Biochimie 122 (2016) 5–30

Contents lists available at ScienceDirect

Biochimie

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

Research paper

Peptidase specificity from the substrate cleavage collection in the *MEROPS* database and a tool to measure cleavage site conservation

Neil D. Rawlings

Wellcome Trust Sanger Institute and the EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK

ARTICLE INFO

Article history:
Received 26 June 2015
Accepted 5 October 2015
Available online 21 October 2015

Keywords:
Peptidase
Substrate
Specificity
Cleavage
Binding pocket
Scissile bond

ABSTRACT

One peptidase can usually be distinguished from another biochemically by its action on proteins, peptides and synthetic substrates. Since 1996, the *MEROPS* database (http://merops.sanger.ac.uk) has accumulated a collection of cleavages in substrates that now amounts to 66,615 cleavages. The total number of peptidases for which at least one cleavage is known is 1700 out of a total of 2457 different peptidases. This paper describes how the cleavages are obtained from the scientific literature, how they are annotated and how cleavages in peptides and proteins are cross-referenced to entries in the UniProt protein sequence database. The specificity profiles of 556 peptidases are shown for which ten or more substrate cleavages are known. However, it has been proposed that at least 40 cleavages in disparate proteins are required for specificity analysis to be meaningful, and only 163 peptidases (6.6%) fulfil this criterion. Also described are the various displays shown on the website to aid with the understanding of peptidase specificity, which are derived from the substrate cleavage collection. These displays include a logo, distribution matrix, and tables to summarize which amino acids or groups of amino acids are acceptable (or not acceptable) in each substrate binding pocket. For each protein substrate, there is a display to show how it is processed and degraded. Also described are tools on the website to help with the assessment of the physiological relevance of cleavages in a substrate. These tools rely on the hypothesis that a cleavage site that is conserved in orthologues is likely to be physiologically relevant, and alignments of substrate protein sequences are made utilizing the UniRef50 database, in which in each entry sequences are 50% or more identical. Conservation in this case means substitutions are permitted only if the amino acid is known to occupy the same substrate binding pocket from at least one other substrate cleaved by the same peptidase.

1. Introduction

In 2007, Barrett & Rawlings [1] proposed a list of criteria to distinguish one peptidase from another. To be considered different, any one of the following bioinformatics tests can be applied: the two peptidases have similar biochemical characteristics but unrelated sequences; the two peptidases have related sequences but different biochemical properties, different domain architectures or the domains are in a different order; or the two peptidases have greater than 50% sequence identity but are derived from nodes on a phylogenetic tree that are not adjacent. In addition, the following biochemical tests can be applied to distinguish two peptidases: the peptidases act under significantly different conditions; the peptidases have different post-translational modifications; the peptidases are sensitive to different inhibitors; the peptidases act on different substrates, or if they act on the same substrates then the cleavage positions are different. It is the last two criteria with which this paper is concerned.

A peptidase cleaves a substrate at the scissile bond, and substrate residues either side of this bond are known as P1 and P1′. Residues towards the N-terminus of the substrate are on the non-prime side, and are numbered P1, P2, P3, P4 and so on. Residues towards the C-terminus are on the prime side and are numbered P1′, P2′, P3′, P4′ and so on. A substrate binding pocket in the peptidase that accommodates a substrate residue is named according to the position the residue occupies in the substrate, except that the “P” is replaced by an “S”. So the S1 binding pocket accommodates the P1 residue, and the S4′ binding pocket accommodates the P4′ residue [2].

A collection of substrate cleavages has been assembled from the scientific literature, annotated, cross-referenced where applicable

E-mail addresses: ndr@ebi.ac.uk, ndr@sanger.ac.uk.

http://dx.doi.org/10.1016/j.biochi.2015.10.003

to the UniProt protein sequence database, and included within the MEROPS database. This collection was originally derived from the CD-ROM version of the first edition of the *Handbook of Proteolytic Enzymes* (1998) [3], which also included a search facility to find the peptidases able to cleave a substrate at a particular position. By knowing where in proteins, peptides or synthetic substrates cleavages occur, it is possible to postulate the specificity of a peptidase. By knowing which amino acids can occupy each substrate binding position, it is also possible to infer whether or not cleavage of a substrate at a particular position is likely to be physiologically relevant from an alignment of protein sequences of closely-related orthologues.

The MEROPS substrate cleavage collection has been widely used to predict cleavages in substrates (for a review see Song et al. (2011) [4]), and to predict what peptidase may be responsible for a known cleavage, for example PROSPER [5]. The MEROPS collection has also been used for the mapping of the human degradome and prediction of "cleavage entropy" as an overall measure of peptidase specificity [6], as well as in the development of the "protease web", the network of peptidase, substrate and inhibitor interactions [7].

This paper describes the MEROPS substrate cleavage collection and the various displays present on the MEROPS website (http://merops.sanger.ac.uk) which aid in understanding peptidase specificity and the processing and degradation of a protein substrate. In order to help determine whether or not a cleavage is physiologically relevant, a service is also described where a user can upload substrate cleavages and receive by E-mail an analysis to show how well conserved, in terms of peptidase binding, each cleavage is.

2. Materials and methods

2.1. Identification of peptidases, homology searching, sequence alignment and phylogenetic tree generation

A peptidase species was defined according to the principles established in Barrett & Rawlings (2007) [1]. The methods for homology searching, family building, and generation of protein sequence alignments and phylogenetic trees are the same as those described in Rawlings et al. (2014) [8]. In brief, the following methods were used. Only the peptidase domain was used for sequence searching and sequence alignment. For each family a *type example* was chosen and for each peptidase species a *holotype* was chosen. The type example and holotype were usually the sequence of the best characterized peptidase in the family or protein species, respectively. A BlastP search [9] of the NCBI non-redundant protein sequence database was performed, using the family type example sequence. Sequences retrieved with an E value of 0.01 or less were considered homologues and included in the family. To find more distant homologues, a HMMER search [10] was performed using a ClustalW alignment [11] of a selection of sequences from the family that included an example from every phylum for which there was a representative. Sequence alignments were built using MAFFT [12]. Phylogenetic trees were built from the family sequence alignment using QuickTree [13].

2.2. Manual substrate cleavage curation

The scientific literature was searched manually for substrate cleavage sites by peptidases. Data were acquired from over 7280 references. The following data were collected, transformed as required and stored in a MySQL database. From the name of the peptidase as given by the authors of the publication, a *MEROPS* identifier and, if possible, a MERNUM indicating the source organism, were assigned. From the name of the substrate and its source, a UniProt accession was assigned where possible, and the name recommended by UniProt was stored in the MySQL database, unless the substrate was a peptide or was processed, in which case a peptide name or a name to indicate that processing had occurred was stored (for example, "Met-enkephalin" would be stored in preference to "pro-opiomelanocortin" if the substrate was just the peptide). Where more than one UniProt entry existed, the annotated SwissProt accession, name and sequence where used in preference. Where isoforms derived from alternative initiation and alternative splicing were indicated in the UniProt database entry, the sequence chosen as the representative sequence by UniProt was selected unless the original publication indicated that a particular isoform had been used. There was no attempt to map a cleavage to all isoforms on the presumption that a change in sequence could lead to a change in cleavage position. The cleavage position (the position of the P1 residue in the substrate) was converted to the equivalent residue number from the respective UniProt entry. Up to four residues either side of the scissile bond (residues P4 to P4′) were stored for each cleavage. The residue range of the substrate used compared to the sequence in the UniProt entry was also stored. This allowed for annotation of peptide substrates derived from full-length proteins and processing events, such as removal of signal and transit peptides and precursor sequences. The CDC checksum for the UniProt entry was also stored so that any changes to the sequence could be identified subsequently. Kinetic data (K_m, K_{cat}, and/or K_m/K_{cat}) were stored where available. Annotations to indicate how the peptidase and cleavage position were identified were also stored. The initials of the curator and the date the cleavage was collected were also stored. The reference was stored in a Reference Manager database (Thomson Reuters) and the PubMed accession was obtained and stored where possible. Any additional data that affected where cleavage occurred, such as reactions conditions, where stored as a comment in the MySQL database.

To ensure that curation was consistent, a Perl program was written to aid cleavage data collection and storage. The user (either the author or a summer student) was asked to enter his or her initials; the UniProt accession of the protein substrate in question; the cleavage position; the residue range of the substrate sequence compared to the UniProt entry; the codes for how the cleavage was identified and how the peptidase was identified; whether the cleavage was physiological, non-physiological, pathological or theoretical; whether the substrate was denatured; the reference and its PubMed identifier; and any comment.

Collection of cleavage data from the literature was also outsourced to Molecular Connections, Bangalore, India. Data were returned to the author as an Excel spreadsheet and a pipeline developed to extract data from the spreadsheet and import it into the MySQL database. Existing substrate cleavage collections were also imported into the MEROPS collection. These included data from the CutDB database [14] and the CASBAH database of caspase substrates [15].

A Perl program to check that the P4–P4′ residues around the cleavage position matched the sequence in the UniProt entry was written as a quality control measure and to identify any subsequent changes in the UniProt sequence.

Cleavage data were also stored for the cleavage of synthetic substrates. These were manually entered into the database. For a synthetic substrate it was not possible to map the sequence to a UniProt database entry. The P4–P4′ positions around the scissile bond were stored where possible (many synthetic substrates do not have residues beyond P1′ or P3), including a unique identifier for each N- or C-terminal blocking or reporter group occurring within that range.

In certain cases, it was not possible to map a cleavage to a single enzyme. This most frequently occurred when cleavage was performed by an enzyme complex, such as the proteasome or

eukaryotic signal peptidase, or where cleavage might have been performed by one or more enzymes, such as removal of the initiating methionine by either of the eukaryotic methionyl aminopeptidases, or where specificity overlapped, for example caspases-3 and -7. Instead of mapping the cleavage to an individual enzyme, the cleavage was mapped to the peptidase family.

2.3. Automated collection of substrate cleavages from SwissProt entries

Entries in the SwissProt section of UniProt are annotated for removal of initiating methionine, signal and transit peptides, and processing events to activate proteins or extract peptides and proteins from polyproteins. These, if not already present in the substrate cleavage collection, were automatically collected using a Perl script. Removal of an initiating methionine was mapped to a MEROPS identifier either for methionyl aminopeptidase (for a prokaryote) or the M24 family (for a eukaryote). Removal of a signal peptidase was mapped to a MEROPS identifier either for signal peptidase or a signal peptidase complex (depending on whether the source species was a prokaryote or a eukaryote). Where the peptidase responsible for the cleavage was not known, the cleavage was not mapped to a specific peptidase or peptidase family. Theoretical cleavages, for example the release of a predicted signal peptide, were not collected.

2.4. Automated collection of substrate cleavages from proteomics experiments

High-throughput identification of proteins in a sample is known as “proteomics”. In many proteomics experiments, proteins are digested with a peptidase so that each protein can be identified from the mass and charge (and sometimes sequence) of its peptides. This requires a prior *in silico* digestion of all proteins in the proteome, and the choice of the peptidase used for the digestion is important: the peptidase must have a very simple specificity that is easily predictable. For this reason, trypsin. which cleaves all lysyl and arginyl bonds in a denatured protein, with the exception of except Lys-Pro and Arg-Pro bonds, is frequently used. There are many tools to perform theoretical cleavages, for example PeptideMass (http://web.expasy.org/peptide_mass/, [16]) which predicts cleaves by peptidases such as trypsin, chymotrypsin, peptidyl-Lys metallopeptidase and glutamyl peptidase I as well as cleavages by chemicals such as CnBr. The neXtProt database includes trypsin digestions of all human proteins and indicates peptides unique to a protein and whether a particular peptide is found in many human proteins [17]. The PRIDE database [[18]; http://www.ebi.ac.uk/pride/archive/] stores the results of proteomics experiments where peptides are identified from a sample by mass spectroscopy. There are a number of different techniques employed of which PICS, TAILS, COFRADIC, Subtiligase, ChaFRADIC and N-terminomics are the most well-known; these have been reviewed by Lai et al. [19] and Schlage & auf dem Keller [20]. TAILS [21] and COFRADIC [22] are techniques that have been employed to identify physiological substrates, including naturally occurring N-termini. Frequently, the peptides derived from a protein can provide evidence of post-translational processing, such as removal of an initiating methionine, signal peptide or transit peptide. A Perl program was written to collect peptides and their corresponding UniProt entries from the PRIDE database and check each to see if the N-terminus matched the site of removal of a hypothetical signal or transit peptide. Because aminopeptidases may further trim an N-terminus once a transit peptidase has been removed, for example aminopeptidase P3 removes an unstable amino terminal tyrosine from mitochondrial proteins following removal of the first transit peptide by mitochondrial processing peptidase but before removal of a second transit peptide by mitochondrial intermediate peptidase [23], the cleavage was only added to the MEROPS collection if the peptide identified and the predicted processing event corresponded. Cleavages from digests performed with more unusual peptidases, such as glutamyl endopeptidase for example, were also collected to boost the number of substrates for these peptidases.

In the last decade or so, there have been a number of attempts to determine peptidase specificity or discover physiological substrates by use of proteomics. In a typical proteomics experiment, a sample, often a human cell lysate, is divided into two portions. The first portion is digested only by trypsin (or a similar peptidase with a known and predictable specificity), and the second with trypsin and a peptidase of choice. The peptides generated are identified by mass spectroscopy. The cleavages that occur only in the second portion are the result of the peptidase of choice. If the second portion is digested first by trypsin and then by the peptidase of choice, then a suite of peptides is prepared to investigate the specificity of the peptidase of choice and the cleavages by this peptidase are not physiologically relevant. Alternatively, if the second portion is digested first with the peptidase of choice and then by trypsin, the digested proteins are likely to include physiological substrates. However, some proteins may be digested because the usual physiological barriers to digestion (compartmentalization, differences in pH, *etc*) have been removed in the experiment, and some proteins may be “bystanders” which are cleaved under physiological conditions but the cleavage is non-functional [24]. The results from analyses by software such as Mascot [25] or PeptideProphet [26] are exported to Excel spreadsheets. These are often deposited as material supplementary to the paper, but spreadsheets have also been kindly supplied by the authors. In a given sample, there can be hundreds or even thousands of peptide products, so manual curation is not possible. A Perl program was written to extract the accessions or identifiers, translate them into UniProt accessions, and then determine the residue number for the P1 residue in each cleavage. All of the cleavages from the seminal paper by Schilling & Overall (2008) [27] were collected, including cleavages by trypsin, but for subsequent papers, trypsin cleavages were not collected.

