteria phyla (including cyanobacteria but not rickettsias) and most eukaryote kingdoms (but with a reduced distribution in animals). Family C110 (kyphoscoliosis peptidase) is widely distributed in bacteria (eleven phyla), animals (ten phyla) and fungi (four phyla), but all protozoan homologues are not active, and presence in archaea is restricted to homologues from euryachaeote species (probably derived from a horizontal gene transfer). Presence of a homologue in the foraminiferan *Reticulomyxa filosa* and another in the stramenopile *Aureococcus anophagefferens* could be evidence of presence in early eukaryotes and subsequent gene loss in other chromists and plants. Homologues are present in cyanobacteria but not rickettsias.

Family C45 (acyl-coenzyme A:6-aminopenicillanic acid acyltransferase precursor) is found in 19 bacterial phyla, but also eukaryotes and archaea, though always less than half the phyla in each kingdom. The six archaean homologues are presumed to be the result of horizontal gene transfers. However, it is unlikely that the presence of this family in eukaryotes is the result of a horizontal gene transfer from an endosymbiont that gave rise to either the mitochondrion or plastid, because there are no known homologues from species of the order Rickettsiales and homologues from only two cyanobacterial species. Family C69 (dipeptidase A) is found in species from 14 bacterial phyla (but not *Rickettsia* and few cyanobacteria), but also eukaryotes. There are also 25 homologues from euryarchaeote archaea, probably the result of horizontal gene transfers. Although widely distributed in animals and protozoa, in other eukaryotes homologues are restricted: in plants to chlorophytes; in fungi to ascomycetes; and in chromists to stramenopiles.

Some peptidase families have a bizarre distribution that can only be explained by horizontal gene transfer, but determining the direction of that transfer is difficult. Family M2 (angiotensin-converting enzyme) is found only in animals (11 phyla) and bacteria (12 phyla, but predominantly Proteobacteria, though not *Rickettsia*). For the transfer to have been from a bacterium to an animal means it happened to an ancestral metazoan and the mechanism is obscure; for the transfer to have been the other way around, implies frequent and recent gene sharing amongst bacteria. Family M49 (dipeptidyl-peptidase III) is found in animals; fungi; single-celled organisms such as *Dictyostelium*, *Leishmania* (but not the closely related *Trypanosoma*), *Paramecium*, *Tetrahymena*, *Giardia* and *Entamoeba*; and bacteria mostly from the phylum Bacteriodetes (including the characterized dipeptidyl-peptidase IIIB from *Bacteroides thetaiotaomicron*, a component of the human intestinal flora) [53]. The homologues in bacteria are probably derived from horizontal gene transfers. The distribution in single-celled organisms that includes bikonts (*Paramecium* and *Tetrahymena*) as well as unikonts implies an origin in very early eukaryotes (and subsequent loss in plants). Family M81 (microcystinase MlrC) is found in Proteobacteria (but not *Rickettsia*) and the ascomycete fungi;

presence in a few bacteria species from other phyla and archaea implies widespread horizontal gene transfer, but again the direction is unclear. Family M96 (Tiki1 peptidase) is found in species from several bacterial phyla (but not *Rickettsia* and few cyanobacteria) and most animals (except insects). Family M98 is found in animals (except insects), some protozoa, and bacteria from the phylum Verrucomicrobia and the genus *Vibrio*. Family S51 (dipeptidase E) is found in the majority of bacterial phyla (including Cyanobacteria but not *Rickettsia*), animals (but not mammals, platyhelminthes or nematodes), ochraphytes, and some protozoa (*Leishmania*, *Naegleria*, *Trichomonas* and choanoflagellates). The eukaryote distribution, with homologues from all three major divisions (Bikonta, Unikonta and Excavata), implies either several horizontal gene transfers or multiple gene losses. Family S82 (autocrine proliferation repressor protein A) homologues are found in species from ten bacterial phyla (but not *Rickettsia* and few cyanobacteria), but apart from Planctomycetes and Thermotogae, are not found in most species in a phylum. Homologues, excluding contaminants, are absent from animals, fungi and plants, but are found in some protozoa (Percolozoa, Lobosa, Mycetozoa and Filizoa), again suggestive of horizontal gene transfers. Surprisingly, the distribution of S82 homologues in protozoa is very similar to that of homologues of family S53.

