Database, Vol. 2013, Article ID bat022, doi:10.1093/database/bat022

Original article

Identification and prioritization of novel uncharacterized peptidases for biochemical characterization

Neil D. Rawlings[1,2,*]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK and [2]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

***Corresponding author:** Tel: +44 1223 494525; Fax: +44 1223 496802; Email: neil.rawlings@ebi.ac.uk or ndr@sanger.ac.uk

Submitted 30 November 2012; Revised 13 February 2013; Accepted 13 March 2013

Citation details: Rawlings,N.D. Identification and prioritization of novel uncharacterized peptidases for biochemical characterization. *Database* (2013) Vol. 2013: article ID bat022; doi: 10.1093/database/bat022.

Genome sequencing projects are generating enormous amounts of biological data that require analysis, which in turn identifies genes and proteins that require characterization. Enzymes that act on proteins are especially difficult to characterize because of the time required to distinguish one from another. This is particularly true of peptidases, the enzymes that activate, inactivate and degrade proteins. This article aims to identify clusters of sequences each of which represents the species variants of a single putative peptidase that is widely distributed and is thus merits biochemical characterization. The MEROPS database maintains large collections of sequences, references, substrate cleavage positions and inhibitor interactions of peptidases and their homologues. MEROPS also maintains a hierarchical classification of peptidase homologues, in which sequences are clustered as species variants of a single peptidase; homologous sequences are assembled into a family; and families are clustered into a clan. For each family, an alignment and a phylogenetic tree are generated. By assigning an identifier to a peptidase that has been biochemically characterized from a particular species (called a holotype), the identifier can be automatically extended to sequences from other species that cluster with the holotype. This permits transference of annotation from the holotype to other members of the cluster. By extending this concept to all peptidase homologues (including those of unknown function that have not been characterized) from model organisms representing all the major divisions of cellular life, clusters of sequences representing putative peptidases can also be identified. The 42 most widely distributed of these putative peptidases have been identified and discussed here and are prioritized as ideal candidates for biochemical characterization.

Database URL: http://merops.sanger.ac.uk

Introduction

Enzymes that act on biopolymers are poorly classified in terms of the reactions they catalyse. It is frequently difficult, or even impossible, to fully describe the signals, which may be sequence- or structure-based, by which these enzyme recognize their target substrates. Classification by reaction catalysed also masks evolutionary relationships because convergent evolution can give rise to enzymes with similar specificities that are otherwise unrelated, and divergent evolution may give rise to homologous enzymes with different specificities. These factors apply to enzymes, such as nucleases, which act on nucleic acids; glycosyl hydrolases, which act on carbohydrates; lipases, which act on lipids; and protein kinases and peptidases, which act on proteins and peptides. For each of these types of enzymes, because the method of classification by reaction catalysed in *Enzyme Nomenclature* (1) has been found to be

Database, Vol. 2013, Article ID bat022, doi:10.1093/database/bat022

insufficient, separate databases for their classification and nomenclature have been developed. These include REBASE [nucleic acid restriction enzymes (2)], CAZy [glycosyl hydrolases (3)], LIPABASE [lipases (4)], KINBASE [protein kinases (5)] and MEROPS [peptidases (6)].

Peptidases are enzymes that hydrolyse the carbon–nitrogen bonds between amino acids (peptide bonds). Peptidases are not only important for the complete degradation of proteins and peptides, so that the component amino acids can be re-used, but also for post-translational processing of proteins, so that activation is delayed until the protein has reached the correct cellular or extracellular compartment, or the necessary developmental, physiological or environmental conditions prevail. Examples of the former include digestion of proteins in food and the turnover of cytoplasmic and phagocytosed proteins in the lysosome. Examples of the latter include activation of enzymes and the processing of prohormone precursors. Proteolytic cascades exist to amplify the original biological signal, such as in blood coagulation and apoptosis. The latter process highlights a function of peptidases in the destruction of proteins and peptidases in switching off biological signals, including the degradation of peptide hormones and neurotransmitters.

A peptidase is targeted to a specific substrate by recognition of one or more binding sites. A substrate may bind directly at the peptidase active site, where the peptidase recognizes a sequence or structural motif in the substrate, or to a site known as an exosite, some distance from the enzyme active site. For a peptidase, such as the lysosomal cysteine peptidase cathepsin B, no clear sequence motif for substrate binding has emerged, despite the identification of hundreds of cleavages sites in substrates (7). This makes it is difficult to classify peptidases by reaction catalysed.

The MEROPS database classifies peptidases along evolutionary principles, so that homologous sequences are included in the same family, and homologous tertiary structures are included in the same clan, according to the methods of Rawlings and Barrett (8). A family consists of sequences that can be shown to be directly or indirectly homologous to a single sequence designated the family-type example. Sequences that represent the same peptidase from different organisms are given a unique identifier, composed of the family name, a dot and a sequential number. For example, the well-characterized lysosomal cysteine peptidase cathepsin B is C01.060, being a member of family C1 (9). A MEROPS identifier is initially applied to a single sequence, known as the *holotype*, representing a peptidase that has been biochemically characterized. Other sequences given the same MEROPS identifier may or may not have been characterized, but all are assumed to represent an enzyme with similar biochemical properties (10). Not all sequences can be assigned to a particular MEROPS identifier, because some sequences are divergent and not closely related to any existing holotype.

A peptidase is assumed to be biochemically characterized when it can be measured quantitatively, which usually means that an assay of some kind is possible. In practical terms, this means that an article has been published, which describes characterization of the peptidase. Ideally, this would include information on substrates cleaved (even if the cleavage positions are not known), inhibitor profiles and the amino acid sequence of the peptidase. MEROPS maintains an extensive collection of references (11), known cleavage sites in substrates [including synthetic peptides, as well as proteins (12)] and a collection of interactions between peptidases and inhibitors, both naturally occurring and synthetic (13). An identifier is only established if the peptidase in question can be shown to differ from an existing peptidase because there is a significant difference in the action on one or more substrates or in the interaction with one or more inhibitors, or, more likely, because the sequence is too divergent to be included within an existing MEROPS identifier.

In the current release of the MEROPS database (release 9.7, 1 August 2012), for peptidases there are 60 clans, 236 families and 3944 identifiers, derived from 264 093 sequences. Of the sequences included in the database, 104 568 sequences (39.6%) cannot be assigned to a MEROPS identifier and have been assigned to miscellaneous family codes. Almost all of these sequences are derived from genome sequencing projects, often from unusual microbes, and are unlikely to be characterized. Besides their physiological roles, many peptidases have biomedical, scientific and industrial uses (14), and several peptidases from unfamiliar organisms have been used, in cheese-making and biological washing powders, for example. It is thus possible that many of the uncharacterized peptidases will prove to have unique specificities and be equally useful, or at least physiologically interesting. The problem of genome sequencing projects yielding large amounts of data relating to uncharacterized proteins has previously been addressed by the curators of the Pfam database because the database contains a considerable number of domain families of unknown function (15).