2.5. Reliability scores

A “reliability score” is calculated which is an average percentage difference for all substrates of a peptidase over the range P4–P4′. All possible pairwise comparisons are made and the number of differences summed for all comparisons. For substrates with non-standard amino acids and blocking and reporter groups such as those found in synthetic substrates, the unusual residues are replaced with “X” before the comparisons are performed. The sum of differences is divided by the number of comparisons times the number of positions considered. For most endopeptidases, the number of positions considered is eight, because at least of one of each of the positions P4 to P4′ will be filled. However, for aminopeptidases, where P4 to P2 are empty, and carboxypeptidases, where P2′ to P4′ are empty, only five positions are considered. For dipeptidyl-peptidases and peptidyl-dipeptidases only six positions are considered, and for dipeptidases only two. The number of positions considered is calculated and not assumed. A peptidase with a reliability score of greater than 75 means that its substrates are varied in sequence and the calculated preferences are likely to be correct. A peptidase with a score in the range 50–74% had a proportion of substrates with similar sequences and thus the calculated preferences were less reliable, and the calculated preferences for a peptidase with a score of less than 50% should be treated with caution.

3. Results and discussion

3.1. Number of peptidases

By using the criteria of Barrett & Rawlings (2007) [1], it was possible to distinguish 2457 different peptidase species. A holotype was established for each of these. In addition, there are a further 250 different peptidase activities in the literature for which either no or too little sequence information exists to be able to map any of them to a UniProt accession. The breakdown into different catalytic types is shown in Table 1. More serine peptidases (554 or 20%) were found than for any other catalytic type.

The holotypes and their sequence homologues, 502,782 sequences in total, were found to be distributed amongst 254 families. This total includes 3669 sequences that are classified as asparagine lyases rather than peptidases, because proteolysis is dependent upon cyclization of an asparagine residue to a succinimide which does not involve hydrolysis [28].

3.2. Numbers of substrates

The total number of cleavages found for all peptidases was 66,615. This number includes substrates cleaved in the same position by different peptidases. Most of these cleavages (59,276 or 89%) were in peptides or proteins that could be mapped to UniProt identifiers: 5821 (or 9%) were in synthetic substrates. The majority of these cleavages (36,229 or 54%) are non-physiological, but 20,264 (30%) are thought to be physiologically relevant, and a further 1349 (2%) are pathological. The totals include self-cleaving reactions, which in the case of asparagine lyases is all the cleavages identified for this catalytic type.

The breakdown of substrate cleavages per catalytic type is shown in Table 2. Most cleavages (30,711 or 46%) are for serine peptidases, a number boosted by the large number of non-physiological cleavages (21,301 or 32%).

Table 3 shows the number of peptidases in each catalytic type for which at least one substrate cleavage is known. The total number is 1700 or 63% of the total number of different peptidases. Comparing the numbers with those in Table 1 shows that for each catalytic type, a substrate cleavage is known for most peptidases. Substrates may be known for any of the other 1007 peptidases, but cleavages were either not found or were not accessible in the literature, or the site of cleavage is unknown. On average a peptidase has 39 substrate cleavages in the MEROPS collection. Table 4 shows the peptidases with 10 or more known cleavages: the large number of non-physiological cleavages for the serine peptidase trypsin, 12,303 of which are derived from the proteomics paper of Schilling & Overall (2008) [27], where it is used for preparation of samples prior to mass spectroscopy analysis, explains the preponderance of cleavages for serine peptidases. To be able to study peptidase specificity and make predictions about where in a protein cleavage might occur, at least 40 cleavages in substrates are required (Robert Pike, personal communication). The number of peptidases with 40 or more cleavages is 163 (or 6% of the total number of peptidases). It is immediately apparent that except for a small number of peptidases, insufficient numbers of cleavages in substrates are known to be able to draw firm conclusions about the specificity of most peptidases. The distribution of cleavages per peptidase is shown in Fig. 1.

There are 36 peptidases with 100 or more known cleavages in physiological substrates. The peptidases with most physiological cleavages are general processing peptidases, such as the animal signal peptidase complex (a complex containing two peptidases with 1879 cleavages) and methionyl aminopeptidase 1 (964 cleavages). Almost all of these cleavages are derived from annotations in the SwissProt section of the UniProt database, and include theoretical cleavages confirmed by proteomics experiments submitted to the PRIDE database. Other peptidases with many physiological cleavages include peptidases involved in protein turnover such as cathepsins E (1553 cleavages) and D (871) and granzyme M (891) almost all of which are derived from proteomics studies [29,30]. There are 27 peptidases with 200 or more known cleavages in substrates thought not to be physiological. As mentioned above, the use of trypsin to produce peptides for mass-spectroscopy in proteomics experiments means that more cleavages (14,201) are known for this peptidase rather than any other. Large numbers of non-physiological cleavages are also known for peptidases regularly used in manual protein sequencing such as pepsin A (350 cleavages), peptidyl-Lys metallopeptidase (2105), chymotrypsin A (1012), glutamyl peptidase 1 (1273) and lysyl endopeptidase (802). Other peptidases with many non-physiological substrates are again the subjects of proteomic studies: matrix metallopeptidase-2 (2726 cleavages) [27]; human granzyme B (1136) [30,31]; glutamyl peptidase I (1003) [27]; cathepsins L (965), S (694) and B (486) [32]; meprin beta (891) and alpha subunits [33]; and the RC1339 protein from *Rickettsia conorii* (799) [34]. There are 57 peptidases with 20 or more cleavages in synthetic substrates and neurolysin (139 cleavages) and thimet oligopeptidase (116) have the most. There are 43 peptidases with cleavages in a hundred or more proteins, including cleavages in peptides that can be mapped to protein sequences in the UniProt database. Once again, because of its use in proteomics, more proteins are known that are susceptible to trypsin (3188) than any other peptidase. The methods by which cleavages were identified include mass spectroscopy (40,729 cleavages), N-terminal sequencing (8284), from knowing the consensus cleavage site (1609), amino acid analysis (205) and site-directed mutagenesis (178).

3.3. Substrate specificity

Table 4 shows the specificity of peptidases based on the occurrence of amino acids in the binding pockets P4–P4′. The preferences for each peptidase are defined in words so that the table represents a classification by specificity and homology. This also allows classification by amino acid type (acidic, basic, aliphatic, aromatic, *etc*) which would not be possible if logos were shown. The MEROPS identifier for each peptidase is a link to the peptidase summary page on the MEROPS website where the specificity logo is displayed. Caution should be exercised where these preferences are derived from fewer than 40 substrates. The number of peptidases

Table 1
Counts of different peptidases (peptidase species) by catalytic type.

	Aspartic	Glutamic	Metallo	Cysteine	Serine	Threonine	Mixed	Asparagine lyases	Unknown	Total
Sequenced and characterized	170	7	633	615	942	46	5	23	16	2457
Sequenced only	118	0	327	297	644	29	0	1	3	1419
Sequence not known	8	0	86	19	87	1	0	0	49	250
Non-peptidase homologues	4	0	103	52	145	26	0	0	0	330
Pseudogenes	24	0	5	21	17	3	0	0	0	70
Total	324	7	1154	1004	1835	105	5	24	68	4526

Table 2
Counts of substrates per catalytic type.

	Aspartic	Glutamic	Metallo	Cysteine	Serine	Threonine	Mixed	Asparagine lyases	Unknown	Total
Physiological	2780	7	4113	7311	5877	33	2	60	81	20,264
Pathological	266	0	345	700	34	0	0	0	4	1349
Non-physiological	2893	70	8762	3112	21,301	84	1	3	3	36,229
Synthetic	364	32	1569	1179	2547	42	37	0	51	5821
Theoretical	176	0	545	106	638	0	0	300	0	1765
Unclassified	77	0	554	194	314	26	1	0	21	1187
Total	6556	109	15,888	12,602	30,711	185	41	363	160	66,615

Table 3
Counts of peptidases with known substrate cleavages by catalytic type.

	Aspartic	Glutamic	Metallo	Cysteine	Serine	Threonine	Mixed	Asparagine lyases	Unknown	Total
Sequenced and characterized	114	4	477	401	520	18	2	17	33	1586
Sequence not known	5	0	41	6	34	0	0	0	28	114
Total	119	4	518	407	554	18	2	17	61	1700

for which the cleavage data are considered reliable (highlighted in green) is 319. The number for which the data are less reliable (yellow highlighting) is 123, and the number for which the data should be treated with caution (red highlighting) is 68. The vast majority of peptidases with a low reliability score are those showing apparent specificity in more than three binding pockets. It should be noted, however, that the reliability score will be affected by the number of binding pockets in which specificity is shown: this is apparent for the deubiquitinating hydrolases where the substrates positions P4–P1 are occupied in physiological substrates by the highly conserved C-terminus of ubiquitin (Leu–Arg–Gly–Gly). Because proteins substrates are unlikely to be homologous, a peptidase with many protein substrates will have a large reliability score. For a peptidase with cleavages in predominantly synthetic substrates, the reliability score will be low, because synthetic substrates tend to be similar with only the blocking and reporter groups differing. The peptidases with the highest reliability scores (96%) are asclepain A (C01.008); the snake venom enzyme jerdohagin (M12.216) and bpr peptidase from *Dichelobacter nodosus* (S08.022). In all three examples, protein cleavages are from the insulin B-chain [35–37]. The peptidase for which the reliability score is lowest (12%) is polyglycine endopeptidase (U9G.075): most known cleavages are from glycine-rich regions in endochitinases [38].

The table shows where one or two amino acids predominate in any binding pocket, or where a defined group of amino acids predominate. Also shown are one or two amino acids that are unknown in any substrate binding pocket where no preference is known to exist. Because some amino acids are rare or rarely encountered around a cleavage site, for example cysteine and tryptophan, these negative preferences are only shown when 200 or more cleavages are known for the peptidase. Where a binding pocket does not exist, for example P4–P2 for an aminopeptidase or P2′–P4′ for a carboxypeptidase, these are shown with a grey background. Please note that for some exopeptidases the name indicates a preference and not a strict specificity. For example, bacterial-type alanyl aminopeptidase (M01.005) has also been shown to act as an endopeptidase [39]. DmpA aminopeptidase (P01.001) also acts as an endopeptidase, processing its own precursor [40]. For many aminopeptidases, only cleavages in synthetic substrates are known and these frequently lack residues in positions P2′ to P4′. These apparent absences in binding pockets S2′–S4′ are not shown in Table 4 because they would be misleading.

The majority of peptidases for which specificity can be inferred (107 endopeptidases and 43 exopeptidases) show a preference in a single substrate binding pocket. Most of these peptidases show a preference in the S1 pocket, but there is a preference in each binding pocket S4–S4′ for at least one endopeptidase. There is an overwhelming bias for basic residues in P1, which probably reflects the large number of characterized peptidases in family S1. Exopeptidases, perhaps unsurprisingly, show preference only in pockets P1 and P1′. The TET aminopeptidase is the only exopeptidase to show a preference beyond the residues either side of the cleavage site.

Several peptidases show a preference in more than one binding pocket. There are 79 peptidases with a preference in two binding pockets, 50 peptidases with a preference in three pockets and 29 peptidases with a preference in four pockets. The table also includes examples of peptidases with preferences in more binding pockets, but as with the case for thimet oligopeptidase (see below) these apparent preferences are more likely to reflect the design of the substrates tested rather than true specificities.

Despite the large number of cleavages in substrates, the specificity of some peptidases cannot be explained in terms of preferences in binding pockets. Table 4 shows the 53 peptidases with more than forty substrate cleavages for which no preference is shown. This list includes some well-known and well-characterized peptidases such as cathepsin G, for which even a large proteomics study was unable to show any preference except that lysine is not acceptable in P4 [27].

Table 4 also shows that peptidases with similar specificities can be unrelated to one another (members of different families or even catalytic types), for example gingipain K and bacterial lysyl endopeptidase both have a preference for Lys in P1, yet the former is a cysteine peptidase and the latter a serine peptidase. There are also examples where specificity differs markedly from peptidases with homologous sequences, with examples from family S1 being the most well known in which chymotrypsin A has a preference for aromatic residues in P1, trypsin 1 for basic residues, elastase-2 for aliphatic residues and granzyme B for aspartic acid. It is more unusual for peptidases within a family to utilise different substrate binding pockets, but an example is granzyme B which also shows a preference for an aliphatic residue in P4. A family may include exo- as well as endopeptidases, for example family S9 includes prolyl oligopeptidase and dipeptidyl-peptidase IV, although both have a preference for Pro in P1. These facts emphasize that peptidases cannot be classified to the protein species by specificity or sequence similarity alone, but that both specificity and homology should be taken into consideration.

There are many peptidases for which the number of physiological substrates is limited. These include proteins that process themselves to remove a propeptide or to expose a N-terminal nucleophile and then cease to function as peptidases. Examples include the amidophosphoribosyltransferase precursor (C44.001) [41] and the penicillin V acylase precursor (C59.001) [42], for both of which the only known cleavage is to expose a new N-terminal nucleophilic cysteine. Human peptidases such as renin (A01.007) have few natural substrates but the number of cleavages is boosted by synthetic substrates and peptides.

3.4. Displays on the MEROPS website

A variety of displays have been implemented on the MEROPS website to aid understanding of peptidase specificity. The displays include the following.

In each peptidase summary, where ten or more substrate cleavages are known, a logo (generated by the Weblogo software, http://weblogo.berkeley.edu/) and a specificity matrix are shown. Fig. 2 shows the logo and specificity matrix for thimet oligopeptidase. In the logo, amino acid residues are shown as single-letter abbreviations, and the greater the height of the letters the greater the preference for the amino acid in that binding pocket [43]. The binding pockets S4–S4′ are numbered along the X axis as 1–8. In the specificity matrix, amino acids are shown in three-letter notation, and the number of times each occurs within the range P4–P4′ for a substrate is shown. The occurrence of each amino acid in each binding pocket is calculated as a percentage of total cleavages, and a different shade of green is used for each tenth percentile. The brighter green the background highlighting, the greater the preference for that amino acid in the corresponding substrate binding pocket. If an amino acid has not been observed in any binding

Table 4
Peptidases with 10 or more known cleavages and peptidase specificity derived from substrate cleavages in the *MEROPS* collection. For each peptidase with ten or more known substrate cleavages are shown. Peptidases are arranged by specificity (number of binding pockets, then preference in each binding pocket in the order P4 to P4′, then by amino acid in alphabetical order) and then by MEROPS identifier. For each peptidase, the MEROPS identifier, the recommended peptidase name, the number of substrate cleavages, and preferences for binding pockets P4 to P4′ are shown. The brighter the shade of green, the greater the preference; five shades are shown ranging from darkest green (50–59% of substrates) to brightest green (90% or greater of substrates). Up to two amino acids are shown in a binding pocket where a preference occurs. Where the preference is for any of a group of amino acids, or the preference for the group is greater than that for a single or amino acid, the following symbols are shown: λ (aliphatic: Ile, Leu, Val), @ (aromatic: Phe, Trp, Tyr), + (acidic: Asp, Glu), − (basic: Arg, His, Lys), Σ (small: Ala, Cys, Gly, Ser) and Ω (other: Asn, Gln, Met, Pro, Thr). Where no preference for an amino acid or group of amino acids exists, and where there are 200 or more cleavages, up to two amino acids that are not acceptable in a binding pocket are shown as white text on a black background. For exopeptidases which act at N- and C-termini of proteins, no residue may be possible in some binding pockets and in these cases the binding pockets are shaded grey. Binding pockets shaded black or grey are ignored for the ordering of items in the table. The "Reliability score" is the percentage difference calculated by counting all the differences between substrates for the same enzyme, dividing the total differences by the number of comparisons times the number of residues P4–P4′ considered, and multiplying by 100. Reliability scores of 75% or greater difference are highlighted in green; scores 50% or greater in yellow, and scores of less than 50% in red. See text for details.