Homologues of family T5 (the self-processing ornithine acetyltransferase precursor) are present in archaea, bacteria and eukaryotes, suggesting an origin in LUCA. However, the archaean homologues are all found in species from a single phylum (Euryarchaeota) and this is probably best explained as an ancient gene transfer from a bacterium to the ancestor of the Euryarchaeota. Homologues are absent from *Rickettsia* but widely distributed in Cyanobacteria, so a possible ancient gene transfer from the protoplastid to the nucleus of an ancestral eukaryote might explain presence in that superkingdom. The distribution in eukaryotes is also unusual, and does not accord with the most recent ideas about eukaryote evolution. Homologues are present in most fungi, the slime mould *Dictyostelium* and the filozoan *Capsaspora*, all of which are opisthokonts. The Opisthokonta also includes animals and microsporidia, from which homologues of T5 are absent. Homologues are also present in the heterokont *Phytophthora*, the rhodophyte *Cyanidioschyzon*, chlorophytes (including *Volvox*, *Chlamydomonas* and *Ostreococcus*), and some, but by no means the majority of, green plants. These are bikonts, which also includes apicomplexans (*e.g. Plasmodium*) and ciliophorans (*e.g. Paramecium*) from which T5 homologues are unknown. T5 homologues are also unknown from the third major division of eukaryotes, the Excavata (which includes *Euglena*, *Trypanosoma*, *Naegleria*, *Giardia* and *Trichomonas*) [25].

4.8. Peptidase families with an origin in eukaryotes

Table 11 lists the families that occur in most phyla of the five eukaryote kingdoms.

Ten other families which were probably also found in the eukaryote ancestor, but which are reduced in one or two kingdoms, are shown in Table 12.

Family A11 (transposon peptidase) probably had its origin in early eukaryotes, but is not present in the majority of any phyla in any kingdom. Other families that also probably originated in eukaryotes but with outliers in other organisms are: C46 (hedgehog protein; reduced in Fungi and Protozoa; outliers in euryarchaeotes, Firmicutes and dsDNA viruses), C89 (acid ceramidase precursor; reduced in all eukaryote kingdoms except animals and with outliers in dsDNA viruses), M43 (cytophagalysin; reduced in all eukaryote kingdoms, absent in plants and with outliers in dsDNA viruses, archaea and bacteria). Families C95 (lysosomal 66.3 kDa protein) and M8 (leishmanolysin) are reduced in some eukaryote kingdoms but absent from fungi.

Table 11
Peptidase families that occur in most phyla of the five eukaryote kingdoms.

Family	Type example
A1	pepsin
C2	calpain
C12	ubiquitinyl hydrolase-L1
C13	legumain
C19	ubiquitin-specific peptidase 14
C48	Ulp1 peptidase
C50	separase
C85	OTLD1 deubiquitinylating enzyme
C97	DeSI-1 peptidase
C115	MINDY-1 protein
M76	Atp23 peptidase
S10	carboxypeptidase Y
S59	nucleoporin 145

Table 12
Peptidase families that occur in all five eukaryote kingdoms but not in most phyla.

Family	Type example	Kingdoms with few represented phyla
A28	DNA-damage inducible protein 1	Chromista
C14	caspase	Fungi
C15	pyroglutamyl-peptidase I	Fungi
C54	autophagin	Fungi
C65	otubain	Chromista
C78	UfSP1 peptidase	Fungi
C86	ataxin	Fungi, Protozoa
C121	MINDY-4 peptidase	Fungi, Protozoa
M12	astacin	Chromista, Protozoa
S28	lysosomal Pro-Xaa carboxypeptidase	Fungi

There are bacterial homologues in families M12 and S10, from several different phyla, but closer examination shows that most bacteria in these phyla do not have homologues, and the presence in bacteria is most likely the result of recent horizontal gene transfers. For example, S10 homologues are found in many species in the bacterial order Alteromonadales including most *Pseudoalteromonas* species, but none are present in some other orders in the same phylum (Proteobacteria), such as Campylobacterales, Nitrosomonadales, Rhodocyclales or Rickettsiales. Homologues of family U74 (neprosin) are found in a selection of plants and fungi as well as bacteria such as *Xanthomonas* that are probably the result of horizontal gene transfers.

Several of these families contain peptidases with functions that are specific for eukaryotes, such as release of protein tags such as ubiquitin, sumo and nedd8 (C12, C19, C48, C65, C78, C85, C86, C97, C115, C121), cytoskeleton remodelling (C2), separation of chromatids during mitosis (C50), cell-cycle control (A28), embryonic development (C46) and apoptosis (C14). Some families contain peptidases restricted to an organelle found only in eukaryotes such as the lysosome (C13, C89, C95, S28 and some A1 homologues), autophagosome (C54), mitochondrion (M76) or nucleus (S59).