Figure 1 shows the number of peptidase homologue sequences in the MEROPS database at the end of each year from 1998 to 2012, and the number of these sequences assigned to a MEROPS identifier. Also shown is the number of MEROPS identifiers, which equates to the number of different characterized peptidases. In 1998, there were 7381 sequences and 117 identifiers, with 3786 sequences assigned to identifiers (51.3%). The growth in sequences is exponential, yet the increase in identifiers is linear, which implies that the gap between the number of

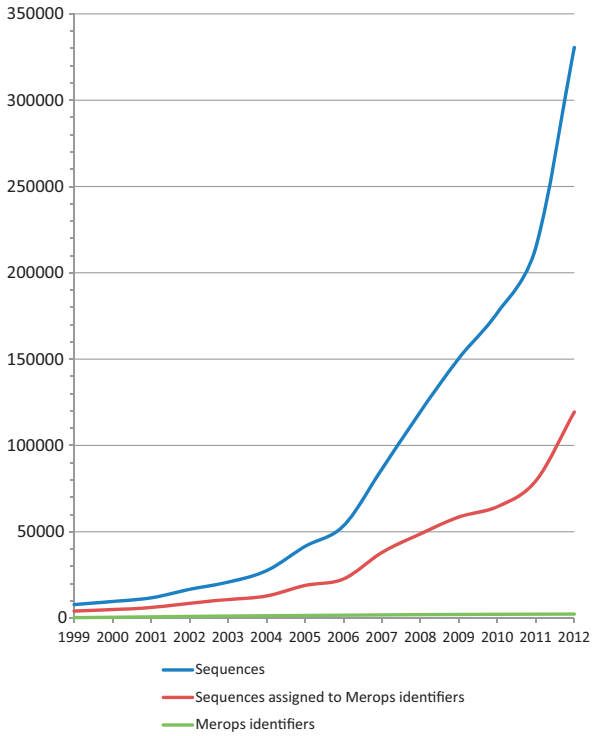

Figure 1. Increase in the number of peptidase homologue sequences, 1998–2012. The graph shows the cumulative number of peptidase homologue sequences added to the MEROPS database per year since 1998, when submission dates were first recorded. Also shown are the number of these sequences that have been assigned to a MEROPS identifier and the number of MEROPS identifiers.

sequences and the proportion that can be assigned to identifiers will only widen. Clearly, characterization of peptidases is lagging behind genome sequencing, and a high-throughput mechanism for characterization is required. An important initial aspect of such a high-throughput system would be to recognize those peptidases most worthy of characterization. In this article, I will identify the uncharacterized peptidases that are most widely distributed among different species because widely distributed peptidases are most likely to have key physiological roles and to be worthy of biochemical characterization.

Materials and Methods

Sequence sources

Sequences homologous to peptidase family-type examples were identified from BLASTP (16) searches of the NCBI non-redundant protein sequence database (17). A Perl program was written to submit that part of the peptidase protein sequence that bears the active site residues (known as the peptidase unit) from each family-type example, to the BLASTP search, with low complexity sequence masked. Searches were performed so that the maximum number of hits in which the expect value (e) was 0.001 or lower was returned. If the same hit sequence was returned in more than one results file, only the match with the lowest E value was retained. A Perl program was written to check that no residues had been filed more than once in the MEROPS MySQL database.

Classification of sequences

A pipeline consisting of a series of Perl programs was assembled to enter sequence data returned by the BLASTP searches into the MEROPS MySQL database. For each BLASTP search, all hits were entered as either a miscellaneous peptidase homologue or a miscellaneous non-peptidase homologue. A non-peptidase homologue was identified as a sequence having one or more active site residues replaced with an amino acid not known to occupy this position in an active peptidase. A list of peptidase active site residues was published by Rawlings and Morton (18). For most peptidases, the active site consists of a catalytic dyad or triad, but in the case of metallopeptidases, the ligands that bind the catalytic metal ion or ions are also considered active site residues. Non-peptidase homologues are not considered further in this article.

An alignment of peptidase units from all homologues in a family considered likely to be active peptidases was prepared using the Muscle software (19). Sequences in which active site residues were misaligned because the sequences were too divergent to align correctly, or missing, because the known sequence is only a fragment, were removed from the alignment. This ensured that false-positive results were removed from the analyses. A difference matrix was calculated from the alignment and converted to accept point mutations using the PAM250 matrix (20). A phylogenetic tree was then computed using the UPGMA algorithm (21) as implemented in the Quicktree program (22).

A bespoke Perl program was written to traverse the tree, to identify the tip that corresponded to each holotype and then to walk down the tree so that all sequences derived from the same node as the holotype were assigned the same MEROPS identifier, provided that the following three conditions were met:

(1) No other holotype was included in the cluster.
(2) No tip in which the sequence length differed from the holotype by a hundred or more residues was included in the cluster.
(3) The node from which all tips in the cluster were derived did not correspond to an average sequence difference >50% of aligned residues.

In practice, the first two of these criteria had exceptions, as discussed later in the text.

Database, Vol. 2013, Article ID bat022, doi:10.1093/database/bat022

Results

Establishment of MEROPS identifiers for all peptidases from model organisms

MEROPS identifiers were set-up for all homologues assumed to be proteolytically active from a range of model organisms (Table 1). The organisms chosen represent the major groups of cellular life: animals, a plant, fungi, a protist, a Gram-negative bacterium, a Gram-positive bacterium and an archaean. For each of the bacteria, a single reference strain was selected: *Escherichia coli* strain K12 substrain MG1655 and *Bacillus subtilis* strain 168, which are also the strains chosen to be representative by the UniProt database (23). The number of peptidases in a model organism varies greatly. It may seem surprising that a mouse has more peptidase genes than any other model organism, or that the plant *Arabidopsis thaliana* has more peptidase genes than the fruit fly *Drosophila melanogaster* or the nematode *Caenorhabditis elegans*. The slime mould *Dictyostelium discoideum* has more peptidase genes than either of the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. The Gram-positive bacterium *B. subtilis* has nearly twice as many peptidases as the Gram-negative bacterium *E. coli*, and more than any of the three unicellular eukaryotes has. The model organism with fewest peptidase genes is the archaean *Pyrococcus furiosus*, although it is entirely possible that with so many uncharacterized proteins in this proteome, many of these will subsequently be discovered to be peptidases from novel families. Of the 2045 proteins from the *P. furiosus* genome, only 1095 have domains included in families in the Pfam database (24), of which 198 are included in domains of unknown function. In other words, 63.2% of proteins from *P. furiosus* have unknown functions.

Table 1 shows the number of holotypes established in each model organism, and the numbers that have been biochemically characterized, i.e. those that have been the subject of a published study and those with known substrates and inhibitor interactions. The holotypes that are biochemically uncharacterized are the putative peptidases that are the subject of this study.