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
Specificity at one binding pocket: P4											
Ω								M10.019	matrix metallopeptidase-20	43	87
Σ								S08.056	cuticle-degrading peptidase of parasitic fungus	14	94
Specificity at one binding pocket: P3											
	P							M10.006	matrix metallopeptidase-10 (*Homo sapiens*-type)	19	89
	λ							S53.006	physarolisin	11	94
	Σ							S08.007	thermitase	13	91
Specificity at one binding pocket: P2											
		P						M03.004	oligopeptidase A	10	68
		λ						C01.002	chymopapain	12	88
		λ						C01.009	cathepsin V	14	83
		λ		P				C01.034	cathepsin S	756	91
		λ						C01.036	cathepsin K	182	90
		λ						C10.001	streptopain	27	91
		Ω						C01.030	crustapain	10	85
		Σ						C01.073	peptidase 1 (mite)	29	92
Specificity at one binding pocket: P1											
			A					M32.005	LmaCP1 carboxypeptidase (*Leishmania major*)	13	42
			A					S10.001	carboxypeptidase Y	17	50
18			A					S10.004	serine carboxypeptidase C	18	48
19			A					S10.005	serine carboxypeptidase D	16	51
			A					S10.008	carboxypeptidase OcpA	14	49
			A					S10.009	serine carboxypeptidase III (plant)	16	48

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
			AP					S09.005	dipeptidyl aminopeptidase A	14	67
			D					M18.002	aspartyl aminopeptidase	10	55
			D					M20.010	DapE peptidase	15	50
			D					T02.002	isoaspartyl dipeptidase (threonine type)	17	48
			E					M01.003	aminopeptidase A	14	88
			F					A01.022	plasmepsin-1	13	83
			F					M28.004	aminopeptidase AC	13	43
			F					S09.070	acylaminoacyl peptidase (*Aeropyrum*-type)	14	70
			G					M23.001	beta-lytic metallopeptidase	15	75
			L		W		W	S01.139	granzyme M	1363	88
			K					A01.085	candidapepsin SAP10	16	90
			K					C25.002	gingipain K	47	82
			K					S01.280	lysyl endopeptidase (bacteria)	809	82
			N					C13.004	legumain, animal-type	81	84
			P					M02.006	angiotensin-converting enzyme-2	14	82
			P					M9E.004	membrane Pro-Xaa carboxypeptidase	20	59
			P					S09.001	prolyl oligopeptidase	59	79
			P					S09.003	dipeptidyl-peptidase IV (eukaryote)	28	82
			P					S09.013	dipeptidyl-peptidase 4 (bacteria-type 2)	17	80
			P					S09.018	dipeptidyl-peptidase 8	10	83
			P					S09.019	dipeptidyl-peptidase 9	15	57
			P					S28.001	lysosomal Pro-Xaa carboxypeptidase	11	52
			P					S28.002	dipeptidyl-peptidase II	30	51
			R					C01.064	RD21 peptidase	20	81
			R					C01.115	fascipain B	10	62
			R					C14.035	metacaspase Yca1 (*Saccharomyces cerevisiae*-type)	11	53
			R					C14.043	metacaspase Ld (*Leishmania*-type)	27	76
			R					C25.003	gingipain R2	16	88
			R					C69.002	arginine aminopeptidase (*Streptococcus*-type)	11	67
			R					M01.014	aminopeptidase B	15	80
			R					M28.021	AaaA aminopeptidase	13	50
			R					S01.017	kallikrein-related peptidase 5	75	87
			R					S01.021	DESC1 peptidase	108	81
			R					S01.033	factor VII-activating peptidase	11	87
			R					S01.047	human airway trypsin-like peptidase	95	82
			R					S01.071	kallikrein 1-related peptidase b9 (*Mus musculus*)	14	70
			R					S01.172	kallikrein 1-related peptidase c2 (*Rattus norvegicus*)	21	80
			R					S01.224	hepsin	110	80
			R					S01.302	matriptase	112	81
			R					S01.308	matriptase-2	161	80
			R					S33.010	SCO7095-type peptidase	11	62
			R					XP01-001	tricorn peptidase complex	14	70
			W					M9A.008	tryptophanyl aminopeptidase (*Trichosporon cutaneum*)	15	73

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
			+					M61.002	BcepAP aminopeptidase (*Burkholderia cepacia*)	10	63
			+	P				S01.269	glutamyl peptidase I	1405	85
			+					T01.010	proteasome catalytic subunit 1	15	71
			λ					S01.131	elastase-2	483	91
			λ					S01.273	DegP peptidase	15	83
	W		λ					S01.278	HtrA2 peptidase	233	89
			λ					S01.494	Htra2 peptidase (*Mycobacterium*-type)	39	90
			@					S01.001	chymotrypsin A (cattle-type)	1057	91
			@					S01.140	chymase (*Homo sapiens*-type)	106	89
			@					S01.149	mast cell peptidase 4 (mouse numbering)	10	91
			@					S01.152	chymotrypsin B	26	76
			@					S01.300	kallikrein-related peptidase 7	17	89
			@					S10.002	serine carboxypeptidase A	10	47
			-					A01.067	candidapepsin SAP9	17	93
			-					C01.006	ficain	15	87
			-					C01.144	cathepsin B-like cysteine peptidase (*Raphanus sativus*)	24	92
			-					C10.004	interpain A	22	91
			-					M12.169	metallopeptidase B (*Bothrops moojeni*)	12	80
			-					M27.002	bontoxilysin	10	87
			-					S01.029	kallikrein-related peptidase 14	74	85
			-					S01.072	matriptase-3	112	85
			-					S01.143	tryptase alpha	23	86
			-					S01.146	granzyme K	20	84
			-					S01.151	trypsin 1	13770	88
			-					S01.160	kallikrein 1	31	87
			-					S01.174	mesotrypsin	22	85
			-					S01.212	plasma kallikrein	40	84
			-					S01.233	plasmin	126	89
			-					S01.236	kallikrein-related peptidase 6	75	85
			-					S01.243	lumbrokinase	25	89
			-					S01.244	kallikrein-related peptidase 8	25	68
			-					S01.251	kallikrein-related peptidase 4	123	86
			-					S01.294	HAT-like 3 peptidase	11	81
			-					S01.357	polyserase-IA unit 1	11	43
			-					S01.407	kallikrein 1-related peptidase c9 (*Rattus* sp.)	14	85
			-					S09.010	oligopeptidase B	42	54
			-					T01.011	proteasome catalytic subunit 2	14	87
			Σ					M10.007	matrix metallopeptidase-11	21	90
			Σ					M12.144	atrolysin C	60	90
			Σ					M12.150	ruberlysin	16	83
			Σ					M12.157	mucrolysin	14	66
			Σ					M12.208	ADAM8 peptidase	23	87
			Σ					M20.006	carnosine dipeptidase I	17	57
			Σ					M28.014	carboxypeptidase Q	33	89
			Σ					M9A.012	glycyl aminopeptidase (*Actinomucor elegans*)	10	54
			Σ					U9B.004	peptidyl-D-amino acid hydrolase (cephalopod)	13	65

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
Specificity at one binding pocket: P1'											
				D				M72.001	peptidyl-Asp metallopeptidase	21	84
				FL				M04.010	vimelysin	12	83
				K				M14.023	CPG70 carboxypeptidase (*Porphyromonas gingivalis*)	27	77
				K				M35.004	peptidyl-Lys metallopeptidase	2105	82
				P				M24.003	Xaa-Pro dipeptidase (bacteria-type)	13	50
				P				M24.009	aminopeptidase P1	18	59
				P				M24.033	aminopeptidase P (*Streptomyces*-type)	12	57
				R				M14.005	carboxypeptidase E	20	67
				+				M12.004	meprin beta subunit	931	91
				+				M9E.007	carboxypeptidase G3	12	31
				λ				M04.001	thermolysin	280	92
				λ				M04.007	coccolysin	23	92
				λ				M04.009	aureolysin	24	91
				λ				M05.001	mycolysin	12	89
				λ				M06.001	immune inhibitor A peptidase (*Bacillus* sp.)	22	89
				λ				M10.001	matrix metallopeptidase-1	75	90
				λ				M10.003	matrix metallopeptidase-2	3417	90
				λ				M10.008	matrix metallopeptidase-7	196	87
				λ				M12.138	jararhagin	40	91
				λ				M12.140	bothropasin	62	90
				λ				M14.017	carboxypeptidase A4	68	90
				λ				M14.018	carboxypeptidase A6	42	88
				λ				M9G.034	metallopeptidase ShpII (*Staphylococcus hyicus*)	21	94
				λ@				M13.008	neprilysin-2	16	86
				@				M12.306	BHRb haemorrhagin (*Bitis arietans*)	11	94
				@				M13.003	endothelin-converting enzyme 2	17	90
				@				M13.009	Zmp1 peptidase (*Mycobacterium*-type)	10	93
				@				M9A.007	Xaa-Trp aminopeptidase	13	68
				-				M14.004	carboxypeptidase N	15	82
				-				M14.006	carboxypeptidase M	11	81
				-				M14.009	carboxypeptidase B2	19	84
			C	Σ	KR			M12.001	astacin	209	91
				Σ				M12.225	ADAMTS5 peptidase	38	92
Specificity at one binding pocket: P2'											
					λ			M12.311	BaP1 peptidase (*Bothrops asper*)	45	90
					λ			M30.001	hyicolysin	15	95
					λ			M73.001	camelysin	11	92
					Ω			M03.007	oligopeptidase F	24	88
Specificity at one binding pocket: P3'											
						Ω		A01.042	syncephapepsin	18	93
						Ω		M42.005	TET aminopeptidase	12	94
						Σ		C01.060	cathepsin B	632	90
						Σ		M04.003	vibriolysin	17	91
						Σ		M12.210	ADAM10 peptidase	40	90
Specificity at one binding pocket: P4'											

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
							Ω	S9G.114	CP790 peptidase (*Lactobacillus helveticus* CP790)	32	93
Specificity at two binding pockets: P4 and P1											
L			Σ					M12.345	moojeni peptidase A	10	91
R			R					S08.077	PCSK7 peptidase	116	72
λ			D					S01.010	granzyme B (*Homo sapiens*-type)	1873	85
λ		CR	D	PR	W			S01.136	granzyme B, rodent-type	487	83
λ			L					A01.084	yolk cathepsin (*Boophilus* sp.)	17	88
Specificity at two binding pockets: P4 and P1'											
Ω				A				M16.004	chloroplast (stromal) processing peptidase	19	87
Specificity at two binding pockets: P3 and P1											
	A		@					S01.262	streptogrisin B	14	57
M	E		D	Q				C14.005	caspase-6	201	77
	λ		λ					A01.004	memapsin-2	24	87
	Ω		-					S01.034	transmembrane peptidase, serine 4	10	85
	Σ		E					S80.001	prohead peptidase gp175 (*Pseudomonas aeruginosa* phage phiKZ)	25	72
	Σ		+					S01.271	glutamyl peptidase BL	72	80
	Σ		Σ					S26.008	thylakoidal processing peptidase	52	76
W	Σ		Σ	P	M			S26.010	signalase (animal) 21 kDa component	363	88
	Σ		Σ					XS26-001	signal peptidase complex (animal)	1878	89
Specificity at two binding pockets: P3 and P1'											
	@			@				M02.002	peptidyl-dipeptidase Acer	11	68
Specificity at two binding pockets: P2 and P1											
		F	R					C9G.030	bothropain	12	38
		F	R					S01.405	kallikrein 1 (*Rattus* sp.)	10	49
		G	R					S01.214	coagulation factor IXa	12	55
		P	F					S01.438	fire ant chymotrypsin	17	46
		R	R					C01.050	histolysain	10	45
		R	R					M10.060	epralysin	10	65
		V	P					S09.073	Xaa-Pro dipeptidylpeptidase	19	35
		+	AP					S09.069	omega peptidase (*Drosophila melanogaster*)	18	73
		λ	+					S01.443	glutamyl peptidase BI	14	89
		λ	+					S46.002	dipeptidyl-peptidase 11 (*Porphyromonas gingivalis*-type)	10	43
		@	-					C01.072	rhodesain	11	47
		-	R					S08.070	kexin	191	71
		-	-					M16.011	falcilysin	10	92
		-	-					S08.072	PCSK1 peptidase	92	78
		-	-					S9G.012	tryase	14	48
		Ω	R					S01.015	tryptase beta	22	48
		Ω	-					S01.218	protein C (activated)	15	82
		Σ	P					S15.001	Xaa-Pro dipeptidyl-peptidase	10	46
		Σ	-					S01.216	coagulation factor Xa	59	78
		Σ	-					S01.231	u-plasminogen activator	17	79
Specificity at two binding pockets: P2 and P1'											
		λ		Σ				C01.032	cathepsin L	1089	91

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
Specificity at two binding pockets: P2 and P2'											
		R			Σ			M16.003	mitochondrial processing peptidase beta-subunit	81	81
		V			λ			A02.007	feline immunodeficiency virus retropepsin	28	84
Specificity at two binding pockets: P1 and P1'											
			D	Σ				C14.001	caspase-1	181	83
			D	Σ				C14.018	caspase-14	17	79
			G	@				M12.136	leucolysin	16	71
			FG	@				M14.001	carboxypeptidase A1	16	61
			L	@				M28.003	aminopeptidase S (*Streptomyces*-type)	10	50
			M	Σ				M24.001	methionyl aminopeptidase 1 (*Escherichia*-type)	978	71
			M	Σ				M24.002	methionyl aminopeptidase 2	135	71
			M	Σ				M24.028	mitochondrial methionyl aminopeptidase	131	67
			M	Σ				M9B.004	Met-Xaa dipeptidase	14	72
			P	Σ				S09.017	prolyl tripeptidyl peptidase	13	74
			R	R				A26.001	omptin	54	81
			R	Σ				C25.001	gingipain R	26	84
			Y	S				M24.026	Icp55 peptidase	39	75
			+	E				M28.010	glutamate carboxypeptidase II	10	50
			λ	λ				M13.004	oligopeptidase O1	10	90
			-	λ				S01.306	kallikrein-related peptidase 13	21	82
			-	Σ				C01.027	comosain	13	47
	W		-	Σ		W	RW	S01.135	granzyme A	441	86
			-	Σ				S01.217	thrombin	186	82
			Σ	F				M9G.018	neutral endopeptidase (*Micrococcus caseolyticus*)	12	78
			Σ	L				M04.005	pseudolysin	70	82
			Σ	L				M12.145	atrolysin E	19	84
			Σ	L				M12.309	hemorrhagic toxin I (*Gloydius halys blomhoffii*)	12	90
			Σ	S				M12.214	ADAM19 peptidase	13	84
			Σ	Σ				P01.001	DmpA aminopeptidase	35	33
Specificity at two binding pockets: P1 and P2'											
			N		λ			C13.001	legumain (plant beta form)	15	79
			P		Σ			S33.008	prolyl aminopeptidase 2	10	52
Specificity at two binding pockets: P1 and P3'											
			-			Σ		S01.281	arginyl peptidase	11	82
W	W	W	Σ	C		Σ		M10.004	matrix metallopeptidase-9	369	88
W		C	Σ			Σ	W	M10.009	matrix metallopeptidase-12	218	86
Specificity at two binding pockets: P1 and P4'											
			-				Σ	A01.030	yapsin-1	45	84
			-				Σ	S01.161	kallikrein-related peptidase 2	66	82
Specificity at two binding pockets: P1' and P2'											
				Q	P			C01.024	endopeptidase-B (*Hordeum*-type)	70	72
				L	λ			M12.178	brevilysin L6	16	81
				P	P			M24.005	aminopeptidase P2	13	44
				Y	λ			S53.004	kumamolisin	11	77
				Σ	Ω			M12.032	LAST peptidase (*Limulus*-type)	78	88