The higher taxonomy of eukaryote kingdoms is currently in a state of flux. The seven-kingdom classification of cellular organisms is now thought to be incorrect because the kingdoms Chromista (plant-like protists) and Protozoa (animal-like protists) are now thought to be paraphyletic [54]. Most of the Chromista and green plants (Viridiplantae) are now included in the taxon Diaphoretickes [55]. Only one family of peptidases has its origin in the ancestor of

green plants and chromists: M11 (gametolysin, which degrades the cell wall in *Chlamydomonas*) [56].

The clade Amorphea includes animal-like protists and is divided into the Amoebozoa and Obozoa. There is no peptidase family that originated in any of these three clades. The Obozoa includes Opisthokonta (animals, fungi and choanoflagellates) and the Apusomonadida, which includes the zooflagellate *Thecamonas trahens*. The only peptidase family that *might* have had an origin in the ancestor of the Opisthokonta is C67 (the deubiquitinating CylD peptidase), but all fungal sequences are non-peptidase homologues and there are no homologues from Basidiomycota and few from Ascomycota: an origin in animals seems more likely. Family M80 (the desumoylating Wss1 peptidase) is present in every eukaryote kingdom except animals; homologues of families C95 and M8 are present in most eukaryote kingdoms but absent from fungi. These also do not support the existence of Opithokonta. The proteome of *Thecamonas trahens* is missing some key peptidase families, namely A11, C14, C121 and M76 common to most eukaryotes and C110 and M17 common to bacteria and eukaryotes. Additionally, *Thecamonas* has a homologue from family C64, otherwise found only in animals, and a homologue from M32, otherwise missing in animals and fungi (presumably lost in the ancestral opisthokont).

4.8.1. Peptidase families with an origin in protozoa

The only family with a possible origin in Protozoa is M66 (StcE peptidase), homologues of which are found in members of the amoebozoan class Dictyostelia. Homologues are also found in bacteria from five phyla, but these are mostly single species implying horizontal gene transfers. The peptidase is, however, more widely distributed in species from the gammaproteobacterial order Vibrionales. A homologue has been characterized as a virulence factor in *E. coli* [57].

4.8.2. Peptidase families with an origin in fungi

Family M77 (dimethylallyltryptophan synthase 7-DMATS) is present only in fungi and with a probable origin in the ancestor of fungi. Although family G1 (scytalidoglutamic peptidase) homologues are found in bacteria from six different phyla, the distribution amongst species with completed genomes in sparse, more indicative of horizontal gene transfers, and the origin of the family also appears to be in fungi. Family S64 (Ssy5 peptidase) is found only from members of Ascomycota and probably had its origin in the ancestor of that phylum. Ssy5 activates a signalling pathway in response to and for the uptake of extracellular amino acids [58]. Homologues of family C100 (agglutinin peptidase, which actives a nematocide [59]) are found, excluding contaminants, in species of Basidiomycota, and the family probably had its origin in the ancestor of that phylum. Although most family M36 (fungalysin, a virulence factor) homologues are from fungi (527), there is a considerable number from bacteria (237) classified into ten phyla: too many to be contaminants. It seems much more likely that an ancient horizontal gene transfer occurred, probably from a fungus to a bacterium. Most homologues from M35 (penicillolysin) are from species of Ascomycota, but with a considerable number of homologues from Proteobacteria, including the characterized EcpA peptidase from *Xanthomonas oryzae* and *Aeromonas* species, and numerous uncharacterized homologues from the sponge *Amphimedon queenslandica*. The non-fungal homologues are assumed to be derived from horizontal gene transfers from a fungus. Again, most N9 (intein-containing V-type proton ATPase catalytic subunit A) homologues are from Ascomycota, but there are also homologues from bacteria and dsDNA viruses, presumably the result of horizontal gene transfers.

4.8.3. Peptidase families with an origin in chromists

No peptidase family appears to have had its origin in chromists.

4.8.4. Peptidase families with an origin in plants

Family M102 (DA1 peptidase), omitting contaminants, is restricted to Streptophyta and had its origin presumably in the ancestor to that phylum. The DA1 peptidase is ubiquitin-dependent and degrades a deubiquitinating enzyme, and regulates cell proliferation [60]. Family N11 (intein-containing chloroplast ATP-dependent peptide lyase) probably originated in the ancestor of chlorophytes, but homologues, presumably the result of horizontal gene transfers, are also found in some euryarchaeotes, bacteria, fungi and dsDNA viruses.