Table 2 shows the number of peptidases common to all model organisms. Equivalent peptidases are tentatively assumed to perform the same physiological roles and to have similar protein architectures and substrate specificities. It is clear from the table that only human and mouse have a significant proportion of equivalent peptidases (77.4% of human peptidases and 72.7% of mouse peptidases). These two organisms are from the same taxonomic class, Mammalia, whereas the other model organisms are from different phyla. This indicates that there are very few sequences from organisms in different phyla that can be considered to be representatives of the same peptidase.

The new identifiers were set-up to be similar to, but easily distinguishable from, existing MEROPS identifiers. For the new identifiers, the first character after the dot was replaced by the letter 'A', and then the letter 'B' if $>$26 new identifiers were created for the family, and then the letter 'C' if more than 52 identifiers were created.

Assignment of the new identifiers to other sequences

Once the new peptidase identifiers had been established, the program to assign identifiers using the phylogenetic

Table 1. Peptidases from model organisms

Organism	Total	Holotypes	Holotypes with references	Holotypes with known substrate cleavage sites	Holotypes with inhibitor interactions	Uncharacterized holotypes
Homo sapiens	597	462	568	261	174	96
Mus musculus	635	221	151	60	30	100
D. melanogaster	467	407	92	27	17	330
C. elegans	367	324	65	11	3	263
A. thaliana	576	520	127	20	14	398
S. cerevisiae	112	97	83	59	22	19
S. pombe	115	55	17	3	1	39
D. discoideum	174	99	9	3	1	90
E. coli K12 MG1655	106	94	91	56	24	33
B. subtilis 168	205	94	48	23	12	50
P furiosus	70	27	9	7	3	17

There may still be some *B. subtilis* holotypes to identify.
For the purposes of this table, peptidases from retrotransposons (families A2 and A11) are excluded. An uncharacterized holotype is one with no references and no known cleavages or inhibitor interactions. In addition to the holotypes established for peptidases from *E. coli* strain K12 substrain MG1655, some 20 peptidases not found in this strain but in other strains of *E. coli* have been characterized, and holotypes have been set-up for these, for example, colicin V processing peptidase (C39.005), which is encoded on a plasmid.

Table 2. Peptidases common to model organisms

	Source organism for holotype										
	Human	Mouse	*D. melanogaster*	*C. elegans*	*A. thaliana*	*S. cerevisiae*	*S. pombe*	*D. discoideum*	*E. coli*	*B. subtilis*	*P. furiosus*
Human	597	462	51	31	28	15	19	31	1	1	1
Mouse		635	49	28	25	15	19	31	1	1	2
D. melanogaster			467	25	19	11	15	19	1	1	1
C. elegans				367	14	13	15	14	1	2	0
A. thaliana					576	17	25	31	2	4	1
S. cerevisiae						112	27	12	0	1	0
S. pombe							115	14	1	1	1
D. discoideum								174	3	2	1
E. coli									106	14	3
B. subtilis										205	7
P. furiosus											70

This table shows the number of times the same MEROPS identifier has been assigned to sequences from different model organisms. Thus, of the 597 human peptidases, 462 sequences are assumed or known to represent peptidases with similar, if not identical, characteristics in mouse, but only 51 in *Drosophila*.

tree was re-run so that the new identifiers could be assigned to sequences from species other than that of the holotype. Figure 2 shows the consequence of running the program on the phylogenetic tree for subfamily M3A. Table 3 shows the new identifiers that were assigned to $\geq$50 different species.

Table 4 shows the peptidase families for which identifiers for either 20 characterized or 20 putative peptidases have been assigned. For many of these families, the numbers for both are approximately equal (for example: A1, M1, M14, M16 and M20). Family A2 includes peptidases from viruses and retrotransposons, neither of which has been included in this study; hence, there are no identifiers for putative peptidases. The number of identifiers for uncharacterized peptidases in families S63 and T3 is also disproportionally small. There are families for which the number of identifiers for uncharacterized peptidases now exceeds the number for those that have been characterized, especially C26, M13, S9, S10, S12, S28 and S33. The numbers of identifiers for uncharacterized peptidases in S9 and S33 (both 100) are exceeded only by the number in S1 (180).

In total, the new identifiers were applied to 15 149 sequences, which had previously been unassigned, which is ~4.6% of all the peptidase homologues currently in the MEROPS database.

Discussion

Criteria for assigning sequences to a MEROPS identifier

The criteria for limiting the extension of a MEROPS identifier to other members of a family, as described in 'Materials and Methods' section, had to be relaxed in some cases. Peptidase holotypes are assumed to represent gene duplications that precede speciation events, but there are known instances where speciation precedes gene duplication. An example is the renin gene in mouse. Most mammals have only one renin gene, expressed in the kidney, and the enzyme is important for controlling blood pressure by processing the precursor of the peptide hormone angiotensin. But mice have two genes, one expressed in the kidney and one expressed only in the salivary gland, where the enzyme is probably more important as a chemical signal (25). A holotype exists for both renin 1 (the kidney enzyme) and renin 2 (the mouse salivary gland enzyme), and that the classification software should work correctly, the renin 2 holotype is considered a subset of the renin 1 cluster.

The condition that a sequence is not included in a cluster if the sequence length differs from that of the holotype by a hundred residues or more has to be relaxed in the few situations where the holotype sequence has been derived from sequencing of the mature protein and is being compared with the sequences of precursors (and *vice versa*). It was also discovered that many sequences derived from eukaryote genome sequencing projects were considerably longer than many holotype sequences, preventing the MEROPS identifier from being applied fully to other sequences in the tree. Gene building in eukaryote sequencing projects is notoriously difficult because of the problem of correctly identifying introns. Misidentifying introns can lead to incorrect chaining together of proteins derived from adjacent genes, or mistranslation of intron sequence as protein coding, both of which inadvertently increase the length of the protein sequence. To avoid this problem, the length criterion was not applied to sequences derived entirely from eukaryote genome sequencing projects. The