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
Specificity at two binding pockets: P1' and P3'											
				λ		Σ		M10.002	matrix metallopeptidase-8	114	88
				λ		Σ		M12.217	ADAM17 peptidase	60	90
Specificity at two binding pockets: P2' and P3'											
					Σ	+		S01.137	granzyme C	19	88
Specificity at three binding pockets: P4, P2 and P1											
L		G	GQ					C30.005	SARS coronavirus picornain 3C-like peptidase	26	71
R		-	-					S08.074	PCSK4 peptidase	103	71
R		-	-					S08.075	PCSK6 peptidase	105	69
R		-	-					S08.076	PCSK5 peptidase	129	72
λ		Σ	Ω					S08.012	PfSUB1 peptidase	23	77
-	CY	-	-	I	CY		W	S08.071	furin	208	72
-	CW	-	-		H			S08.073	PCSK2 peptidase	202	77
Specificity at three binding pockets: P4, P2 and P1'											
L		T		G				C03.020	enterovirus picornain 2A	12	75
λ		λ		Σ				C47.004	staphopain C	10	91
Specificity at three binding pockets: P4, P1 and P1'											
D			D	G				C14.002	CED-3 peptidase	14	74
+			D	Σ				C14.003	caspase-3	651	77
+			D	Σ				C14.004	caspase-7	179	75
Specificity at three binding pockets: P4, P2' and P4'											
Ω					λ		Ω	S01.480	HtrA peptidase (*Chlamydia trachomatis*-type)	14	93
Specificity at three binding pockets: P3, P2 and P1											
	E	Ω	D					C14.019	caspase DRONC (*Drosophila melanogaster*)-type peptidase	11	57
	L	-	-					S53.007	aorsin	10	45
	λ	L	-					C01.057	vinckepain-2	28	64
	λ	λ	λ					S01.150	mast cell peptidase 5 (mouse numbering)	44	84
	Σ	Σ	@					S33.005	tricorn interacting factor F1	16	49
Specificity at three binding pockets: P3, P1 and P1'											
	E		D	Σ				C14.007	caspase-4	10	79
	V		A	Σ				S21.002	cytomegalovirus assemblin	13	58
	Ω		Ω	Σ				M03.006	mitochondrial intermediate peptidase	16	87
	+		D	Σ				C14.009	caspase-8	68	76
	+		D	Σ				C14.010	caspase-9	14	85
	+		D	Σ				C14.011	caspase-10	10	80
	@		λ	Σ				S01.526	SplD peptidase (*Staphylococcus aureus*)	22	59
	@		@	Σ				S01.503	SplA peptidase (*Staphylococcus aureus*)	23	57
W	Σ	P	A	Σ	R		C	S26.001	signal peptidase I	342	75
Specificity at three binding pockets: P3, P1' and P2'											
	Σ			λ	Ω			M10.015	membrane-type matrix metallopeptidase-2	14	92
Specificity at three binding pockets: P3, P1' and P4'											
	λ			@			Ω	A01.019	polyporopepsin	11	90
Specificity at three binding pockets: P2, P1 and P1'											
		G	G	Σ				M23.002	staphylolysin	12	45
		L	Q	S				C30.001	coronavirus picornain 3C-like peptidase-1	44	66
		L	Q	S				C30.003	human coronavirus 229E main peptidase	11	69

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
		-	R	Σ				S07.001	flavivirin	21	74
		Σ	Σ	@				M9G.029	MAP1 peptidase (*Myxococcus xanthus*, *Pimelobacter* sp.)	17	80
Specificity at three binding pockets: P2, P1 and P2'											
		A	K		A			A01.011	penicillopepsin	17	66
		λ	F		Σ			A01.023	plasmepsin-2	15	83
Specificity at three binding pockets: P2, P1' and P2'											
		Σ		Σ	A			S53.011	scytalidolisin	20	47
Specificity at three binding pockets: P2, P1' and P3'											
		L		Σ		-		C01.075	cruzipain	65	77
Specificity at three binding pockets: P1, P1' and P2'											
			G	L	A			M04.017	griselysin	36	44
			G	L	Σ			M04.014	bacillolysin	41	74
			R	R	I			A26.003	plasminogen activator Pla	16	36
			Σ	D	+			M12.016	vertebrate tolloid-like 1 protein	17	77
			Σ	Σ	A			C13.005	glycosylphosphatidylinositol:protein transamidase	19	82
			Σ	Σ	Σ			M12.209	ADAM9 peptidase	18	84
			Σ	Σ	Σ			M20.003	peptidase T	12	77
Specificity at three binding pockets: P1, P1' and P4'											
			E	Σ			Σ	S32.001	equine arteritis virus serine peptidase	15	78
			T	V			Ω	M01.018	endoplasmic reticulum aminopeptidase 1	76	75
			-	λ			G	S01.257	kallikrein-related peptidase 11	16	68
Specificity at three binding pockets: P1', P2' and P3'											
				λ	-	-		M08.001	leishmanolysin	20	86
Specificity at three binding pockets: P1', P2' and P4'											
				Q	Ω		D	C01.001	papain	87	73
Specificity at four binding pockets: P4, P3, P2 and P1											
A	A	P	F					S08.037	subtilisin DY	16	40
L	R	G	G					C12.003	ubiquitinyl hydrolase-L3	26	30
L	R	G	G					C19.001	ubiquitin-specific peptidase 5	10	41
L	R	G	G					C65.001	otubain-1	10	43
Q	T	G	G					C48.002	SENP1 peptidase	11	40
Σ	T	F	G					C54.003	autophagin-1	15	37
Specificity at four binding pockets: P4, P3, P1 and P1'											
L	P		T	G				C60.001	sortase A (*Staphylococcus*-type)	16	45
Specificity at four binding pockets: P4, P3, P2' and P4'											
Ω	Ω				Σ		Ω	A02.005	bovine immunodeficiency virus retropepsin	11	91
Specificity at four binding pockets: P4, P2, P1 and P1'											
D		Ω	D	G				C14.006	caspase-2	16	73
λ		G	G	Σ				C05.001	adenain	26	76
Specificity at four binding pockets: P4, P2, P1 and P3'											
G		Σ	Ω			Q		M03.001	thimet oligopeptidase	124	72
Specificity at four binding pockets: P4, P2, P1' and P2'											
λ		Ω		G	Ω			C03.021	rhinovirus picornain 2A	12	81
Specificity at four binding pockets: P4, P1, P1' and P4'											
A			Q	G			Ω	C03.001	poliovirus-type picornain 3C	14	70

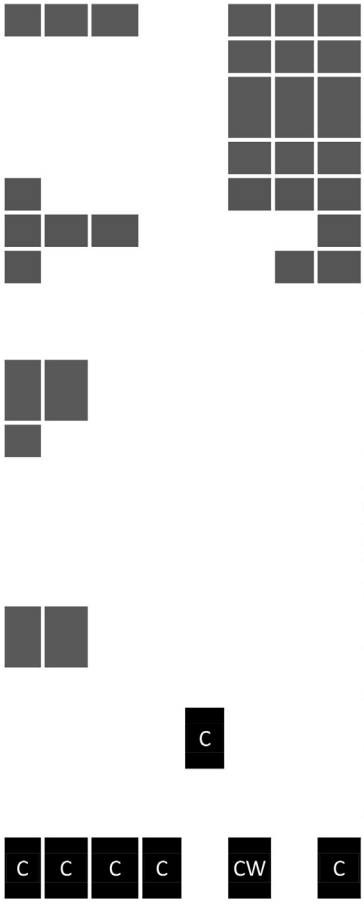
P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
Specificity at four binding pockets: P3, P2, P1 and P1'											
	Ω	λ	Σ	λ				M10.029	matrix metallopeptidase-26	17	87
	Σ	-	-	Σ				S08.109	KPC2-type peptidase	115	71
Specificity at four binding pockets: P3, P2, P1 and P2'											
	E	λ	Y		λ			M10.063	PrtA peptidase (*Photorhabdus*-type)	29	77
Specificity at four binding pockets: P3, P2, P1' and P4'											
	Ω	Σ		L			Σ	M10.066	karilysin	38	80
Specificity at four binding pockets: P3, P1, P1' and P3'											
	P		Σ	λ		Σ		M10.013	matrix metallopeptidase-13	147	83
Specificity at four binding pockets: P3, P1', P2' and P4'											
	Σ			L	V		Σ	M12.153	fibrinolytic peptidase (*Philodryas olfersii*)	15	72
Specificity at four binding pockets: P2, P1, P1' and P2'											
		H	N	C	λ			N10.002	intein-containing replicative DNA helicase precursor	158	75
		P	L	G	P			M03.010	Pz-peptidase A	13	71
Specificity at four binding pockets: P2, P1, P2' and P4'											
		Ω	-		λ		Σ	S01.213	coagulation factor XIa	12	75
Specificity at four binding pockets: P2, P1, P3' and P4'											
		+	-			Σ	Σ	S01.156	enteropeptidase	21	74
Specificity at four binding pockets: P2, P1', P2' and P4'											
		L		Q	P		D	C01.168	endopeptidase-A (*Hordeum*-type)	20	21
		λ		Q	Q		+	C01.010	vignain	80	76
Specificity at four binding pockets: P1, P1', P2' and P3'											
			L	G	P	A		M09.003	bacterial collagenase H	18	29
			S	D	Σ	Σ		M13.091	PHEX peptidase	14	71
			@	S	λ	Σ		S01.458	chymase 1 (*Mus musculus*-type)	44	78
			Σ	D	+	Ω		M12.005	procollagen C-peptidase	41	80
Specificity at four binding pockets: P1, P1', P2' and P4'											
			KR	λ	λ		G	S01.020	kallikrein-related peptidase 12	20	80
Specificity at four binding pockets: P1, P1', P3' and P4'											
			R	Σ		A	Ω	C14.044	metacaspase TbMCA2	11	67
Specificity at five binding pockets: P4, P3, P2, P1 and P1'											
L	λ	S	R	G				C14.026	paracaspase	17	64
R	-	λ	L	Σ				S08.063	site-1 peptidase	30	70
Ω	Σ	-	A	Σ				S26.016	SpsB signal peptidase	49	71
Specificity at five binding pockets: P4, P3, P2, P1 and P2'											
λ	V	W	F		G			S01.005	mast cell peptidase 4 (*Rattus*)	40	76
Specificity at five binding pockets: P4, P3, P2, P1 and P4'											
I	λ	G	G				Σ	C39.001	bacteriocin-processing peptidase	13	65
Specificity at five binding pockets: P4, P3, P2, P1' and P2'											
K	T	K		λ	I			M79.002	RCE1 peptidase (*Homo sapiens*-type)	22	53
Specificity at five binding pockets: P4, P2, P1', P2' and P4'											
Σ		Ω		P	I		Ω	A02.063	walleye dermal sarcoma virus retropepsin	25	73
Specificity at five binding pockets: P4, P1, P2', P2' and P4'											
Σ			Ω		L	λ	Ω	S14.001	peptidase Clp (type 1)	15	78
Specificity at five binding pockets: P3, P2, P1, P3' and P4'											
	λ	Σ	F			λ	Ω	A01.006	chymosin	20	81

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
Specificity at five binding pockets: P2, P1, P1', P2' and P3'											
		L	Q	G	P	E		C37.001	calicivirin	11	66
		-	Ω	C	F	Ω		N10.005	intein-containing DNA polymerase II large subunit DP2 precursor Mername-AA281	26	68
Specificity at six binding pockets: P4, P3, P2, P1, P2' and P3'											
E	λ	G	R		Σ	Σ		C50.001	separase (yeast-type)	12	69
P	I	+	F		R	L		S53.002	sedolisin-B	47	24
P	I	+	F		R	L		S53.005	kumamolisin-B	61	30
P	Σ	K	@		-	λ		S53.001	sedolisin	22	46
P	Σ	Σ	F		R	L		A01.002	pepsin B	25	60
Specificity at six binding pockets: P4, P3, P1', P2, P3' and P4'											
λ	λ			λ	λ	λ	λ	A22.001	presenilin 1	16	88
Specificity at six binding pockets: P4, P2, P1, P1', P2' and P3'											
G		Σ	P	@	-	Q		M03.002	neurolysin	128	69
-		λ	N	Ω	λ	T		A09.001	spumapepsin	11	79
Ω		λ	L	P	λ	Ω		A02.012	retropepsin (human T-cell leukemia virus)	12	65
Specificity at six binding pockets: P4, P2, P1, P1', P2' and P4'											
-		A	A	AE	Σ		M	S50.004	blotched snakehead birnavirus Vp4 peptidase	38	62
λ		GH-	N	C	λ		P	N10.004	intein-containing translation initiation factor IF-2 precursor	51	70
Specificity at six binding pockets: P4, P1, P2, P2', P3' and P4'											
Σ			M	Σ	L	λ	Ω	S14.003	peptidase Clp (type 3)	10	75
Specificity at six binding pockets: P3, P2, P1, P1', P2' and P4'											
	A	A	E	V	F		D	S01.267	streptogrisin E	23	46
Specificity at six binding pockets: P3, P2, P1', P2', P3' and P4'											
	K	L		Σ	S	K	Q	G01.001	scytalidoglutamic peptidase	37	45
Specificity at six binding pockets: P3, P1, P1', P2', P3' and P4'											
	Q		G	F	T	L	I	A24.001	type 4 prepilin peptidase 1	16	41
	Σ		E	Σ	-	Σ	Σ	M12.222	ADAMTS1 peptidase	13	76
Specificity at seven binding pockets: P4, P3, P2, P1, P1' P2' and P3'											
F	A	P	@	Σ	Σ	Σ		S08.112	ARA12 peptidase	53	38
G	F	S	P	F	R	Q		M03.011	tropolysin	34	53
W	E	λ	Q	Σ	S	Ω		S01.282	SplB peptidase (*Staphylococcus aureus*)	19	62
Specificity at seven binding pockets: P4, P3, P2, P1, P1' P3' and P4'											
λ	VY	HV	GN	Σ		E	-	N09.001	intein-containing V-type proton ATPase catalytic subunit A	45	48
Specificity at seven binding pockets: P4, P3, P2, P1, P2' P3 and P4'											
L	V	E	L		Y	L	V	A01.060	candidapepsin SAP2	22	55
L	V	E	λ		Y	L	V	A01.014	candidapepsin SAP1	26	66
Specificity at seven binding pockets: P4, P2, P1, P1' P2' P3 and P4'											
F		E	EQ	G	Ω	T	D	C24.003	sapovirus 3C-like peptidase	10	60
S		N	Y	P	I	V	Q	A02.013	bovine leukemia virus retropepsin	17	58
S		N	@	P	I	V	Q	A02.010	mouse mammary tumor virus retropepsin	13	44
S		N	@	Ω	I	V	Q	A02.002	HIV-2 retropepsin	30	75
Specificity at seven binding pockets: P4, P3, P1, P1' P2' P3 and P4'											
M	K		E	I	A	S	E	A25.001	gpr peptidase	32	41
Specificity at seven binding pockets: P3, P2, P1, P1', P2' P3 and P4'											