4.8.5. Peptidase families with an origin in animals

The following families are found only in animals and presumably had their origin in the ancestral animal: C98 (USPL1 peptidase), C101 (OTULIN peptidase) and S71 (MUC1 self-cleaving mucin). Omitting possible false positives and contaminants, families C111 (coagulation factor XIIIa), M87 (chloride channel accessory protein 1), S60 (lactoferrin; absent from nematodes of the class Chromadorea), S72 (dystroglycan) and S81 (destabilase; missing from chordates) can also be added to this list. Other families that predominate in animals, but with outliers in other kingdoms are: C64 (outliers in Chromista, Protozoa and plants) and P2 (EGF-like module containing mucin-like hormone receptor-like 2; outliers in Protozoa).

Peptidases from these ten families do not have roles specific to animals, but are further refinements of functions found in eukaryotes. Peptidases from families C64, C98 and C101 are required for releasing protein tags such as sumo and ubiquitin [61–63]. The proteins from M87, P2, S71 and S72 are self-cleaving. Peptidases in families C111, S60 and S81 are all involved in animal defence systems. Lactoferrin has been shown to cleave and inactivate peptidases from bacterial pathogens [64]. Coagulation factor XIIIa and destabilase are multifunctional, but both can act as isopeptidases to break the isopeptide bonds in fibrin and the peptidoglycan of bacterial cell walls [65,66].

Of all the many animal phyla, Chordata is the only one to which peptidase families are (mostly) restricted. The three families are M100 (Spartan peptidase; with homologues from the crustacean *Daphnia pulex* and the mesozoan *Intoshia linei*); S68 (PIDD autoprocessing protein) and S79 (CARD8 self-cleaving protein; with a single homologue from the nematode *Litomosoides sigmodontis*). Spartan is related to Wss1 from yeast (family M80) and probably processes DNA-protein crosslinks [67].

5. Conclusions

Although there are over 270 different families of peptidases, and proteolytic processing is a process required by all living things, no family of peptidases has homologues from species in every phyla of organisms. This includes supposedly ubiquitous peptidase families such as M24 (which includes methionyl aminopeptidase), which is absent from species of Cryptophyta, and S26 (signal peptidase) which is absent from species of Mollicutes and Pinopsida.

From the analysis of the distribution of peptidase homologues amongst phyla of organisms, some 33 families of peptidases are identified as most likely to have originated in the last universal common ancestor (LUCA). LUCA is predicted to have included peptidases of different tertiary structures, different catalytic types, as well as peptidases of different activities, such as aminopeptidases, carboxypeptidases and endopeptidases. These must have developed in organisms that predate LUCA. No family of aspartic or glutamic peptidases or asparagine peptide lyases are predicted to

have had an origin in LUCA.

Many families of peptidases have originated since LUCA and many of them are restricted to particular organism kingdoms, phyla or even families. There are many families of peptidases restricted only to viruses. None of these families includes viruses with different genetic material (double-stranded DNA, double-stranded RNA or single-stranded RNA) and in most cases all the homologues in one of these families is restricted to viruses from a single viral family. There are 37 families predicted to have had their origin in the ancestral bacterium, with a further 38 originating in the ancestors of five bacterial phyla. Peptidase family A5 is the only one to have originated in the ancestral archaean, but two other families (A37 and C106) originate in archaeal phyla. There are 32 families predicted to have had their origin in the ancestral eukaryote, plus a further 11 that originated in the ancestral animal, three in the ancestral fungus, and one (M11) in the ancestor of the Diaphoretickes. A further ten families had their origins in the ancestor of a phylum (three in a chordate, three in an ascomycete, and one each in an amoebazoan, basidiomycete, streptophyte and chlorophyte).

Recent hypotheses have suggested that eukaryotes are more closely related in archaea than bacteria, and may be derived from a recently identified group of archaeans from the superphylum Asgardarchaeota. However, more peptidase families are common to bacteria and eukaryotes (17) than to archaea and eukaryotes (3), and there are no peptidase families common to only species of Asgardarchaeota and eukaryotes. It has also been suggested that some genes occur in eukaryotes as a result of horizontal transfers from organelles such as the mitochondrion and the plastid which are derived from endosymbiont bacteria. This may be true for eight of the 17 families common to bacteria and eukaryotes.