Database, Vol. 2013, Article ID bat022, doi:10.1093/database/bat022

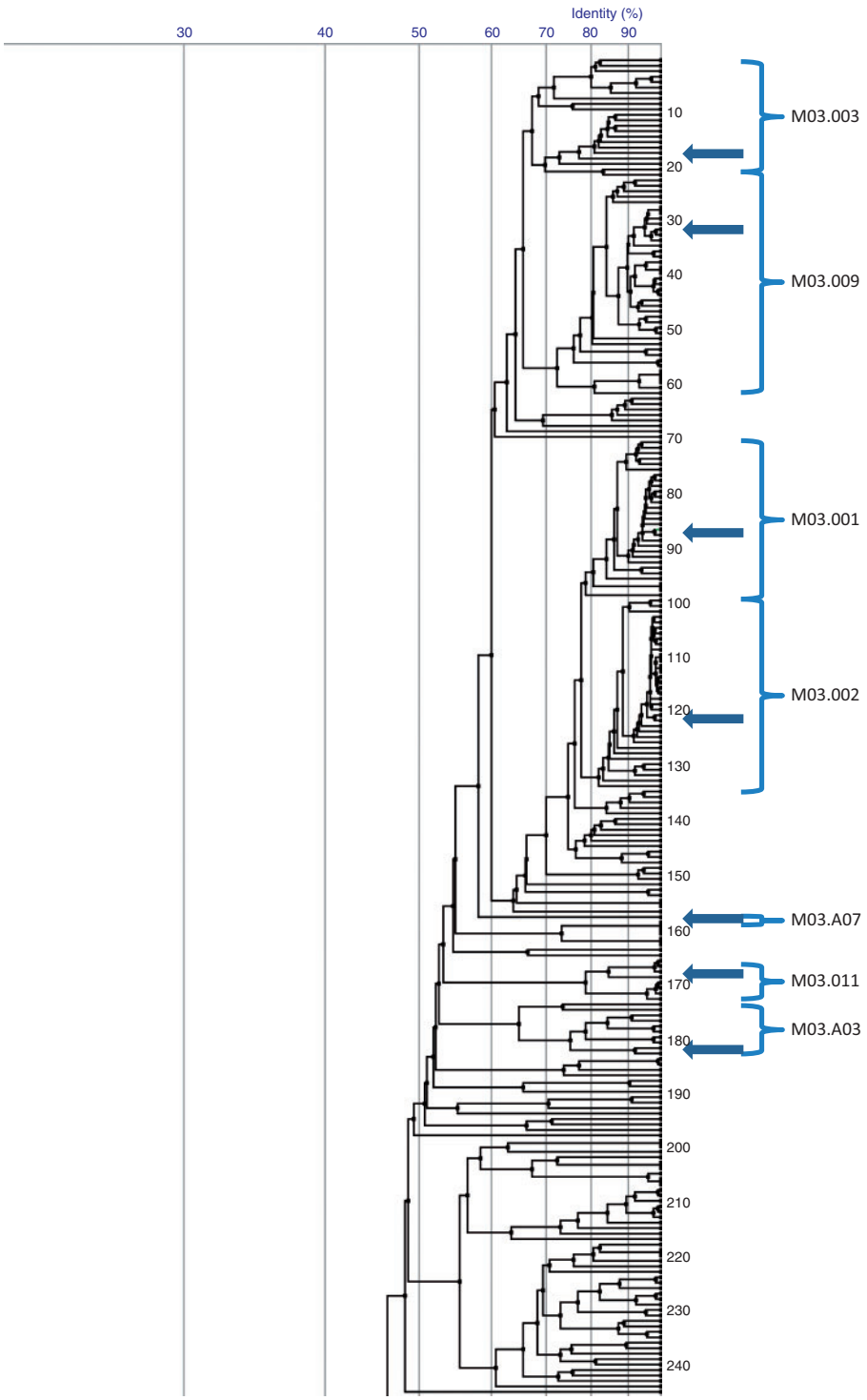

Figure 2. Use of a phylogenetic tree to assign MEROPS identifiers. Part of the tree for family M3 subfamily A is shown. The tips correspond to individual sequences. An arrow indicates the tip corresponding to the sequence that is the holotype for a particular MEROPS identifier. Tips assigned to the same MEROPS identifier are bracketed together, and the MEROPS identifier is shown. Sequences that are not included within a bracket are unassigned. Tips 63–70 are derived from a node ancestral to M03.003 and M03.009; hence, they cannot be assigned to either identifier; similar situations apply to the unassigned tips 135–157, 158–166 and 184–246. The tree shows two identifiers for putative peptidases. M03.A03 was initially assigned to the At1g67690 gene product from *A. thaliana* and has been extended to include sequences from nine other plants. On the other hand, M03.A07, which was originally assigned to the DDB_G0292362 gene product from the slime mould *D. discoideum*, but cannot be extended to sequences from other species because no others are derived from the same node on the tree.

Table 3. New holotypes for putative peptidases with the widest organism distribution

MEROPS identifier	Peptidase name	Number of species	Phyla
C26.A05	γ-Glutamyl peptidase 1 (*A. thaliana*)	78	Bacteria (Bacteroidetes, Proteobacteria, Xenobacteria) Archaea (Euryarchaeota) Fungi (Ascomycota, Basidiomycota, Chytridiomycota) Plantae (Tracheophyta) Animalia (Annelida)
C26.A28	SPBPB2B2.05 (*S pombe*)	297	Bacteria (Acidobacteria, Bacteroidetes, Chlamydiae, Chloroflexi, Chrysiogenetes, Cyanobacteria, Firmicutes, Fusobacteria, Gemmatimonadetes, Nitrospirae, Planctomycetes, Proteobacteria, Spirochaetes, Thermotogae, Xenobacteria) Fungi (Ascomycota) Plantae (Tracheophyta)
C26.A31	GMP synthase (*P. furiosus*)	72	Bacteria (Chlamydiae) Archaea (Crenarchaeota, Euryarchaeota)
C26.A32	Imidazoleglycerol-phosphate synthase (*P. furiosus*)	176	Bacteria (Acidobacteria, Bacteroidetes, Chloroflexi, Dictyoglomi, Fibrobacteres, Firmicutes, Fusobacteria, Proteobacteria, Synergistetes, Thermotogae, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota, Korarchaeota) Fungi (Ascomycota) Plantae (Tracheophyta) Animalia (Chordata)
C40.A01	NlpC protein (*E. coli*)	60	Bacteria (Proteobacteria) Animalia (Porifera)
C44.A08	Glutamine-fructose-6-phosphate transaminase precursor (*D. discoideum*)	74	Bacteria (Bacteroidetes, Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes, Proteobacteria, Verrucomicrobia, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota) Protozoa (Alveolata, Apicomplexa, Microspora, Sarcomastigophora) Plantae (Heterokontophyta, Ochrophyta) Animalia (Arthropoda, Chordata)
C56.A06	DDB_G0276405 (*D. discoideum*)	350	Bacteria (Acidobacteria, Bacteroidetes, Chloroflexi, Cyanobacteria, Deferribacteres, Firmicutes, Ignavibacteria, Lentisphaerae, Planctomycetes, Proteobacteria, Spirochaetes, Thermodesulfobacteria, Verrucomicrobia, Xenobacteria) Archaea (Euryarchaeota) Protozoa (Sarcomastigophora) Fungi (Ascomycota) Plantae (Chlorophyta) Animalia (Arthropoda, Nematoda)
C82.A01	YafK protein (*E. coli*)	56	Bacteria (Bacteroidetes, Proteobacteria)
C82.A07	Murein transglycosylase (*B. subtilis*)	74	Bacteria (Cyanobacteria, Firmicutes, Proteobacteria)
M03.A08	BSSC8_09230 protein (*B. subtilis*)	234	Bacteria (Chlamydiae, Chloroflexi, Cyanobacteria, Firmicutes, Planctomycetes, Proteobacteria, Spirochaetes, Thermotogae, Xenobacteria) Plantae (Tracheophyta)