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
	K	K	G	A	ST	GS	IT	A24.017	PibD peptidase	18	50
	K	L	R	S	S	K	Q	C01.038	cathepsin P	10	20
	K	L	R	F	S	K	Q	C01.074	CPB peptidase	17	30
	K	L	R	F	S	K	Q	C01.100	cruzipain 2	26	54
	Y	-	R	E	A	E	A	A01.031	yapsin-2	16	46
Specificity at eight binding pockets											
A	L	F	Q	G	P	P	V	C03.011	coxsackievirus-type picornain 3C	10	35
I	A	Ω	Σ	Σ	F	G	S	S54.004	aarA-type peptidase	58	24
I	L	G	A	K	G	G	S	C39.007	lacticin 481 processing peptidase (*Lactococcus lactis*)	16	60
K	R	GR	NQ	C	DF	G	D	N10.006	intein-containing DNA polymerase II large subunit DP2 precursor Mername-AA282	18	54
L	A	A	R	R	G	A	G	A26.002	OmpP peptidase	11	47
L	V	E	L	λ	Y	L	V	A01.061	candidapepsin SAP3	21	24
L	V	E	L	L	Y	L	V	A01.064	candidapepsin SAP6	22	18
P	A	L	N	S	P	Ω	λ	N08.001	poliovirus capsid VP0-type self-cleaving protein	38	20
P	F	H	L	L	V	HY	S	A01.007	renin	10	37
S	F	L	Y	Q	V	S	T	A28.004	skin SASPase	15	65
S	Q	N	Y	P	I	V	Q	A02.004	equine infectious anaemia virus retropepsin	13	16
S	Q	N	Y	P	I	V	Q	A02.009	Mason-Pfizer leukemia virus retropepsin	18	52
S	Ω	N	Y	P	I	V	Q	A02.008	murine leukemia virus-type retropepsin	20	63
S	Ω	N	Y	P	I	V	Q	A02.015	Rous sarcoma virus retropepsin	18	52
V	R	F	Q	S	G	T	R	C04.003	tobacco vein mottling virus-type NIa peptidase	13	31
V	RY	F	Q	S	G	T	R	C04.004	tobacco etch virus NIa peptidase	14	53
-	Ω	L	Ω	λ	Ω	L	Ω	M34.001	anthrax lethal factor	12	78
λ	λ	N	P	P	V	P	P	M34.002	Zmp1 peptidase (*Clostridium difficile*-type)	13	22
Ω	P	Σ	Q	Σ	Ω	Ω	Σ	S9G.118	*Nysius* proteinase	11	67
No positive specificity											
			C					A01.001	pepsin A	417	93
								A01.003	gastricsin	50	93
			H	H	W			A01.009	cathepsin D	897	90
			P	H		H		A01.010	cathepsin E	1596	90
C	C	C	C	C	C	C	C	A01.012	rhizopuspepsin	249	94
								A01.016	aspergillopepsin I	157	94
								A01.017	endothiapepsin	11	87
								A01.018	saccharopepsin	14	94
								A01.020	phytepsin	62	92
								A01.040	nepenthesin	19	95
								A01.041	memapsin-1	33	92
								A01.053	necepsin-1	133	93
								A01.068	necepsin-2	45	91
								A01.071	pepsin A5 (*Homo sapiens*)	10	93
		KP	KT		R			A02.001	HIV-1 retropepsin	1045	90
KR	KR	KR						A32.002	RC1339 g.p. (*Rickettsia conorii*)	785	92
								C01.005	stem bromelain	26	94
								C01.007	actinidain	19	91
								C01.008	asclepain A	10	96

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
								C01.013	cathepsin X	46	90
								C01.026	ananain	17	58
								C01.033	cathepsin L1 (*Fasciola* sp.)	196	93
								C01.040	cathepsin H	48	90
								C01.046	falcipain-2	152	92
								C01.062	cathepsin B-like peptidase (platyhelminth)	54	93
								C01.063	falcipain-3	126	92
								C01.070	dipeptidyl-peptidase I	34	89
								C01.084	bleomycin hydrolase (animal)	19	59
■	■	■			■	■	■	C01.086	aminopeptidase C	20	51
								C01.092	cathepsin L1 (arthropod-type)	12	95
								C01.097	phytolacain	13	95
								C01.101	cathepsin B-like peptidase, nematode	47	93
								C01.129	cathepsin L3 (*Fasciola* sp.)	72	90
								C01.130	CSCP3 peptidase (*Clonorchis*-type)	56	93
								C01.142	cathepsin L2 (*Fasciola* sp.)	54	92
								C02.001	calpain-1	106	91
								C02.002	calpain-2	170	91
								C02.004	calpain-3	19	93
								C9G.001	cathepsin N	14	92
								G01.002	aspergilloglutamic peptidase	68	93
■	■	■						M01.001	aminopeptidase N	42	58
■	■	■						M01.002	lysyl aminopeptidase (bacteria)	22	75
■	■	■						M01.004	leukotriene A4 hydrolase	14	52
								M01.005	alanyl aminopeptidase (bacterial-type)	17	56
■	■	■						M01.009	aminopeptidase N (actinomycete-type)	11	48
■	■	■						M01.010	cytosol alanyl aminopeptidase	56	68
■	■	■						M01.011	cystinyl aminopeptidase	11	82
■	■	■						M01.013	aminopeptidase N (insect)	14	49
■	■	■						M01.016	aminopeptidase Ey	18	73
■	■	■						M01.022	arginyl aminopeptidase-like 1	14	50
■	■	■						M01.024	ERAP2 aminopeptidase	23	89
■	■	■						M01.026	aminopeptidase Q	17	70
								M10.005	matrix metallopeptidase-3	180	90
								M10.014	membrane-type matrix metallopeptidase-1	132	92
								M10.016	membrane-type matrix metallopeptidase-3	20	92
								M10.017	membrane-type matrix metallopeptidase-4	27	92
								M10.024	membrane-type matrix metallopeptidase-6	50	93
								M10.056	aeruginolysin	13	93
								M10.057	mirabilysin	32	95
								M12.002	meprin alpha subunit	776	92
								M12.033	LAST_MAM peptidase (*Limulus*-type)	432	91
								M12.137	BHRa hemorrhagin (*Bitis arietans*)	14	89
								M12.164	lebetase	14	89
								M12.216	jerdohagin (*Trimeresurus jerdonii*)	15	96
								M12.221	ADAMTS4 peptidase	59	91
								M12.310	metallopeptidase MTP-1	18	92

P4	P3	P2	P1	P1'	P2'	P3'	P4'	*MEROPS* ID	Peptidase name	Total cleavages	Reliability
								M12.327	leucurolysin-a	25	90
								M13.001	neprilysin	123	92
								M13.002	endothelin-converting enzyme 1	34	94
								M13.011	MEP peptidase (nematode)	57	91
								M16.001	pitrilysin	23	94
								M16.002	insulysin	101	88
								M16.009	eupitrilysin	60	92
								M16.016	YMXG peptidase	40	90
								M16.017	plinsulysin	17	94
■	■	■						M17.015	aminopeptidase pepZ (*Staphylococcus* sp.)	16	47
■	■	■			■	■	■	M20.013	Xaa-methyl-His dipeptidase	13	76
■	■	■						M28.001	aminopeptidase Y	28	41
■	■	■						M28.002	aminopeptidase Ap1	66	93
■	■	■						M28.006	Mername-AA063 peptidase	19	37
■	■	■						M28.022	leucine aminopeptidase-1 (*Aspergillus fumigatus*)	10	50
								M35.001	penicillolysin	20	93
								M35.002	deuterolysin	22	91
								M41.001	FtsH peptidase	24	92
								M41.013	Ftsh peptidase (*Thermus*-type)	32	91
■	■	■						M42.003	PhTET1 aminopeptidase	10	44
■	■	■						M42.004	PhTET2 aminopeptidase	23	83
■	■							M49.001	dipeptidyl-peptidase III	24	81
■	■	■						M61.001	glycyl aminopeptidase	26	84
■	■	■						M9A.011	neuron-specific aminopeptidase	10	48
								M9G.026	dactylysin	23	85
								S01.122	brachyurin-C	36	88
K								S01.133	cathepsin G	464	89
								S01.134	myeloblastin	52	90
								S01.142	duodenase	19	77
								S01.157	chymotrypsin C	36	93
								S01.162	kallikrein-related peptidase 3	83	88
								S01.422	fibrinolytic enzyme A (Annelida-type)	19	90
								S08.001	subtilisin Carlsberg	43	92
								S08.003	subtilisin lentus	16	94
								S08.008	Mername-AA053 peptidase	16	95
								S08.009	subtilisin Ak1	12	63
								S08.010	M-peptidase	14	95
								S08.014	ALE1 peptidase	16	95
								S08.019	lactocepin I	102	92
								S08.022	bpr peptidase (*Dichelobacter nodosus*)	12	96
								S08.024	dentilisin	17	85
								S08.028	high alkaline protease (*Alkaliphilus transvaalensis*)	28	94
								S08.034	subtilisin BPN'	17	93
								S08.051	aqualysin 1	11	93
								S08.052	cerevisin	20	82



P4	P3	P2	P1	P1'	P2'	P3'	P4'	MEROPS ID	Peptidase name	Total cleavages	Reliability
								S08.054	peptidase K	47	94
								S08.057	thermomycolin	28	95
								S08.092	cucumisin	31	89
								S08.116	lactocepin-3	158	93
								S08.135	tengconlysin	19	94
								S08.142	collagenolytic endopeptidase (*Geobacillus* sp. MO-1)	11	95
								S08.148	keratinase (*Doratomyces microsporus*)	19	95
								S08.151	SlSBT3 g.p. (*Solanum lycopersicum*)	19	91
								S09.004	acylaminoacyl-peptidase	14	59
								S09.072	puromycin hydrolase	17	56
								S10.006	Mername-AA083 peptidase	72	94
								S10.010	serine carboxypeptidase Z (*Absidia zachae*) and similar	16	78
								S10.014	carboxypeptidase O	89	78
								S12.001	D-Ala-D-Ala carboxypeptidase B	11	22
								S12.002	aminopeptidase DmpB	17	31
								S12.003	alkaline D-peptidase	12	26
								S16.001	Lon-A peptidase	21	92
								S41.007	ctpB peptidase	18	93
								S46.001	dipeptidyl-peptidase 7 (*Porphyromonas gingivalis*-type)	21	69
								S53.003	tripeptidyl-peptidase I	45	81
								S53.010	grifolisin	13	93
								T01.002	archaean proteasome, beta component	44	93
								T01.005	bacterial proteasome, beta component	31	93
								T01.012	proteasome catalytic subunit 3	24	90
								XM02-001	angiotensin-converting enzyme compound peptidase	24	82
								XM12-001	meprin A complex peptidase	71	94
				C				XT01-001	20 S constitutive proteasome peptidase complex (eukaryote)	468	94
								XT01-002	26 S proteasome peptidase complex (eukaryote)	67	91
C	C	C	C		CW		C	XT01-003	20 S immunoproteasome peptidase complex (eukaryote)	228	94

pocket, then it is shown as white letters on a black background. Obviously, with some amino acids that occur less frequently (Cys and Trp, for example), the number of cleavages should be taken into consideration before any conclusions are drawn about negative preferences by the peptidase.

The example in Fig. 2 is based on 125 cleavages. However, 80 of these cleavages are derived from one reference [44] in which the substrates based on Abz-GFSPFRQ-EDDnp were synthesized to help distinguish thimet oligopeptidase from its close relative neurolysin. It is immediately obvious that the preferences shown in binding pockets S4–S3′ (Gly, Phe, Ser, Pro, Phe, Arg, Gln) reflect the residues that were kept constant during the residue-scanning experiments. This example serves as a cautionary note to show that apparent preferences may reflect the methodology employed rather than represent the true specificity of the peptidase.

For each peptidase where substrate cleavages are known there is a page dedicated to listing the substrates and the known cleavage positions. Fig. 3 shows sections of the table of substrates for thimet oligopeptidase. For each cleavage, the substrate name is given, the UniProt accession if the substrate is from a naturally occurring protein, the residue range of the substrate with respect to the Uniprot entry, the cleavage site, the nature of the substrate ("cleavage type"), the evidence for the cleavage position (if known), a reference and a cross-reference to the CutDB database [14].

There are also displays for protein substrates. One important question when considering whether a particular cleavage is

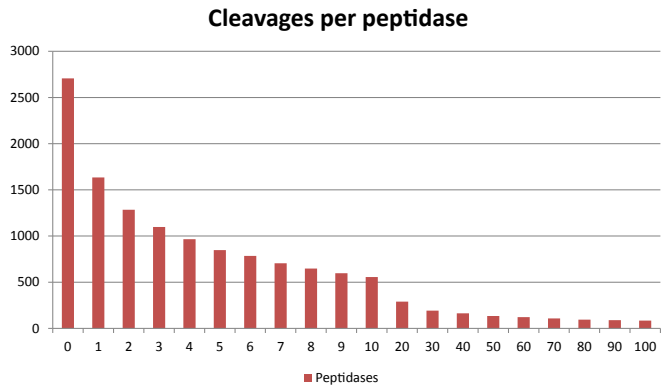


Fig. 1. Cleavages per peptidase. The bar chart shows the number of known substrate cleavages per peptidase on the Y axis and the count of peptidases with this number of cleavages on the X axis.

physiologically relevant is whether the cleavage site is conserved in close homologues. A cleavage site that is not conserved is unlikely to be physiologically relevant, although it may be pathologically relevant in the species in which it occurs. Conservation should be in terms of which residues are accepted in the binding pockets of the peptidase and not just sequence conservation in the substrate orthologues. This is only possible if a large number of substrate cleavages are known for the peptidase in question. Fig. 4 shows part of the alignment for orthologues of the Ebola virus envelope glycoprotein. The known cleavage is from the Zaire strain (UniProt P87671) and cleavage is by ADAM17 (MEROPS ID M12.217) at residue 637; the cleavage results in shedding of the ectodomain which circulates in the blood of the patient and may interfere with antibodies thus helping to prolong infection [45]. The alignment is dynamically generated from the sequences clustered in the UniRef50 database entry that contains this sequence. The UniRef50 database entry contains sequences that share 50% or more sequence identity [46]. Alignments are generated using MUSCLE [47]. The sequence containing the known cleavage site is highlighted in green. Residues in the range P4–P4′ are highlighted in pink if they are identical to that from the Zaire strain; substituted residues are highlighted in orange if the amino acid from another ADAM17 substrate is known to occupy the same binding pocket; and substituted residues are shown as white on black if the amino acid is not known to occupy the same binding pocket from any ADAM17 substrate ("unacceptable replacements").