Authors contributions

NDR devised the paper, collected the data, performed the analyses and wrote the initial draft. AGB oversaw the project and approved the final version.

Declarations of interest

None.

Acknowledgements

The authors would like to thank Dr Alan Barrett for his help in maintaining the references and identification of new families and peptidases in the MEROPS database. We would also like to thank the web team at EBI for help in maintaining the MEROPS website. Alex Bateman is funded by the European Molecular Biology Laboratory.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biochi.2019.07.026.

References

[1] Enzyme Nomenclature, Academic Press, San Diego, California, 1992.
[2] N.D. Rawlings, A.J. Barrett, A. Bateman, Asparagine peptide lyases: a seventh catalytic type of proteolytic enzymes, J. Biol. Chem. 286 (2011) 38321–38328, https://doi.org/10.1074/jbc.M111.260026.
[3] G. Dodson, A. Wlodawer, Catalytic triads and their relatives, Trends Biochem. Sci. 23 (1998) 347–352.
[4] I. Botos, E.E. Melnikov, S. Cherry, J.E. Tropea, A.G. Khalatova, F. Rasulova, Z. Dauter, M.R. Maurizi, T.V. Rotanova, A. Wlodawer, A. Gustchina, The catalytic domain of *Escherichia coli* Lon protease has a unique fold and a Ser-Lys dyad in the active site, J. Biol. Chem. 279 (2004) 8140–8148.
[5] B.L. Vallee, D.S. Auld, Cocatalytic zinc motifs in enzyme catalysis, Proc. Natl. Acad. Sci. U. S. A. 90 (1993) 2715–2718.
[6] A.J. Barrett, Classification of peptidases, Methods Enzymol. 244 (1994) 1–15.
[7] N.D. Rawlings, A.J. Barrett, Evolutionary families of peptidases, Biochem. J. 290 (1993) 205–218. PMID: 8439290.
[8] N.D. Rawlings, A.J. Barrett, P.D. Thomas, X. Huang, A. Bateman, R.D. Finn, The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database, Nucleic Acids Res. 46 (D1) (2018) D624–D632, https://doi.org/10.1093/nar/gkx1134. PMID:29145643.
[9] UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (D1) (2017) D158–D169, https://doi.org/10.1093/nar/gkw1099.
[10] NCBI Resource Coordinators, Database resources of the national center for biotechnology information, Nucleic Acids Res. 46 (D1) (2018) D8–D13, https://doi.org/10.1093/nar/gkx1095.
[11] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. PMID: 2231712.
[12] R.D. Finn, J. Clements, W. Arndt, B.L. Miller, T.J. Wheeler, F. Schreiber, A. Bateman, S.R. Eddy, HMMER web server: 2015 update, Nucleic Acids Res. 43 (W1) (2015) W30–W38, https://doi.org/10.1093/nar/gkv397. PMID: 25943547.
[13] N.D. Rawlings, F.R. Morton, The MEROPS batch BLAST: a tool to detect peptidases and their non-peptidase homologues in a genome, Biochimie 90 (2008) 243–259. PMID: 17980477.
[14] wwPDB consortium, Protein Data Bank: the single global archive for 3D macromolecular structure data, Nucleic Acids Res. 47 (D1) (2019) D520–D528, https://doi.org/10.1093/nar/gky949. PMID:30357364.
[15] L. Holm, C. Sander, Dali: a network tool for protein structure comparison, Trends Biochem. Sci. 20 (1995) 478–480. PMID: 8578593.
[16] R. Sanjuán, P. Domingo-Calap, Mechanisms of viral mutation, Cell. Mol. Life Sci. 73 (2016) 4433–4448.
[17] Y. Luo, P. Pfister, T. Leisinger, A. Wasserfallen, Pseudomurein endoisopeptidases PeiW and PeiP, two moderately related members of a novel family of proteases produced in *Methanothermobacter* strains, FEMS Microbiol. Lett. 208 (1) (2002) 47–51. Feb 19.