(Continued)

Table 3. Continued

MEROPS identifier	Peptidase name	Number of species	Phyla
M13.A32	T25B6.2 protein (*C. elegans*)	140	Bacteria (Bacteroidetes, Firmicutes, Proteobacteria) Fungi (Ascomycota, Basidiomycota) Plantae (Chlorophyta, Rhodophyta, Tracheophyta) Animalia (Arthropoda, Chordata, Echinodermata, Hemichordata, Nematoda)
M15.A04	YokZ protein (*B. subtilis*)	274	Bacteria (Bacteroidetes, Chloroflexi, Cyanobacteria, Firmicutes, Spirochaetes, Xenobacteria)
M16.A04	Insulysin homologue (*S. cerevisiae*)	73	Fungi (Ascomycota, Basidiomycota)
M16.A05	PqqL protein (*E. coli*)	103	Bacteria (Bacteroidetes, Chlorobi, Deferribacteres, Fusobacteria, Gemmatimonadetes, Proteobacteria, Spirochaetes, Xenobacteria)
M20.A08	YgeY protein (*E. coli*)	61	Bacteria (Bacteroidetes, Caldiserica, Chloroflexi, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes, Synergistetes, Thermotogae)
M20.A18	DDB_G0279291 protein (*D. discoideum*)	72	Bacteria (Firmicutes, Proteobacteria) Protozoa (Alveolata, Parabasalidea, Sarcomastigophora) Plantae (Tracheophyta)
M20.A21	YodQ protein (*B. subtilis*)	81	Bacteria (Chloroflexi, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes) Archaea (Euryarchaeota)
M20.A23	Bsubs1_010100013116 protein (*B. subtilis*)	52	Bacteria (Firmicutes, Fusobacteria, Planctomycetes, Proteobacteria, Synergistetes)
M20.A27	YkuR protein (*B. subtilis*)	82	Bacteria (Firmicutes, Fusobacteria, Thermotogae)
M24.A09	YFR006W protein (*S. cerevisiae*)	57	Fungi (Ascomycota, Basidiomycota) Plantae (Chlorophyta)
M24.A11	SPBC4F6.19c protein (*S. pombe*)	77	Bacteria (Acidobacteria, Bacteroidetes, Proteobacteria) Fungi (Ascomycota, Basidiomycota) Plantae (Tracheophyta)
M48.A02	At3g27110 (*A. thaliana*)	50	Bacteria (Cyanobacteria, Firmicutes, Proteobacteria, Spirochaetes) Archaea (Euryarchaeota) Plantae (Tracheophyta)
M50.A04	PF0392 protein (*P. furiosus*)	79	Bacteria (Acidobacteria, Bacteroidetes, Chlorobi, Chloroflexi, Cyanobacteria, Nitrospirae, Planctomycetes, Proteobacteria) Archaea (Crenarchaeota, Euryarchaeota)
M50.A05	YwhC protein (*B. subtilis*)	203	Bacteria (Acidobacteria, Aquificae, Chloroflexi, Chrysiogenetes, Deferribacteres, Dictyoglomi, Firmicutes, Fusobacteria, Nitrospirae, Proteobacteria, Spirochaetes, Synergistetes, Thermodesulfobacteria, Verrucomicrobia, Xenobacteria) Archaea (Nanoarchaeota)

(Continued)

Table 3. Continued

MEROPS identifier	Peptidase name	Number of species	Phyla
M50.A07	PF0457 protein (*P. furiosus*)	76	Bacteria (Bacteroidetes, Chloroflexi, Cyanobacteria, Firmicutes, Gemmatimonadetes, Proteobacteria, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota) Plantae (Chlorophyta)
M79.A04	At3g26085 (*A. thaliana*)	58	Bacteria (Chloroflexi, Cyanobacteria, Firmicutes, Proteobacteria, Xenobacteria) Archaea (Euryarchaeota) Plantae (Tracheophyta)
M79.A09	YdiL protein (*B. subtilis*)	140	Bacteria (Bacteroidetes, Chlamydiae, Chloroflexi, Cyanobacteria, Deferribacteres, Firmicutes, Fusobacteria, Proteobacteria, Spirochaetes, Verrucomicrobia) Archaea (Crenarchaeota, Euryarchaeota) Protozoa (Sarcomastigophora)
N06.A01	FlhB protein (*E. coli*)	450	Bacteria (Acidobacteria, Aquificae, Chloroflexi, Deferribacteres, Firmicutes, Proteobacteria, Spirochaetes, Synergistetes, Thermotogae)
S01.A08	At5g27660 (*A. thaliana*)	58	Bacteria (Acidobacteria, Bacteroidetes, Caldiserica, Chlamydiae, Chloroflexi, Firmicutes, Fusobacteria, Ignavibacteria, Planctomycetes, Proteobacteria, Spirochaetes, Thermotogae, Xenobacteria) Archaea (Crenarchaeota) Plantae (Heterokontophyta, Tracheophyta) Animalia (Chordata)
S09.A43	YpfH protein (*E. coli*)	101	Bacteria (Acidobacteria, Chlamydiae, Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes, Proteobacteria) Fungi (Ascomycota)
S09.A77	dpf-6 protein (*C. elegans*)	334	Bacteria (Acidobacteria, Bacteroidetes, Chlamydiae, Chloroflexi, Cyanobacteria, Firmicutes, Gemmatimonadetes, Nitrospirae, Planctomycetes, Proteobacteria, Synergistetes, Thermotogae, Verrucomicrobia, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota) Protozoa (Sarcomastigophora) Fungi (Ascomycota) Plantae (Ochrophyta, Tracheophyta) Animalia (Chordata, Nematoda)
S09.B04	BSU23640 protein (*B. subtilis*)	56	Bacteria (Acidobacteria, Bacteroidetes, Chloroflexi, Firmicutes, Planctomycetes, Proteobacteria, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota)
S10.A47	At2g22960 (*A. thaliana*)	122	Protozoa (Alveolata, Apicomplexa, Sarcomastigophora) Fungi (Ascomycota, Basidiomycota) Plantae (Chlorophyta, Ochrophyta, Oomycota, Rhodophyta, Streptophyta, Tracheophyta) Animalia (Arthropoda, Chordata, Echinodermata, Hemichordata, Nematoda)
S12.A03	EcHS_A2566 protein (*E. coli*)	180	Bacteria (Bacteroidetes, Chlorobi, Chloroflexi, Cyanobacteria, Firmicutes, Fusobacteria, Nitrospirae, Planctomycetes, Proteobacteria, Spirochaetes, Thermotogae, Xenobacteria) Archaea (Euryarchaeota) Fungi (Ascomycota)

(Continued)