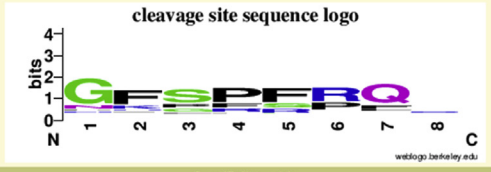
In this example, in one sequence the residues around the cleavage site have not been determined and are replaced by Xs, and in sequences from a number of strains of the virus the P1′ residue is replaced by His and the P2′ residue by Asp or Asn, which have not been observed in any of the 59 other substrates for ADAM17. The possible conclusions are: 1) ADAM17 is not the physiological peptidase that performs this cleavage, 2) some Ebola virus envelope glycoproteins are processed by a different peptidase, 3) the specificity of ADAM17 has not been fully explored and His is permissible in P1′ and Asp or Asn in P2′, or 4) it doesn't matter if the cleavage is inefficient in some cases provided some processing occurs.

3.5. A service to test conservation of cleavage sites

The production of this cleavage collection has allowed the installation of a service on the MEROPS website to help researchers assess whether or not a particular cleavage is physiologically relevant. Following on from the display described above, the philosophy behind this service is that a cleavage site in a protein substrate is most likely to be physiologically relevant if it is conserved. The user can submit a list of cleavage sites in a file the structure of

Cleavage pattern

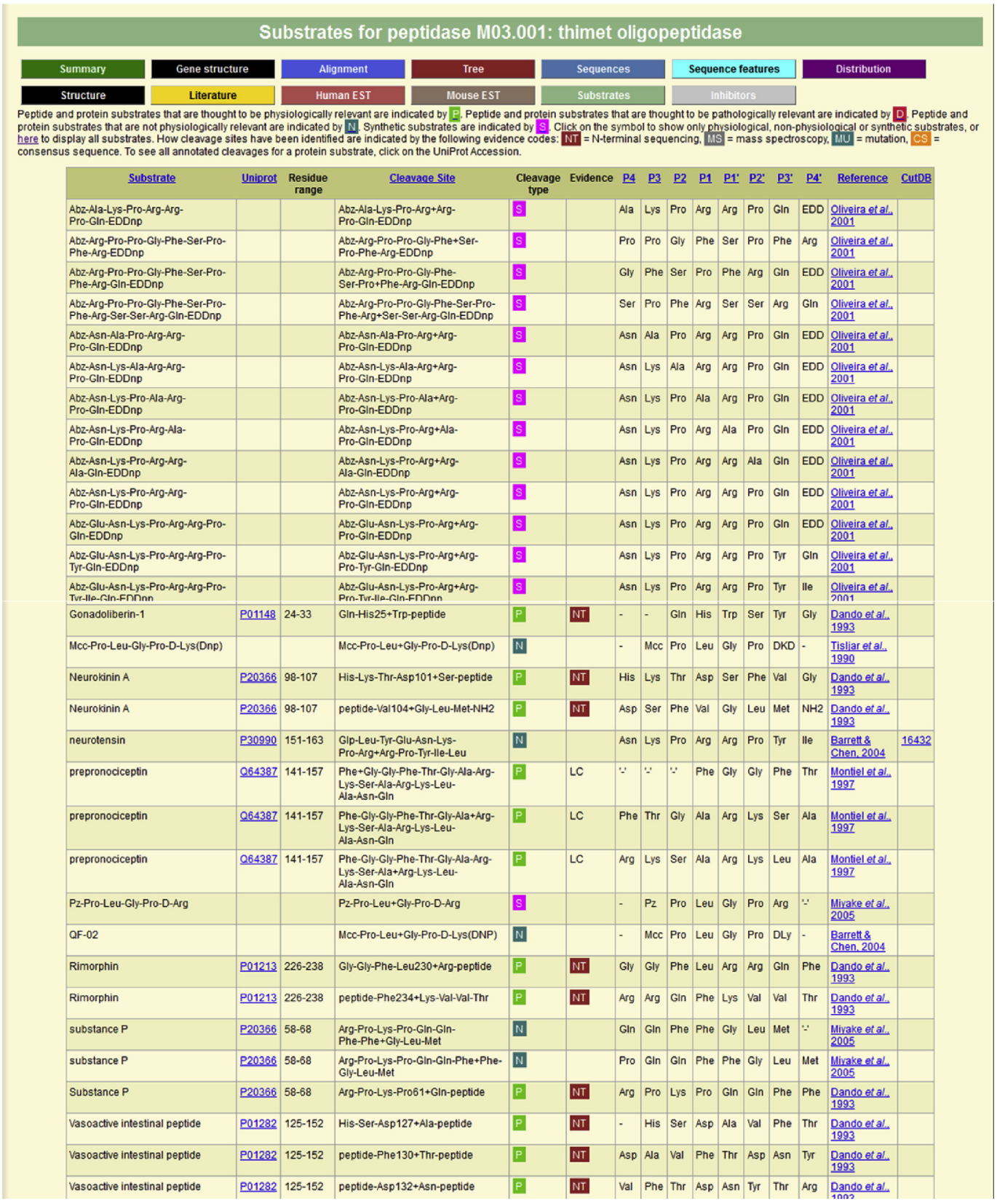
G/Fk/spg/Pfr+Fsr/rp/Q/- (based on 124 cleavages)



Specificity matrix

Amino acid	P4	P3	P2	P1	P1'	P2'	P3'	P4'
Gly	65	2	16	1	6	3	1	2
Pro	5	6	18	57	0	31	2	1
Ala	1	5	2	4	3	3	1	2
Val	1	1	4	1	0	3	2	0
Leu	0	2	3	6	1	5	10	0
Ile	0	2	1	0	4	5	3	3
Met	2	0	0	1	0	0	2	1
Phe	3	53	7	20	56	2	14	4
Tyr	1	3	0	2	2	3	4	1
Trp	0	0	0	0	1	0	0	0
Ser	2	3	56	2	20	4	2	1
Thr	0	1	2	0	1	0	1	4
Cys	0	0	0	0	0	0	0	0
Asn	11	1	1	3	2	1	2	4
Gln	1	3	4	1	2	2	68	5
Asp	3	1	1	4	2	1	1	1
Glu	1	1	1	1	1	1	4	2
Lys	2	14	2	2	3	2	0	0
Arg	5	7	3	17	19	55	5	9
His	1	2	2	2	1	3	0	1

Fig. 2. Example of a specificity logo and distribution matrix. The specificity logo and distribution matrix are shown for thimet oligopeptidase. In the logo, the taller the character the greater the preference in substrate binding pockets S4 to S4′ (numbered as 1 to 8 on the X axis). In the specificity matrix the number of times an amino acid occurs in the residue range P4 to P4′ in substrates is shown. The brighter the green highlighting, the greater the preference for an amino acid in that position. An amino acid that has not been observed to occupy a specific binding pocket is shown as white text on a black background. Amino acids are ordered so that amino acids with similar properties are grouped together.



Substrates for peptidase M03.001: thimet oligopeptidase

Summary | Gene structure | Alignment | Tree | Sequences | Sequence features | Distribution
Structure | Literature | Human EST | Mouse EST | Substrates | Inhibitors

Peptide and protein substrates that are thought to be physiologically relevant are indicated by P. Peptide and protein substrates that are thought to be pathologically relevant are indicated by D. Peptide and protein substrates that are not physiologically relevant are indicated by N. Synthetic substrates are indicated by S. Click on the symbol to show only physiological, non-physiological or synthetic substrates, or here to display all substrates. How cleavage sites have been identified are indicated by the following evidence codes: NT = N-terminal sequencing, MS = mass spectroscopy, MU = mutation, CS = consensus sequence. To see all annotated cleavages for a protein substrate, click on the UniProt Accession.

Substrate	Uniprot	Residue range	Cleavage Site	Cleavage type	Evidence	P4	P3	P2	P1	P1'	P2'	P3'	P4'	Reference	CutDB
Abz-Ala-Lys-Pro-Arg-Arg-Pro-Gln-EDDnp			Abz-Ala-Lys-Pro-Arg+Arg-Pro-Gln-EDDnp	S		Ala	Lys	Pro	Arg	Arg	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-EDDnp			Abz-Arg-Pro-Pro-Gly-Phe+Ser-Pro-Phe-Arg-EDDnp	S		Pro	Pro	Gly	Phe	Ser	Pro	Phe	Arg	Oliveira et al., 2001	
Abz-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-Gln-EDDnp			Abz-Arg-Pro-Pro-Gly-Phe-Ser-Pro+Phe-Arg-Gln-EDDnp	S		Gly	Phe	Ser	Pro	Phe	Arg	Gln	EDD	Oliveira et al., 2001	
Abz-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg-Ser-Ser-Arg-Gln-EDDnp			Abz-Arg-Pro-Pro-Gly-Phe-Ser-Pro-Phe-Arg+Ser-Ser-Arg-Gln-EDDnp	S		Ser	Pro	Phe	Arg	Ser	Ser	Arg	Gln	Oliveira et al., 2001	
Abz-Asn-Ala-Pro-Arg-Arg-Pro-Gln-EDDnp			Abz-Asn-Ala-Pro-Arg+Arg-Pro-Gln-EDDnp	S		Asn	Ala	Pro	Arg	Arg	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Asn-Lys-Ala-Arg-Arg-Pro-Gln-EDDnp			Abz-Asn-Lys-Ala-Arg+Arg-Pro-Gln-EDDnp	S		Asn	Lys	Ala	Arg	Arg	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Asn-Lys-Pro-Ala-Arg-Pro-Gln-EDDnp			Abz-Asn-Lys-Pro-Ala+Arg-Pro-Gln-EDDnp	S		Asn	Lys	Pro	Ala	Arg	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Asn-Lys-Pro-Arg-Ala-Pro-Gln-EDDnp			Abz-Asn-Lys-Pro-Arg+Ala-Pro-Gln-EDDnp	S		Asn	Lys	Pro	Arg	Ala	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Asn-Lys-Pro-Arg-Arg-Ala-Gln-EDDnp			Abz-Asn-Lys-Pro-Arg+Arg-Ala-Gln-EDDnp	S		Asn	Lys	Pro	Arg	Arg	Ala	Gln	EDD	Oliveira et al., 2001	
Abz-Asn-Lys-Pro-Arg-Arg-Pro-Gln-EDDnp			Abz-Asn-Lys-Pro-Arg+Arg-Pro-Gln-EDDnp	S		Asn	Lys	Pro	Arg	Arg	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Glu-Asn-Lys-Pro-Arg-Arg-Pro-Gln-EDDnp			Abz-Glu-Asn-Lys-Pro-Arg+Arg-Pro-Gln-EDDnp	S		Asn	Lys	Pro	Arg	Arg	Pro	Gln	EDD	Oliveira et al., 2001	
Abz-Glu-Asn-Lys-Pro-Arg-Arg-Pro-Tyr-Gln-EDDnp			Abz-Glu-Asn-Lys-Pro-Arg+Arg-Pro-Tyr-Gln-EDDnp	S		Asn	Lys	Pro	Arg	Arg	Pro	Tyr	Gln	Oliveira et al., 2001	
Abz-Glu-Asn-Lys-Pro-Arg-Arg-Pro-Tyr-Ile-Gln-EDDnp			Abz-Glu-Asn-Lys-Pro-Arg+Arg-Pro-Tyr-Ile-Gln-EDDnp	S		Asn	Lys	Pro	Arg	Arg	Pro	Tyr	Ile	Oliveira et al., 2001	
Gonadoliberin-1	P01148	24-33	Gln-His25+Trp-peptide	P	NT	-	-	Gln	His	Trp	Ser	Tyr	Gly	Dando et al., 1993	
Mcc-Pro-Leu-Gly-Pro-D-Lys(Dnp)			Mcc-Pro-Leu+Gly-Pro-D-Lys(Dnp)	N		-	Mcc	Pro	Leu	Gly	Pro	DKD	-	Tisljar et al., 1990	
Neurokinin A	P20366	98-107	His-Lys-Thr-Asp101+Ser-peptide	P	NT	His	Lys	Thr	Asp	Ser	Phe	Val	Gly	Dando et al., 1993	
Neurokinin A	P20366	98-107	peptide-Val104+Gly-Leu-Met-NH2	P	NT	Asp	Ser	Phe	Val	Gly	Leu	Met	NH2	Dando et al., 1993	
neurotensin	P30990	151-163	Glp-Leu-Tyr-Glu-Asn-Lys-Pro-Arg+Arg-Pro-Tyr-Ile-Leu	N		Asn	Lys	Pro	Arg	Arg	Pro	Tyr	Ile	Barrett & Chen, 2004	16432
prepronociceptin	Q64387	141-157	Phe+Gly-Gly-Phe-Thr-Gly-Ala-Arg-Lys-Ser-Ala-Arg-Lys-Leu-Ala-Asn-Gln	P	LC	'-'	'-'	'-'	Phe	Gly	Gly	Phe	Thr	Montiel et al., 1997	
prepronociceptin	Q64387	141-157	Phe-Gly-Gly-Phe-Thr-Gly-Ala+Arg-Lys-Ser-Ala-Arg-Lys-Leu-Ala-Asn-Gln	P	LC	Phe	Thr	Gly	Ala	Arg	Lys	Ser	Ala	Montiel et al., 1997	
prepronociceptin	Q64387	141-157	Phe-Gly-Gly-Phe-Thr-Gly-Ala-Arg-Lys-Ser-Ala+Arg-Lys-Leu-Ala-Asn-Gln	P	LC	Arg	Lys	Ser	Ala	Arg	Lys	Leu	Ala	Montiel et al., 1997	
Pz-Pro-Leu-Gly-Pro-D-Arg			Pz-Pro-Leu+Gly-Pro-D-Arg	S		-	Pz	Pro	Leu	Gly	Pro	Arg	'-'	Miyake et al., 2005	
QF-02			Mcc-Pro-Leu+Gly-Pro-D-Lys(DNP)	N		-	Mcc	Pro	Leu	Gly	Pro	DLy	-	Barrett & Chen, 2004	
Rimorphin	P01213	226-238	Gly-Gly-Phe-Leu230+Arg-peptide	P	NT	Gly	Gly	Phe	Leu	Arg	Arg	Gln	Phe	Dando et al., 1993	
Rimorphin	P01213	226-238	peptide-Phe234+Lys-Val-Val-Thr	P	NT	Arg	Arg	Gln	Phe	Lys	Val	Val	Thr	Dando et al., 1993	
substance P	P20366	58-68	Arg-Pro-Lys-Pro-Gln-Gln-Phe-Phe+Gly-Leu-Met	N		Gln	Gln	Phe	Phe	Gly	Leu	Met	'-'	Miyake et al., 2005	
substance P	P20366	58-68	Arg-Pro-Lys-Pro-Gln-Gln-Phe+Phe-Gly-Leu-Met	N		Pro	Gln	Gln	Phe	Phe	Gly	Leu	Met	Miyake et al., 2005	
Substance P	P20366	58-68	Arg-Pro-Lys-Pro61+Gln-peptide	P	NT	Arg	Pro	Lys	Pro	Gln	Gln	Phe	Phe	Dando et al., 1993	
Vasoactive intestinal peptide	P01282	125-152	His-Ser-Asp127+Ala-peptide	P	NT	-	His	Ser	Asp	Ala	Val	Phe	Thr	Dando et al., 1993	
Vasoactive intestinal peptide	P01282	125-152	peptide-Phe130+Thr-peptide	P	NT	Asp	Ala	Val	Phe	Thr	Asp	Asn	Tyr	Dando et al., 1993	
Vasoactive intestinal peptide	P01282	125-152	peptide-Asp132+Asn-peptide	P	NT	Val	Phe	Thr	Asp	Asn	Tyr	Thr	Arg	Dando et al., 1993	

Fig. 3. Example of a substrate page. Part of the substrates page for thimet oligopeptidase is shown. For each substrate the following are shown: name; a cross-reference and link to the entry in the UniProt database where appropriate; the residue range of the substrate as used in the experiment with reference to the numbering in the UniProt entry; a description of the cleavage where the scissile bond is represented by the symbol '+'; whether the cleavage is physiological, non-physiological, pathological or in a synthetic substrate; the evidence by which the cleavage site was determined; the residues occupying residues P4 to P4′ in the substrate; the source reference; and a cross-reference and link to the CutDB database [14]. By default substrates are listed alphabetically, but the order can be changed by clicking the column heading. It is possible to filter the results for physiological, nonphysiological, pathological or cleavages in synthetic substrates by clicking on the appropriate letter in the table legend.