[18] C. Schlieker, G.A. Korbel, L.M. Kattenhorn, H.L. Ploegh, A deubiquitinating activity is conserved in the large tegument protein of the herpesviridae, J. Virol. 79 (2005) 15582–15585.
[19] M. Allaire, M.M. Chernaia, B.A. Malcolm, M.N. James, Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinases, Nature 369 (1994) 72–76. PMID:8164744.
[20] B. Goulet, A. Baruch, N.S. Moon, M. Poirier, L.L. Sansregret, A. Erickson, M. Bogyo, A. Nepveu, A cathepsin L isoform that is devoid of a signal peptide localizes to the nucleus in S phase and processes the CDP/Cux transcription factor, Mol. Cell 14 (2004) 207–219. PMID:15099520.
[21] J. Shen, Y. Zeng, X. Zhuang, L. Sun, X. Yao, P. Pimpl, L. Jiang, Organelle pH in the *Arabidopsis* endomembrane system, Mol. Plant 6 (2013) 1419–1437, https://doi.org/10.1093/mp/sst079.
[22] K.B. Sieber, R.E. Bromley, J.C. Dunning Hotopp, Lateral gene transfer between prokaryotes and eukaryotes, Exp. Cell Res. 358 (2017) 421–426, https://doi.org/10.1016/j.yexcr.2017.02.009.
[23] J.D. Gary, S. Clarke, Purification and characterization of an isoaspartyl dipeptidase from *Escherichia coli*, J. Biol. Chem. 270 (1995) 4076–4087.
[24] J.P. Simmer, R.E. Kelly, A.G. Rinker Jr., B.H. Zimmermann, J.L. Scully, H. Kim, D.R. Evans, Mammalian dihydroorotase: nucleotide sequence, peptide sequences, and evolution of the dihydroorotase domain of the multifunctional protein CAD, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 174–178.
[25] T. Cavalier-Smith, The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa, Int. J. Syst. Evol. Microbiol. 52 (2002) 297–354.
[26] C.R. Woese, O. Kandler, M.L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 4576–4579. PMID: 2112744.
[27] K.C. Keiler, P.R. Waller, R.T. Sauer, Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA, Science 271 (1996) 990–993.
[28] M. Kaser, M. Kambacheld, B. Kisters-Woike, T. Langer, Oma1, a novel membrane-bound metallopeptidase in mitochondria with activities overlapping with the m-AAA protease, J. Biol. Chem. 278 (2003) 46414–46423.
[29] J. Skórko-Glonek, A. Wawrzynów, K. Krzewski, K. Kurpierz, B. Lipińska, Site-directed mutagenesis of the HtrA (DegP) serine protease, whose proteolytic activity is indispensable for *Escherichia coli* survival at elevated temperatures, Gene 163 (1995) 47–52.
[30] D.R. Wessner, The origins of viruses, Nature Ed 3 (2010) 37.
[31] B. La Scola, S. Audic, C. Robert, L. Jungang, X. de Lamballerie, M. Drancourt, R. Birtles, J.M. Claverie, D. Raoult, A giant virus in amoebae, Science 299 (2003) 2033. PMID:12663918.
[32] V.H. Tang, G.A. Stewart, B.J. Chang, *Dermatophagoides pteronyssinus* lytFM encoding an NlpC/P60 endopeptidase is also present in mite-associated bacteria that express LytFM variants, FEBS Open Bio 7 (2017) 1267–1280, https://doi.org/10.1002/2211-5463.12263. eCollection 2017 Sep. PMID:28904857.
[33] N.D. Rawlings, Evolution of the thermopsin peptidase family (A5), PLoS One 8 (2013), https://doi.org/10.1371/journal.pone.0078998 e78998.
[34] M.F. Abdul Halim, F. Pfeiffer, J. Zou, A. Frisch, D. Haft, S. Wu, N. Tolić, H. Brewer, S.H. Payne, L. Paša-Tolić, M. Pohlschroder, *Haloferax volcanii* archaeosortase is required for motility, mating, and C-terminal processing of the S-layer