Table 3. Continued

MEROPS identifier	Peptidase name	Number of species	Phyla
S12.A23	GYO_0385 protein (*B. subtilis*)	51	Bacteria (Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes, Xenobacteria) Archaea (Euryarchaeota)
S16.A10	YcbZ (*E. coli*)	110	Bacteria (Firmicutes, Proteobacteria, Spirochaetes) Animalia (Chordata)
S16.A12	PF1438 (*P. furiosus*)	270	Bacteria (Firmicutes, Proteobacteria) Archaea (Crenarchaeota, Euryarchaeota, Korarchaeota) Protozoa (Apicomplexa)
S49.A08	BSn5_05605 protein (*B. subtilis*)	189	Bacteria (Chrysiogenetes, Cyanobacteria, Deferribacteres, Firmicutes, Proteobacteria, Spirochaetes, Thermodesulfobacteria, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota)
S49.A09	PF0240protein (*P. furiosus*)	67	Bacteria (Aquificae, Bacteroidetes, Chloroflexi, Cyanobacteria, Deferribacteres, Dictyoglomi, Firmicutes, Nitrospirae, Proteobacteria, Synergistetes, Thermodesulfobacteria, Thermotogae, Xenobacteria) Archaea (Crenarchaeota, Euryarchaeota)
S54.A18	YdcA protein (*B. subtilis*)	54	Bacteria (Firmicutes, Proteobacteria)
T03.A09	GYO_3966 protein (*B. subtilis*)	761	Bacteria (Acidobacteria, Bacteroidetes, Chloroflexi, Cyanobacteria, Firmicutes, Fusobacteria, Lentisphaerae, Planctomycetes, Proteobacteria, Xenobacteria) Archaea (Euryarchaeota) Protozoa (Sarcomastigophora) Fungi (Ascomycota, Basidiomycota, Oomycota) Plantae (Chlorophyta, Heterokontophyta, Ochrophyta, Tracheophyta) Animalia (Annelida, Arthropoda, Chordata, Cnidaria, Placozoa, Porifera)
U32.A01	YhbV protein (*E. coli*)	131	Bacteria (Proteobacteria)

Columns are the MEROPS identifier (linked to the peptidase summary page in the MEROPS database); the name used in the MEROPS database, which is often derived from the gene name; the number of species containing a sequence to which the identifier has been assigned; and the phyla that includes these species, grouped by kingdom. When one of these putative peptidases is characterized, the MEROPS identifier will be replaced with an identifier that follows the normal naming convention used in the MEROPS database. The obsolete identifier will not be re-used, and so that the links below will remain useful the user will be automatically redirected to the summary page for the replacement identifier.

Table 4. Totals of identifiers for characterized and uncharacterized peptidases

Family	Type example	Characterized peptidases	Putative peptidases
A1	Pepsin	84	92
A2	Retropepsin	32	0
C1	Papain	148	59
C2	Calpain	25	10
C19	Ubiquitin-specific peptidase 14	94	66
C26	γ-Glutamyl hydrolase	1	20
C48	Ulp1 peptidase	32	19
M1	Aminopeptidase N	31	31
M12	Astacin	180	51
M13	Neprilysin	17	32
M14	Carboxypeptidase A1	34	34
M16	Pitrilysin	17	20
M20	Glutamate carboxypeptidase	20	21
M41	FtsH peptidase	22	14
S1	Chymotrypsin	462	180
S8	Subtilisin	147	58
S9	Prolyl oligopeptidase	45	100
S10	Carboxypeptidase Y	17	69
S12	D-Ala-D-Ala carboxypeptidase B	10	21
S26	Signal peptidase I	27	10
S28	Lysosomal Pro-Xaa carboxypeptidase	5	26
S33	Prolyl aminopeptidase	18	100
S54	Rhomboid-1	28	18
S63	EGF-like module containing mucin-like hormone receptor-like 2	35	2
T3	γ-Glutamyltransferase 1	21	9

The total number of identifiers for characterized and uncharacterized peptidases is shown for all families where there are ≥20 examples in either category. The name of the type example peptidase is given for each family.

length criterion was introduced so that the domain architecture would be more or less the same for all sequences assigned to the same identifier because sequences with similar domain architectures are unlikely to have different exosites. Relaxing the length criterion for eukaryotes until automated gene assembly improves does carry the risk of assigning sequences with genuinely different domain architectures, and possibly different exosites, to the same identifier.

Species-distribution of new peptidase identifiers

The species-distribution of a peptidase varies greatly, from just one known species, such as renin-2, which is known only from the mouse, to many species from all major groups of organic life. The most widely distributed peptidases are peptidase Clp (present in organisms from 51 of the ~100 known phyla), methionyl aminopeptidase 1 (48 phyla) and the self-processing ornithine acetyltransferase precursor (42 phyla). T03.A09 is the new identifier assigned to the sequences from the most species. This is a homologue of γ-glutamyltransferase, which degrades glutathione (γ-Glu-Cys-Gly) by cleavage of the γ-glutamyl bond. T03.A09 is present in species from all six major kingdoms of organisms (bacteria, archaea, protozoa, fungi, plants and animals), and it is widely distributed in bacteria, plants and animals. Despite having not been characterized, crystal structures have been deposited for T03.A09 from *Bacillus halodurans* (PDB: 2NLZ) and *Thermoplasma acidophilum* (PDB: 2I3O), but neither has been published. T03.A09 from *B. halodurans* is described as 'cephalosporin acylase' (cephalosporin is a β-lactam–based antibiotic and a bacteriocide), whereas in *T. acidophilum,* it is described as a 'γ-glutamyltransferase–related protein'.

More genomes from bacteria have been sequenced than for any other kingdom of organisms; thus, it is no surprise that most of the identifiers listed in Table 3 are from bacteria. The identifiers with the widest distribution among bacterial phyla are C26.A28 (15), C56.A06 (14), S09.A77

(14), S01.A08 (13), S12.A03 (12), C26.A32 (11), M79.A09 (10) and T03.A09 (10). The few identifiers that are exclusively bacterial are found only in species from a few phyla. An exception is N06.A01, which is found in species from nine bacterial phyla. Members of family N6 are not peptidases, but self-processing proteins that use asparagine as a nucleophile and have been described as 'asparagine peptide lyases' (26). Members of family N6 are known to be important for type III secretion by which virulence factors are injected into host cells through a flagellum-like filament (27).