63 64 65 66 67
12A 45678901234567890123456789012345678901234567890123456
M12.217 <- -------------†--------------------------------------->
A0A0D5W927 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0D5W952 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A7PCD5 MG- DQIIHDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXFCICKFVF
A0A0D5W8P4 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVXIAXIALFCICKFVF
A8TSF7 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A8TSH3 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
L7QI57 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A8TSI4 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVL
A8TSI9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A8TSH6 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A8TSH9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
G8DB49 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
P87671 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
Q05320 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A9QPL9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
L7QHR7 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A7P6T9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A096YGX4 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
P87666 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
L7QHT1 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
I0DHA9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
L7QHV7 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
G0YZP1 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
O11457 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
L7QI08 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A068J4A4 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A0VB57 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A1EDT0 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0D5W8E1 MG- XQIIXDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0D5W9K1 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0D5W994 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0D5W896 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A068J419 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0D5W882 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A0UG51 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
X5HMX4 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
X5H5A9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A0UDX5 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
A0A0A0ULB9 MG- DQIIHDFVDKTLPDQGDNDNWWTGWRQWIPAGIGVTGVIIAVIALFCICKFVF
B8XCN9 MG- DQIIHDFVDNNLPNQNDGSNWWTGWKQWVPAGIGITGVIIAIIALLCICKFML
Q66810 MG- DQIIHDFVDNNLPNQNDGSNWWTGWKQWVPAGIGITGVIIAIIALLCICKFML
B8XCN0 MV- DQIIHDFIDKPLPDQTDNDNWWTGWRQWVPAGIGITGVIIAVIALLCICKFLL
R4QGV6 MV- DQIIHDFIDKPLPDQTDNDNWWTGWRQWVPAGIGITGVIIAVIALLCICKFLL
R4QRC0 MV- DQIIHDFIDKPLPDQTDNDNWWTGWRQWVPAGIGITGVIIAVIALLCICKFLL
R4P4N7 MG- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
R4QJ45 MG- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
I7F2J9 MG- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
Q7T9D9 MG- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
Q66798 ME- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
Q66814 ME- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
C4PK58 ME- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
B0LPL7 ME- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
M4JBD3 ME- NQIIHDFIDNPLPNQDNDDNWWTGWRQWIPAGIGITGIIIAIIALLCVCKLLC
C6G8E4 MGS NQIRHDFIDNPLPNHGDDLILWTGWRQWIPAGIGIIGVIIAIIALLCICKILC
C6G8F2 MGS NQIKHDFIDNPLPDHGDDLNLWTGWRQWIPAGIGIIGVIIAIIALLCICKILC
C6G8D6 MGS NQIKHDFIDNPLPDHGDDLNLWTGWRQWIPAGIGIIGVIIAITALLCICKILC
L7REV7 MGS NQIKHDFIDNPLPDHGDDLNLWTGWRQWIPAGIGIIGVIIAITALLCICKILC
Q91DD8 MGS NQIKHDFIDNPLPDHGDDLNLWTGWRQWIPAGIGIIGVIIAIIALLCICKILC
Q66799 MGS NQIKHDFIDNPLPDHGDDLNLWTGWRQWIPAGIGIIGVIIAIIALLCICKILC
Q89853 MGS NQIKHDFIDNPLPDHGDDLNLWTGWRQWIPAGIGIIGVIIAIIALLCICKILC

which is shown in Table 5. This should be a text file created with software such as Notepad and not a document file created with a word processing package such as Word. The data required per line are: MEROPS identifier of the peptidase performing the cleavage, the UniProt accession for the substrate protein, and the position of the residue occupying the S1 binding pocket (the residue after which cleavage occurs) taken from the full coding sequence in the UniProt entry. For each line, an alignment is generated from the sequences in the UniRef50 entry containing the UniProt accession of the substrate protein, and the number of unacceptable replacements is counted. A file is generated and is E-mailed to the user. An example of a results file is shown in Table 6. For each line in the table the following are included: the MEROPS identifier of the peptidase; the total number of cleavages known for that peptidase; the UniProt accession of the substrate protein; the number of homologues aligned from the UniRef50 entry; the cleavage position; the number of unacceptable replacements at P4, P3, P2, P1, P1′, P2′, P3′ and P4′; and a URL to display the substrate alignment at the MEROPS website (to conserve space, the URL column is not shown in Table 6). The number of known substrate cleavages for the peptidase is returned because this will help the user assess the reliability of the analysis: the more cleavages the more reliable the analysis. Similarly, the number of sequences in the alignment will help the user assess the results: if the alignment contains many sequences and the cleavage site is well conserved, then the likelihood that the cleavage site is physiologically relevant is greater. If on the other hand the cleavage site is not conserved, and there are many sequences in the alignment, then it is much more likely that the cleavage is not physiologically relevant.

If the number of unacceptable replacements in positions P4–P4′ is zero, then the cleavage site is extremely well conserved and is likely to represent a physiological cleavage by the peptidase concerned. If the number of unacceptable replacements is small, then the cleavage may still be physiological. The most common explanation is that the UniRef50 entry on which the alignment is based contains two or more very closely related paralogues, one of which is a physiological substrate, but the other is not. A second possibility is that the sequences in the UniRef50 entry represent species variants of only one protein, but one or more is derived from genomic sequencing and is either a fragment (perhaps because the initiating methionine has been misidentified or an anomalous frame-shift has been introduced which truncates the C-terminus), or is missing the exon which codes for all or part of the cleavage site. The user is advised to consult the alignment via the URL provided to check. If the number of unaccepted replacements is high, then the cleavage site is not conserved and the cleavage is unlikely to be physiologically significant. However, the user should check the number of cleavages for the peptidase: if this is less than 40, then it is possible that there is not enough variation amongst the known cleavage site sequences to account for the substitutions that have occurred. It is possible that many homologues in the alignment have the same replacement which has not been observed in any substrate for the peptidase, simply because not enough substrates have been found.

The user should also be aware that if there is a high number of unacceptable replacements in one position, this requires further investigation, because substitution to a rare amino acid may have taken place. For a peptidase with no preference in a binding pocket, for example S3′, almost any amino acid can occupy the P3′ position.

Table 5
An example of a file for submission to the Analyse Substrates service.

MEROPS ID	UniProt	Cleavage position
A01.004	P05067	671
A01.009	P05067	705
A01.009	P05067	713
A01.009	P05067	714
A01.009	P05067	719
A01.009	P05067	720
A01.041	P05067	690
A01.041	P05067	691
A22.001	P05067	711
A22.001	P05067	713
A22.001	P05067	714
C01.060	P05067	704
C01.060	P05067	708
C01.060	P05067	711
C01.084	P05067	685
C01.084	P05067	685
C01.084	P05067	685
C01.084	P05067	689
C01.084	P05067	690
C01.084	P05067	690
C14.003	P05067	739
C14.005	P05067	672
C14.005	P05067	739
M02.001	P05067	711
M10.003	P05067	687
M10.003	P05067	705
M10.003	P05067	706
M10.004	P05067	687
M10.004	P05067	691
M10.004	P05067	694
M10.004	P05067	701
M10.004	P05067	704
M10.004	P05067	705
M10.014	P05067	579
M10.014	P05067	687
M10.016	P05067	463
M10.016	P05067	579
M10.016	P05067	622
M10.016	P05067	685
M10.017	P05067	685
M10.017	P05067	687

In the alignment of substrate homologues, the count of unacceptable replacements is therefore likely to be small. However, if no substrate is known with cysteine in P3′, for example, yet many homologues of a known substrate have a replacement cysteine in P3′, then the unacceptable replacement count will be high. In such an example, it is simply not known if cysteine is acceptable in P3′, so it is impossible to say if the cleavage is physiological or not. The user should check the alignment for such a replacement.

I would like to reiterate that a cleavage site that is not conserved could still be important pathologically. The user is advised to check the preferences for the peptidase on the peptidase summary page in MEROPS to see if, for example, a large number of mismatches in P4′ is relevant.

This service is known as “Analyse Substrates” and is accessed from the main menu on the website. It is particularly useful for proteomics studies in which hundreds of potential physiological substrates are found. The limitations on the service are that the maximum number of cleavages per file uploaded is 5000 and the maximum file size is 10 Mbytes.

Fig. 4. Example of a substrate alignment. Part of the alignment for orthologues of the Ebola virus envelope glycoprotein is shown, with the known cleavage of the glycoprotein from the Zaire strain (UniProt P87671) by ADAM17 (MEROPS ID M12.217) at residue 637 highlighted in green. Residues in the range P4–P4′ are highlighted in pink if they are identical to that from the Zaire strain; substituted residues are highlighted in orange if the amino acid from another ADAM17 substrate is known to occupy the same binding pocket; and substituted residues are shown as white on black if the amino acid is not known to occupy the same binding pocket from any ADAM17 substrate.

Table 6
Results from the Analyse Substrates service.

MEROPS identifier	Total cleavages known	Substrate UniProt accession	Homologues	Cleaved at	P4 count	P3 count	P2 count	P1 count	P1′ count	P2′ count	P3′ count	P4′ count
A01.004	24	P05067	352	671	10	7	25	5	0	0	0	1
A01.009	897	P05067	352	705	0	0	0	0	0	0	1	0
A01.009	897	P05067	352	713	0	0	1	1	6	5	5	5
A01.009	897	P05067	352	714	0	1	1	6	5	5	5	5
A01.009	897	P05067	352	719	0	0	0	0	0	1	1	1
A01.009	897	P05067	352	720	0	0	0	0	1	1	1	3
A01.041	33	P05067	352	690	0	2	2	1	1	3	1	10
A01.041	33	P05067	352	691	2	2	1	1	1	1	1	1
A22.001	16	P05067	352	711	1	0	0	1	1	2	6	6
A22.001	16	P05067	352	713	4	1	1	2	6	5	5	5
A22.001	16	P05067	352	714	0	1	1	6	5	5	5	8
C01.060	632	P05067	352	704	0	0	0	0	0	0	0	1
C01.060	632	P05067	352	708	0	0	0	1	0	0	0	1
C01.060	632	P05067	352	711	1	0	0	0	1	1	5	5
C01.084	19	P05067	352	685	1	0	30	1	6	1	3	3
C01.084	19	P05067	352	685	1	0	30	1	6	1	3	3
C01.084	19	P05067	352	685	1	0	30	1	6	1	3	3
C01.084	19	P05067	352	689	0	1	3	2	2	6	4	2
C01.084	19	P05067	352	690	1	3	3	1	5	4	4	8
C01.084	19	P05067	352	690	1	3	3	1	5	4	4	8
C14.003	651	P05067	352	739	1	1	4	4	3	3	3	3
C14.005	201	P05067	352	672	7	5	4	16	0	0	0	0
C14.005	201	P05067	352	739	1	1	4	4	3	3	3	3
M02.001	5	P05067	352	711	1	1	4	1	1	2	173	173
M10.003	3417	P05067	352	687	2	0	0	0	0	1	1	1
M10.003	3417	P05067	352	705	0	0	0	0	0	0	1	0
M10.003	3417	P05067	352	706	0	0	0	0	0	1	0	0
M10.004	369	P05067	352	687	2	0	0	0	0	1	1	1
M10.004	369	P05067	352	691	0	1	1	1	1	1	1	1
M10.004	369	P05067	352	694	0	0	0	0	0	0	0	0
M10.004	369	P05067	352	701	0	0	0	0	0	0	0	0
M10.004	369	P05067	352	704	0	0	0	0	0	0	0	1
M10.004	369	P05067	352	705	0	0	0	0	0	0	1	0
M10.014	132	P05067	352	579	0	0	0	0	0	0	0	0
M10.014	132	P05067	352	687	11	0	0	0	0	1	1	1
M10.016	20	P05067	352	463	0	0	0	0	0	0	0	9
M10.016	20	P05067	352	579	7	0	18	0	2	4	0	1
M10.016	20	P05067	352	622	13	6	19	2	1	6	11	6
M10.016	20	P05067	352	685	0	0	11	1	5	0	0	2
M10.017	27	P05067	352	685	0	0	11	1	5	0	0	2
M10.017	27	P05067	352	687	11	1	0	0	0	3	1	1

4. Conclusions

Cleavages in substrates (proteins, peptides and synthetic substrates) by proteolytic enzymes have been collected from the literature. In total, 66,615 cleavages have been annotated for 1700 different peptidases (69% of the 2457 different peptidases so far identified). The number of cleavages per peptidase varies greatly, from zero to 13,770 for trypsin 1. Peptidases with the most known cleavages are derived from proteomics experiments.

Users should be aware of the biases that may exist in the substrate cleavage data for a peptidase. If the data are predominantly derived from amino acid scanning experiments in which all but one amino acid is changed in a variety of synthetic peptides, then the apparent specificity will be affected by the amino acids that are kept constant. Similarly, if the cleavages are in synthetic substrates, then the amino acid variety may be limited with changes only to blocking and reporter groups. To address this problem, a "reliability score" has been introduced, which is the average percentage difference in the residues P4–P4′ for all substrates for a peptidase. A high reliability score will indicate a higher variety in residues occupying substrate binding sites. However, this score is affected when several of the substrate binding pockets that confer specificity.