glycoprotein, Mol. Microbiol. 88 (2013) 1164–1175, https://doi.org/10.1111/mmi.12248.
[35] A.C. Martín, R. López, P. García, Pneumococcal bacteriophage Cp-1 encodes its own protease essential for phage maturation, J. Virol. 72 (1998) 3491–3494. PMID: 9525689.
[36] E.A. Wall, J.H. Caufield, C.E. Lyons, K.A. Manning, T. Dokland, G.E. Christie, Specific N-terminal cleavage of ribosomal protein L27 in *Staphylococcus aureus* and related bacteria, Mol. Microbiol. 95 (2015) 258–269, https://doi.org/10.1111/mmi.12862.
[37] J.C. Chen, A.K. Hottes, H.H. McAdams, P.T. McGrath, P.H. Viollier, L. Shapiro, Cytokinesis signals truncation of the PodJ polarity factor by a cell cycle-regulated protease, EMBO J. 25 (2006) 377–386.
[38] Y.T. Yu, L. Snyder, Translation elongation factor Tu cleaved by a phage-exclusion system, Proc. Natl. Acad. Sci. U. S. A. 91 (1994) 802–806.
[39] M. Mallozzi, V.K. Viswanathan, G. Vedantam, Spore-forming bacilli and clostridia in human disease, Future Microbiol. 5 (2010) 1109–1123, https://doi.org/10.2217/fmb.10.60.
[40] R. Qiu, W. Pei, L. Zhang, J. Lin, G. Ji, Identification of the putative staphylococcal AgrB catalytic residues involving the proteolytic cleavage of AgrD to generate autoinducing peptide, J. Biol. Chem. 280 (2005) 16695–16704.
[41] I.Y. Khudyakov, J.W. Golden, Different functions of HetR, a master regulator of heterocyst differentiation in *Anabaena* sp. PCC 7120, can be separated by mutation, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 16040–16045.
[42] A. Spang, J.H. Saw, S.L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A.E. Lind, R. van Eijk, C. Schleper, L. Guy, T.J.G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes, Nature 521 (2015) 173–179, https://doi.org/10.1038/nature14447.
[43] Z. Zhang, P. Nadeau, W. Song, D. Donoviel, M. Yuan, A. Bernstein, B.A. Yankner, Presenilins are required for gamma-secretase cleavage of beta-APP and transmembrane cleavage of Notch-1, Nat. Cell Biol. 2 (2000) 463–465.
[44] A.J. Barrett, N.D. Rawlings, Evolutionary origins of the families of peptidases in eukaryotes, in: J.S. Bond, A.J. Barrett (Eds.), Proteolysis and Protein Turnover, Portland Press, London, 1993, pp. 73–79.
[45] T. Cavalier-Smith, J.J. Lee, Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles, J. Protozool. 32 (1985) 376–379.
[46] S.G. Andersson, A. Zomorodipour, J.O. Andersson, T. Sicheritz-Pontén, U.C. Alsmark, R.M. Podowski, A.K. Näslund, A.S. Eriksson, H.H. Winkler, C.G. Kurland, The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria, Nature 396 (1998) 133–140.
[47] J.N. Timmis, M.A. Ayliffe, C.Y. Huang, W. Martin, Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes, Nat. Rev. Genet. 5 (2004) 123–135.
[48] T. Cavalier-Smith, Molecular phylogeny. Archaebacteria and Archezoa, Nature 339 (1989) 100–101.
[49] H. Brinkmann, H. Philippe, The diversity of eukaryotes and the root of the eukaryotic tree, Adv. Exp. Med. Biol. 607 (2007) 20–37.
[50] K. Leonhard, J.M. Herrmann, R.A. Stuart, G. Mannhaupt, W. Neupert, T. Langer, AAA proteases with catalytic sites on opposite membrane surfaces comprise a proteolytic system for the ATP-dependent degradation of inner membrane proteins in mitochondria, EMBO J. 15 (1996) 4218–4229.
[51] J. Narváez-Vásquez, C.J. Tu, S.Y. Park, L.L. Walling, Targeting and localization of wound-inducible leucine aminopeptidase A in tomato leaves, Planta 227 (2008) 341–351.
[52] A.K. Clarke, The chloroplast ATP-dependent Clp protease in vascular plants - new dimensions and future challenges, Physiol. Plant. 145 (2012) 235–244, https://doi.org/10.1111/j.1399-3054.2011.01541.x.
[53] I. Sabljić, N. Meštrović, B. Vukelić, P. Macheroux, K. Gruber, M. Luić, M. Abramić, Crystal structure of dipeptidyl peptidase III from the human gut symbiont *Bacteroides thetaiotaomicron*, PLoS One 12 (2017), e0187295, https://doi.org/10.1371/journal.pone.0187295.