There are three other identifiers found only in bacteria for which examples are known from over a hundred species. These are M15.A04, M16.A05 and U32.A01. Family M15 includes predominantly D-Ala-D-Ala carboxypeptidases, important for processing of the precursor of the cross-linking peptide of a bacterial cell wall (28). The cross-linking peptide precursor in both *B. subtilis* (from which the holotype is derived) and *E. coli* is L-Ala-D-Glu-L-*meso*-diaminopimelic acid-D-Ala-D-Ala (29), so presumably the M15.A04 peptidase is not involved in processing of this precursor because this peptidase is not present in *E. coli*. Peptidase family M16 includes the inverzincins, metallopeptidases in which the zinc-binding residues are the histidines in the motif HXXEH, the reverse of the motif found in thermolysin (30), and includes the oligopeptidases pitrilysin, which is periplasmic in *E. coli*, and insulysin, which is cytoplasmic in vertebrates. Insulysin consists of two homologous domains, of which only the N-terminal one bears the active site (31). The PqqL protein (M16.A05), which was sequenced independently of the genome sequencing project (32) but never characterized, is similarly composed of two homologous domains. Unlike pitrilysin, the PqqL protein does not have an N-terminal signal peptide and is presumably cytosolic. Peptidases in family U32 are of unknown catalytic type, and the only characterized peptidases have been shown to degrade soluble collagen, such as the PrtC protein from *Porphyromonas gingivalis* (33). U32.A01 is restricted entirely to Proteobacteria. However, not only is U32.A01 the only identifier in Table 3 for a peptidase of unknown catalytic type, it is also the only one for which active site residues are unknown. Therefore, it is impossible to tell whether this protein represents an active peptidase or a non-peptidase homologue.

Table 3 lists three identifiers that are not found in bacteria: M16.A04 (found in 73 species), M24.A09 (found in 57) and S10.A47 (found in 122). M16.A04, which is restricted to fungi, presumably represents a novel oligopeptidase and M24.A09, found only in fungi and green algae, may be a novel exopeptidase because family M24 contains aminopeptidases (such as methionyl aminopeptidase), dipeptidases (such as Xaa-Pro dipeptidase) and aminopeptidases (such as aminopeptidase P) (34, 35). Family S10 contains carboxypeptidases (35), and S10.A47 is presumably also a carboxypeptidase. An example of S10.A47 was partially sequenced, but never characterized, from barley and was shown to be expressed in the aleurone layer of the seed (36).

Examination of the distribution of source organisms among the sequences assigned to a particular MEROPS identifier can highlight the problem of contamination when the genome was being sequenced, or reveal possible examples of horizontal gene transfer if the unusual distribution is consistent. Examples of possible contamination include M03.A08 from the moss *Physcomitrella patens* (the identifier has otherwise only been assigned to sequences from bacteria); M79.A04 from the plants *A. thaliana* and poplar (otherwise the identifier has been assigned only to sequences from bacteria and archaea); M79.A09 from the protozoan *Naegleria gruberi* (otherwise the identifier has been assigned exclusively to prokaryote sequences); S09.A43 from the yeast *Gibberella zeae* (otherwise the identifier has been assigned only to sequences from bacteria); and S16.A10 from the armadillo (otherwise the identifier has been assigned only to sequences from bacteria). The presence of M48.A02 in *A. thaliana* and rice also probably represents bacterial contamination, but there is also a sequence from the uncultured methanogenic archaeon RC-I, which could be an example of a horizontal gene transfer. M50.A07 is again mostly prokaryotic apart from three sequences from three different green algae species, which may represent a single horizontal gene transfer to an ancestral alga, because the alga sequences cluster together on the phylogenetic tree, from a bacterium.

Non-peptidase homologues

It should be stressed that not all the identifiers listed in Table 3 may relate to peptidases. Almost all families have non-peptidase homologues where one or more active site residues are missing or have been replaced. However, there are members of some families in which the active site residues are conserved but are known to perform reactions other than proteolysis. Examples include aminoacylase and acetylornithine deacetylase (from family M20), ureases (family M38), esterases (family S9) and lipases (family S33). Succinyl-diaminopimelate desuccinylase (family M20) was initially characterized as a non-peptidase homologue, but it was then shown to possess proteolytic activity (37). Leukotriene A4 hydrolase is an example of a protein with dual enzymatic activity, including that of a peptidase (38). Table 3 lists five members of family C26, which includes not only the mammalian peptidase γ-glutamyl hydrolase but is also known to include several activities other than peptidases, including guanine 5′-monophosphate (GMP), carbamyl phosphate and aminodeoxychorismate synthases. The holotypes for C26.A31 and C26.A32, both from *P. furiosus*, are described as a GMP synthase and an imidazoleglycerol-phosphate synthase, respectively, even though neither has

been biochemically characterized. It remains likely that some or even all of the C26 homologues listed in Table 3 are not peptidases.

Characterization

Although none of the putative peptidases listed in Table 3 has been biochemically characterized, some have been subjects of study, such as PqqL (M16.A05, see earlier in the text). Additionally, C26.A05 has been implicated, along with other members of the γ-glutamyl hydrolase family, in the production of glucosinolates, which are defence-related plant metabolites using glutathione as a source of sulphur (39, 40), although the pathway has not been established. Tertiary structures have been solved for some of the putative peptidases listed in Table 3 by consortia who are trying to solve structures for all protein folds without necessarily knowing the functions of the proteins being solved. Structures have been solved for M32.A01 from *B. subtilis* [PDB: 3HQ2; (41)] and M13.A32 (see earlier in the text).

The identifiers for these putative peptidases will be established in the next release of the MEROPS database (release 9.8, due December 2012). When one of these putative peptidases becomes characterized, the identifier will be replaced with a standard MEROPS identifier. The obsolete identifier will not be re-used, and an automatic redirection to the new identifier will be set up in the MEROPS database.

Conclusions

Methods have been established for distinguishing a cluster of sequences within a peptidase family that represents species variants of a common peptidase. The methods involve detection of homologues by a sequence similarity search using that part of the sequence bearing the catalytic residues from a designated type example, assembly of an alignment, generation of a phylogenetic tree from the alignment, followed by computerized assessment of the tree to identify clusters of related sequences. By using these methods and the extensive collections of references, substrate cleavages and inhibitor interactions in the MEROPS database, it has been possible to distinguish a cluster of sequences representing a biochemically characterized peptidase from a cluster that represents a putative peptidase that has not been characterized. By assigning identifiers to every peptidase homologue from the complete genome sequences of 11 model organisms, both those that are characterized and those that are not, the methods have been applied to detect a cluster of sequences that includes each of the putative peptidases from these model organisms. The clusters of putative peptidases with the widest organism distribution are presented in this article as ideal targets for biochemical characterization. Data derived from experimentation should then be transferable to all members of the cluster.

Versions of these methods should be applicable to any collection of sequence information where uncharacterized proteins or genes need to be distinguished from those that are characterized and should be especially applicable to enzymes that act on biological polymers.

By combining these methods with an analysis of the taxonomic distribution within the cluster, it is possible to identify potential contaminants in genome sequencing projects and examples of possible horizontal gene transfer.

The new identifiers for putative peptidases will be included in release 9.8 of the MEROPS database.

Acknowledgements

The author thanks Dr Alan Barrett who maintains the MEROPS reference collection, Dr Alex Bateman for support and guidance, members of the Wellcome Trust Sanger Institute web team, especially Matthew Waller and Paul Bevan, for maintaining the MEROPS website and Pfam, Rfam and Treefam colleagues for helpful discussions.