By analysing the residues that occupy the P4–P4′ substrate binding sites, it has been possible to categorize the specificity of some peptidases. A logo and specificity matrix is shown on the MEROPS website for each of the 556 peptidases (22.7%) with ten or more substrate cleavages. Of these, 107 endopeptidases and 43 exopeptidases show a preference in one substrate binding pocket, 79 peptidases show a preference in two binding pockets, 50 show a preference at three binding pockets, 29 show a preference at four binding pockets and 53 show no positive preference in any substrate binding site. However, it has been proposed that at least 40 cleavages in disparate proteins are required for specificity analysis to be meaningful, and only 163 peptidases (6.6%) fulfil this criterion.

Proteomics experiments in which cleavage sites are identified by mass spectroscopy provides the bulk of the substrate cleavage data. However, a common problem is that in order to extract the proteins from the sample, the natural, physical boundaries that separate peptidase from substrate may have been removed, and a large number of substrate cleavages are generated. While this is not a problem if no claims are made about the physiological relevance of the data and the suite of peptides are generated just to examine peptidase specificity, there are issues when it is claimed that the experiment will reveal physiological substrates. The biggest concern is that not all substrates will be physiological, but distinguishing these from artefactual and bystander substrates is difficult. A bioinformatics approach that has been adopted here is to consider whether the cleavage is conserved in orthologues of the substrate protein. For each cleavage site, the UniProt accession of

the substrate protein is determined, as well as the UniRef50 cluster of sequences to which it belongs. All protein sequences from the UniRef50 cluster are then aligned, and the residues P4–P4′ for each cleavage site are examined for conservation. The number of replacements that are not found to occupy the same binding pocket in any of the substrates for the peptidase is counted. If this number is high, then the likelihood that the substrate is physiologically relevant is low. A server has been set up at the MEROPS website whereby the results from a proteomics experiment to study peptidase specificity can be uploaded, and the conservation around each cleavage site examined. Results are returned by E-mail.

Acknowledgements

The author would like to thank Dr Alan Barrett for maintaining the literature collection for the MEROPS database; Dr Usha Mahadevan (Molecular Connections) and summer students Matthew Jenner, Jack Feltham and Danielle Weaver for all their hard work in collecting substrate cleavages; Jung Kong for implementing the WebLogo software; Wei Yang for help with collecting data from the PRIDE database; and Matthew Waller and Paul Bevan from the Wellcome Trust Sanger Institute web team and former colleagues Emmett O'Brien, Dominic Tolle, Fraser Morton and Joyce Kok for help with the design, development and maintenance of the MEROPS website. I would also like to thank John Marioni, statistician at EMBL-EBI, and colleagues from the Protein Families team at EMBL-EBI for helpful discussions and Penny Coggill for proof-reading of the paper. This project was funded by the Wellcome Trust (grant number WT077044/Z/05/Z).

References

[1] A.J. Barrett, N.D. Rawlings, 'Species' of peptidases, Biol. Chem. 388 (2007) 1151–1157, http://dx.doi.org/10.1515/BC.2007.151.
[2] I. Schechter, A. Berger, On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain, Biochem. Biophys. Res. Commun. 32 (1968) 898–902, http://dx.doi.org/10.1016/0006-291X(68)90326-4.
[3] J.F. Woessner, A.J. Barrett, N.D. Rawlings, Handbook of Proteolytic Enzymes, first ed., Academic, London, 1998.
[4] J. Song, H. Tan, S.E. Boyd, H. Shen, K. Mahmood, G.I. Webb, T. Akutsu, J.C. Whisstock, R.N. Pike, Bioinformatic approaches for predicting substrates of proteases, J. Bioinformatics Comput. Biol. 9 (2011) 149–178. J Bioinform Comput Biol. Feb;9(1):149–78.
[5] J. Song, H. Tan, A.J. Perry, T. Akutsu, G.I. Webb, J.C. Whisstock, R.N. Pike, PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites, PLoS One 7 (11) (2012) e50300, http://dx.doi.org/10.1371/journal.pone.0050300.
[6] J.E. Fuchs, S. von Grafenstein, R.G. Huber, M.A. Margreiter, G.M. Spitzer, H.G. Wallnoefer, K.R. Liedl, Cleavage entropy as quantitative measure of protease specificity, PLoS Comput. Biol. 9 (2013) e1003007, http://dx.doi.org/10.1371/journal.pcbi.1003007.
[7] N. Fortelny, J.H. Cox, R. Kappelhoff, A.E. Starr, P.F. Lange, P. Pavlidis, C.M. Overall, Network analyses reveal pervasive functional regulation between proteases in the human protease web, PLoS Biol. 12 (2014) e1001869, http://dx.doi.org/10.1371/journal.pbio.1001869.
[8] N.D. Rawlings, A.J. Barrett, A. Bateman, Using the MEROPS database for proteolytic enzymes and their inhibitors and substrates, Curr. Protoc. Bioinformatics 48 (2014), http://dx.doi.org/10.1002/0471250953.bi0125s48, 1.25.1-1.25.33.
[9] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402, http://dx.doi.org/10.1093/nar/25.17.3389.
[10] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37, http://dx.doi.org/10.1093/nar/gkr367.
[11] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948, http://dx.doi.org/10.1093/bioinformatics/btm404.
[12] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780, http://dx.doi.org/10.1093/molbev/mst010.
[13] K. Howe, A. Bateman, R. Durbin, QuickTree: building huge neighbour-joining trees of protein sequences, Bioinformatics 18 (2002) 1546–1547, http://dx.doi.org/10.1093/bioinformatics/18.11.1546.
[14] Y. Igarashi, A. Eroshkin, S. Gramatikova, K. Gramatikoff, Y. Zhang, J.W. Smith, A.L. Osterman, A. Godzik, CutDB: a proteolytic event database, Nucleic Acids Res. 35 (Database issue) (2007) D546–D549, http://dx.doi.org/10.1093/nar/gkl813.
[15] A.U. Lüthi, S.J. Martin, The CASBAH: a searchable database of caspase substrates, Cell Death Differ. 14 (2007) 641–650, http://dx.doi.org/10.1038/sj.cdd.4402103.
[16] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, A. Bairoch, Protein identification and analysis tools on the ExPASy server, in: J.M. Walker (Ed.), The Proteomics Protocols Handbook, Humana Press, 2005, pp. 571–607.
[17] P. Gaudet, P.A. Michel, M. Zahn-Zabal, I. Cusin, P.D. Duek, O. Evalet, A. Gateau, A. Gleizes, M. Pereira, D. Teixeira, Y. Zhang, L. Lane, A. Bairoch, The neXtProt knowledgebase on human proteins: current status, Nucleic Acids Res. 43 (Database issue) (2015) D764–D770, http://dx.doi.org/10.1093/nar/gku1178.
[18] J.A. Vizcaíno, R.G. Côté, A. Csordas, J.A. Dianes, A. Fabregat, J.M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger, D. Ovelleiro, Y. Pérez-Riverol, F. Reisinger, D. Ríos, R. Wang, H. Hermjakob, The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013, Nucleic Acids Res. 41 (Database issue) (2013) D1063–D1069, http://dx.doi.org/10.1093/nar/gks1262.
[19] Z.W. Lai, A. Petrera, O. Schilling, Protein amino-terminal modifications and proteomic approaches for N-terminal profiling, Curr. Opin. Chem. Biol. 24 (2015) 71–79, http://dx.doi.org/10.1016/j.cbpa.2014.10.026.
[20] P. Schlage, U. auf dem Keller, Proteomic approaches to uncover MMP function, Matrix Biol. 44–46 (2015) 232–238, http://dx.doi.org/10.1016/j.matbio.2015.01.003.
[21] P.F. Lange, C.M. Overall, Protein TAILS: when termini tell tales of proteolysis and function, Curr. Opin. Chem. Biol. 17 (2013) 73–82, http://dx.doi.org/10.1016/j.cbpa.2012.11.025.
[22] K. Gevaert, M. Goethals, L. Martens, J. Van Damme, A. Staes, G.R. Thomas, J. Vandekerckhove, Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides, Nat. Biotechnol. 21 (2003) 566–569, http://dx.doi.org/10.1038/nbt810.
[23] F.N. Vögtle, C. Prinz, J. Kellermann, F. Lottspeich, N. Pfanner, C. Meisinger, Mitochondrial protein turnover: role of the precursor intermediate peptidase Oct1 in protein stabilization, Mol. Biol. Cell 22 (2011) 2135–2143, http://dx.doi.org/10.1091/mbc.E11-02-0169.
[24] J.C. Timmer, G.S. Salvesen, Caspase substrates, Cell Death Differ. 14 (2007) 66–72, http://dx.doi.org/10.1038/sj.cdd.4402059.
[25] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (1999) 3551–3567, http://dx.doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.
[26] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 (2002) 5383–5392.
[27] O. Schilling, C.M. Overall, Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites, Nat. Biotechnol. 26 (2008) 685–694, http://dx.doi.org/10.1038/nbt1408.
[28] 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, http://dx.doi.org/10.1074/jbc.M111.260026.
[29] F. Impens, N. Colaert, K. Helsens, B. Ghesquière, E. Timmerman, P.J. De Bock, B.M. Chain, J. Vandekerckhove, K. Gevaert, A quantitative proteomics design for systematic identification of protease cleavage events, Mol. Cell Proteomics 9 (2010) 2327–2333, http://dx.doi.org/10.1074/mcp.M110.001271.
[30] P. Van Damme, A. Staes, S. Bronsoms, K. Helsens, N. Colaert, E. Timmerman, F.X. Aviles, J. Vandekerckhove, K. Gevaert, Complementary positional proteomics for screening substrates of endo- and exoproteases, Nat. Methods 7 (2010) 512–515, http://dx.doi.org/10.1038/nmeth.1469.
[31] P. Van Damme, S. Maurer-Stroh, K. Plasman, J. Van Durme, N. Colaert, E. Timmerman, P.J. De Bock, M. Goethals, F. Rousseau, J. Schymkowitz, J. Vandekerckhove, K. Gevaert, Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs, Mol. Cell Proteomics 8 (2009) 258–272, http://dx.doi.org/10.1074/mcp.M800060-MCP200.
[32] M.L. Biniossek, D.K. Nägler, C. Becker-Pauly, O. Schilling, Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S, J. Proteome Res. 10 (2011) 5363–5373, http://dx.doi.org/10.1021/pr200621z.
[33] C. Becker-Pauly, O. Barré, O. Schilling, U. Auf dem Keller, A. Ohler, C. Broder, A. Schütte, R. Kappelhoff, W. Stöcker, C.M. Overall, Proteomic analyses reveal an acidic prime side specificity for the astacin metalloprotease family reflected by physiological substrates, Mol. Cell Proteomics 10 (2011), http://dx.doi.org/10.1074/mcp.M111.009233. M111.009233.
[34] R. Cruz, P. Huesgen, S.P. Riley, A. Wlodawer, C. Faro, C.M. Overall, J.J. Martinez, I. Simões, RC1339/APRc from *Rickettsia conorii* is a novel aspartic protease with properties of retropepsin-like enzymes, PLoS Pathog. 10 (2014) e1004324, http://dx.doi.org/10.1371/journal.ppat.1004324.
[35] K.R. Lynn, W.J. Brockbank, N.A. Clevette, Multiple forms of the asclepains. Cysteinyl proteases from milkweed, Biochim. Biophys. Acta 612 (1980) 119–125.

[36] R.Q. Chen, Y. Jin, J.B. Wu, X.D. Zhou, D.S. Li, Q.M. Lu, W.Y. Wang, Y.L. Xiong, A novel high molecular weight metalloproteinase cleaves fragment F1 of activated human prothrombin, Toxicon 44 (2004) 281–287.
[37] A.A. Kortt, J.B. Caldwell, G.G. Lilley, R. Edwards, J. Vaughan, D.J. Stewart, Characterization of a basic serine proteinase (pI approximately 9.5) secreted by virulent strains of *Dichelobacter nodosus* and identification of a distinct, but closely related, proteinase secreted by benign strains, Biochem. J. 299 (1994) 521–525.
[38] T.A. Naumann, D.T. Wicklow, N.P. Price, Polyglycine hydrolases secreted by Pleosporineae fungi that target the linker region of plant class IV chitinases, Biochem. J. 460 (2014) 187–198, http://dx.doi.org/10.1042/BJ20140268.
[39] D. Chandu, A. Kumar, D. Nandi, PepN, the major Suc-LLVY-AMC-hydrolyzing enzyme in *Escherichia coli*, displays functional similarity with downstream processing enzymes in Archaea and Eukarya. Implications in cytosolic protein degradation, J. Biol. Chem. 278 (2003) 5548–5556, http://dx.doi.org/10.1074/jbc.M207926200.
[40] L. Fanuel, C. Goffin, A. Cheggour, B. Devreese, G. Van Driessche, B. Joris, J. Van Beeumen, J.M. Frère, The DmpA aminopeptidase from *Ochrobactrum anthropi* LMG7991 is the prototype of a new terminal nucleophile hydrolase family, Biochem. J. 341 (1999) 147–155, http://dx.doi.org/10.1042/0264-6021:3410147.
[41] S.J. Vollmer, R.L. Switzer, M.A. Hermodson, S.G. Bower, H. Zalkin, The glutamine-utilizing site of *Bacillus subtilis* glutamine phosphoribosylpyrophosphate amidotransferase, J. Biol. Chem. 258 (1983) 10582–10585.
[42] C.G. Suresh, A.V. Pundle, H. SivaRaman, K.N. Rao, J.A. Brannigan, C.E. McVey, C.S. Verma, Z. Dauter, E.J. Dodson, G.G. Dodson, Penicillin V acylase crystal structure reveals new Ntn-hydrolase family members, Nat. Struct. Biol. 6 (1999) 414–416.
[43] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190, http://dx.doi.org/10.1101/gr.849004.
[44] V. Oliveira, M. Campos, R.L. Melo, E.S. Ferro, A.C. Camargo, M.A. Juliano, L. Juliano, Substrate specificity characterization of recombinant metallo oligopeptidases thimet oligopeptidase and neurolysin, Biochemistry 40 (2001) 4417–4425, http://dx.doi.org/10.1021/bi002715k.
[45] O. Dolnik, V. Volchkova, W. Garten, C. Carbonnelle, S. Becker, J. Kahnt, U. Ströher, H.D. Klenk, V. Volchkov, Ectodomain shedding of the glycoprotein GP of Ebola virus, EMBO J. 23 (2004) 2175–2184, http://dx.doi.org/10.1038/sj.emboj.7600219.
[46] B.E. Suzek, Y. Wang, H. Huang, P.B. McGarvey, C.H. Wu, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, Bioinformatics 31 (2015) 926–932, http://dx.doi.org/10.1093/bioinformatics/btu739.
[47] R.C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5 (2004) 113, http://dx.doi.org/10.1186/1471-2105-5-113.