[54] T. Cavalier-Smith, Eukaryote kingdoms: seven or nine? Biosystems 14 (1981) 461–481.
[55] F. Burki, K. Shalchian-Tabrizi, J. Pawlowski, Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes, Biol. Lett. 4 (2008) 366–369, https://doi.org/10.1098/rsbl.2008.0224.
[56] T. Kinoshita, H. Fukuzawa, T. Shimada, T. Saito, Y. Matsuda, Primary structure and expression of a gamete lytic enzyme in *Chlamydomonas reinhardtii*: similarity of functional domains to matrix metalloproteases, Proc. Natl. Acad. Sci. U. S. A. 89 (1992) 4693–4697.
[57] W.W. Lathem, T.E. Grys, S.E. Witowski, A.G. Torres, J.B. Kaper, P.I. Tarr, R.A. Welch, StcE, a metalloprotease secreted by *Escherichia coli* O157:H7, specifically cleaves C1 esterase inhibitor, Mol. Microbiol. 45 (2002) 277–288.
[58] F. Abdel-Sater, M. El Bakkoury, A. Urrestarazu, S. Vissers, B. André, Amino acid signaling in yeast: casein kinase I and the Ssy5 endoprotease are key determinants of endoproteolytic activation of the membrane-bound Stp1 transcription factor, Mol. Cell. Biol. 24 (2004) 9771–9785.
[59] T. Wohlschlager, A. Butschi, K. Zurfluh, S.C. Vonesch, U. auf dem Keller, P. Gehrig, S. Bleuler-Martinez, M.O. Hengartner, M. Aebi, M. Künzler, Nematotoxicity of *Marasmius oreades* agglutinin (MOA) depends on glycolipid binding and cysteine protease activity, J. Biol. Chem. 286 (2011) 30337–30343, https://doi.org/10.1074/jbc.M111.258202.
[60] H. Dong, J. Dumenil, F.H. Lu, L. Na, H. Vanhaeren, C. Naumann, M. Klecker, R. Prior, C. Smith, N. McKenzie, G. Saalbach, L. Chen, T. Xia, N. Gonzalez, M. Seguela, D. Inze, N. Dissmeyer, Y. Li, M.W. Bevan, Ubiquitylation activates a peptidase that promotes cleavage and destabilization of its activating E3 ligases and diverse growth regulatory proteins to limit cell proliferation in *Arabidopsis*, Genes Dev. 31 (2017) 197–208, https://doi.org/10.1101/gad.292235.116.
[61] P.C. Evans, T.S. Smith, M.J. Lai, M.G. Williams, D.F. Burke, K. Heyninck, M.M. Kreike, R. Beyaert, T.L. Blundell, P.J. Kilshaw, A novel type of deubiquitinating enzyme, J. Biol. Chem. 278 (2003) 23180–23186.
[62] S. Schulz, G. Chachami, L. Kozaczkiewicz, U. Winter, N. Stankovic-Valentin, P. Haas, K. Hofmann, H. Urlaub, H. Ovaa, J. Wittbrodt, E. Meulmeester, F. Melchior, Ubiquitin-specific protease-like 1 (USPL1) is a SUMO isopeptidase with essential, non-catalytic functions, EMBO Rep. 13 (2012) 930–938, https://doi.org/10.1038/embor.2012.125.
[63] E. Rivkin, S.M. Almeida, D.F. Ceccarelli, Y.C. Juang, T.A. MacLean, T. Srikumar, H. Huang, W.H. Dunham, R. Fukumura, G. Xie, Y. Gondo, B. Raught, A.C. Gingras, F. Sicheri, S.P. Cordes, The linear ubiquitin-specific deubiquitinase gumby regulates angiogenesis, Nature 498 (2013) 318–324, https://doi.org/10.1038/nature12296.
[64] A.G. Plaut, J. Qiu, J.W. St Geme 3rd, Human lactoferrin proteolytic activity: analysis of the cleaved region in the IgA protease of *Haemophilus influenzae*, Vaccine 19 (Suppl 1) (2000) S148–S152.
[65] A. Heil, J. Weber, C. Büchold, R. Pasternack, M. Hils, Differences in the inhibition of coagulation factor XIII-A from animal species revealed by Michael Acceptor- and thioimidazol based blockers, Thromb. Res. 131 (2013) e214–e222, https://doi.org/10.1016/j.thromres.2013.02.008.
[66] R. Josková, M. Silerová, P. Procházková, M. Bilej, Identification and cloning of an invertebrate-type lysozyme from Eisenia andrei, Dev. Comp. Immunol. 33 (2009) 932–938, https://doi.org/10.1016/j.dci.2009.03.002.
[67] J. Stingele, B. Habermann, S. Jentsch, DNA-protein crosslink repair: proteases as DNA repair enzymes, Trends Biochem. Sci. 40 (2015) 67–71, https://doi.org/10.1016/j.tibs.2014.10.012.