Funding

This work and funding for open access publication was supported by the Wellcome Trust (WT098051).

Conflict of interest. None declared.

References

1. NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology) (1992) *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, Orlando.
2. Roberts,R.J., Vincze,T., Posfai,J. *et al*. (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
3. Cantarel,B.L., Coutinho,P.M., Rancurel,C. *et al*. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.*, **37**, D233–D238.
4. Messaoudi,A., Belguith,H., Ghram,I. *et al*. (2011) LIPABASE: a database for 'true' lipase family enzymes. *Int. J. Bioinform. Res. Appl.*, **7**, 390–401.
5. Manning,G., Plowman,G.D., Hunter,T. *et al*. (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.*, **27**, 514–520.
6. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, **40**, D343–D350.
7. Biniossek,M.L., Nagler,D.K., Becker-Pauly,C. *et al*. (2011) Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S. *J. Proteome Res.*, **10**, 5363–5373.

8. Rawlings,N.D. and Barrett,A.J. (1993) Evolutionary families of peptidases. *Biochem. J.*, **290**, 205–218.
9. Rawlings,N.D. and Barrett,A.J. (1999) *MEROPS*: the peptidase database. *Nucleic Acids Res.*, **27**, 325–331.
10. Barrett,A.J. and Rawlings,N.D. (2007) 'Species' of peptidases. *Biol. Chem.*, **388**, 1151–1157.
11. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2010) MEROPS: the peptidase database. *Nucleic Acids Res.*, **38**, D227–D233.
12. Rawlings,N.D. (2009) A large and accurate collection of peptidase cleavages in the MEROPS database. *Database*, **2009**, bap015.
13. Rawlings,N.D., Morton,F.R., Kok,C.Y. *et al*. (2008) MEROPS: the peptidase database. *Nucleic Acids Res.*, **36**, D320–D325.
14. Rawlings,N.D., Morton,F.R. and Barrett,A.J (2007) An introduction to peptidases and the MEROPS database. In: Polaina,J. and MacCabe,A.P. (eds), *Industrial Enzymes*. Springer, Dordrecht, pp. 161–179.
15. Bateman,A., Coggill,P. and Finn,R.D. (2010) DUFs: families in search of function. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **66**, 1148–1152.
16. Altschul,S.F., Gish,W., Miller,W. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. Sayers,E.W., Barrett,T., Benson,D.A. *et al*. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
18. Rawlings,N.D. and Morton,F.R. (2008) The MEROPS batch BLAST: a tool to detect peptidases and their non-peptidase homologues in a genome. *Biochimie*, **90**, 243–259.
19. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
20. George,D.G., Barker,W.C. and Hunt,L.T. (1990) Mutation data matrix and its uses. *Methods Enzymol.*, **183**, 333–351.
21. Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.*, **38**, 1409–1438.
22. Howe,K., Bateman,A. and Durbin,R. (2002) QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.
23. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
24. Finn,R.D., Mistry,J., Tate,J. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
25. Suzuki,F., Murakami,K., Nakamura,Y. *et al*. (1998) Mouse submandibular renin. In: Barrett,A.J., Rawlings,N.D. and Woessner,J.F. (eds), *Handbook of Proteolytic Enzymes*. Academic Press, London, pp. 856–858.
26. Rawlings,N.D., Barrett,A.J. and Bateman,A. (2011) Asparagine peptide lyases: a seventh catalytic type of proteolytic enzymes. *J. Biol. Chem.*, **286**, 38321–38328.
27. Edqvist,P.J., Olsson,J., Lavander,M. *et al*. (2003) YscP and YscU regulate substrate specificity of the *Yersinia* type III secretion system. *J. Bacteriol.*, **185**, 2259–2266.
28. Bochtler,M., Odintsov,S.G., Marcyjaniak,M. *et al*. (2004) Similar active sites in lysostaphins and D-Ala-D-Ala metallopeptidases. *Protein Sci.*, **13**, 854–861.
29. Atrih,A., Bacher,G., Allmaier,G. *et al*. (1999) Analysis of peptidoglycan structure from vegetative cells of *Bacillus subtilis* 168 and role of PBP 5 in peptidoglycan maturation. *J. Bacteriol.*, **181**, 3956–3966.
30. Hooper,N.M. (1994) Families of zinc metalloproteases. *FEBS Lett.*, **354**, 1–6.
31. Shen,Y., Joachimiak,A., Rosner,M.R. *et al*. (2006) Structures of human insulin-degrading enzyme reveal a new substrate recognition mechanism. *Nature*, **443**, 870–874.
32. Turlin,E., Gasser,F. and Biville,F. (1996) Sequence and functional analysis of an *Escherichia coli* DNA fragment able to complement pqqE and pqqF mutants from *Methylobacterium organophilum*. *Biochimie*, **78**, 823–831.
33. Kato,T., Takahashi,N. and Kuramitsu,H.K. (1992) Sequence analysis and characterization of the *Porphyromonas gingivalis prtC* gene, which expresses a novel collagenase activity. *J. Bacteriol.*, **174**, 3889–3895.
34. Bazan,J.F., Weaver,L.H., Roderick,S.L. *et al*. (1994) Sequence and structure comparison suggest that methionine aminopeptidase, prolidase, aminopeptidase P, and creatinase share a common fold. *Proc. Natl Acad. Sci. USA*, **91**, 2473–2477.
35. Remington,S.J. (1993) Serine carboxypeptidases: a new and versatile family of enzymes. *Curr. Opin. Biotechnol.*, **4**, 462–468.
36. Dal Degan,F., Rocher,A., Cameron-Mills,V. *et al*. (1994) The expression of serine carboxypeptidases during maturation and germination of the barley grain. *Proc. Natl Acad. Sci. USA*, **91**, 8209–8213.
37. Broder,D.H. and Miller,C.G. (2003) DapE can function as an aspartyl peptidase in the presence of Mn2+. *J. Bacteriol.*, **185**, 4748–4754.
38. Griffin,K.J., Gierse,J., Krivi,G. *et al*. (1992) Opioid peptides are substrates for the bifunctional enzyme LTA_4 hydrolase/aminopeptidase. *Prostaglandins*, **44**, 251–257.
39. Geu-Flores,F., Moldrup,M.E., Bottcher,C. *et al*. (2011) Cytosolic gamma-glutamyl peptidases process glutathione conjugates in the biosynthesis of glucosinolates and camalexin in *Arabidopsis*. *Plant Cell*, **23**, 2456–2469.
40. Moldrup,M.E., Geu-Flores,F., Olsen,C.E. *et al*. (2011) Modulation of sulfur metabolism enables efficient glucosinolate engineering. *BMC Biotechnol.*, **11**, 12.
41. Lee,M.M., Isaza,C.E., White,J.D. *et al*. (2009) Insight into the substrate length restriction of M32 carboxypeptidases: characterization of two distinct subfamilies. *Proteins*, **77**, 647–657.