The *MEROPS* database of peptidases and proteolysis

Neil D. Rawlings & Alan J. Barrett
The Wellcome Trust Sanger Institute
Genome Campus, Hinxton Hall, Cambridge, CB10 1SA, UK

http://merops.sanger.ac.uk

wellcome trust

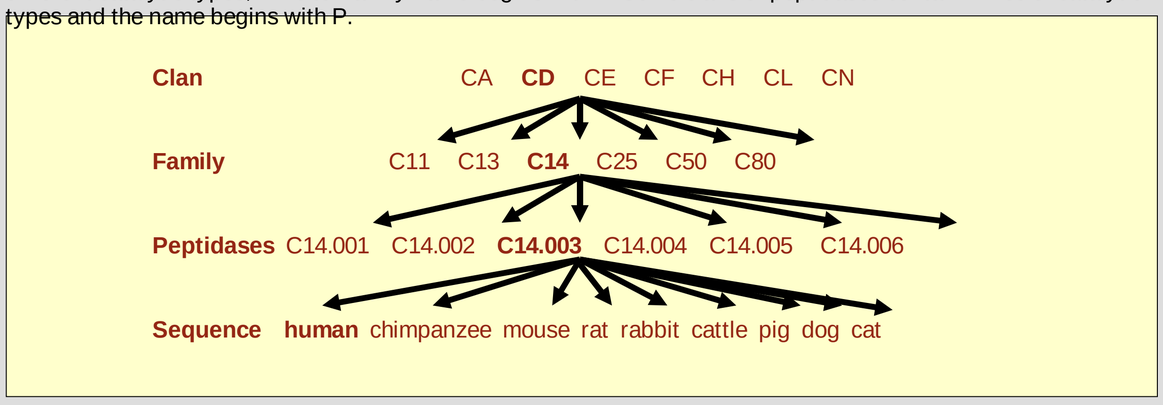
Peptidases

A peptidase is an enzyme that catalyses the hydrolysis of a peptide bond that forms between amino acids in proteins and peptides. Peptidases are important for numerous post-translational modifications such as removal of the initiating methionine of cytoplasmic proteins, release of signal peptides and other transit peptides that target proteins for secretion or to organelles, activation of proteins by removal of propeptides (*e.g.* enzymes) and processing of polyproteins (*e.g.* peptide hormone precursors). Peptidases are also involved in protein catabolism, either to acquire or recycle components during feeding, protein turnover and degradation of misfolded proteins, or to switch off biological signals, as in apoptosis and receptor shedding. Peptidases are important components of pathogens (*e.g.* parasite invasion and feeding, viral polyprotein processing, processing of bacterial toxins) and are involved in pathology (*e.g.* cancer, rheumatoid arthritis, Alzheimer's disease).

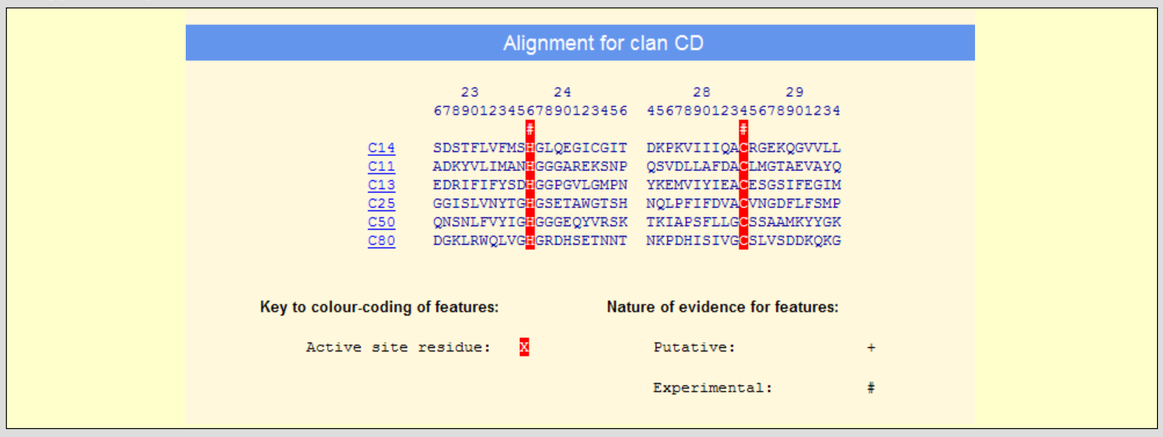
The control of proteolysis is important because unchecked degradation can lead to pathological conditions, and the *MEROPS* database (Rawlings *et al*., 2008) includes not only peptidases but also their inhibitors (Rawlings *et al*., 2004). The **latest release** of the database is 8.4 (April 2009).

MEROPS Classification

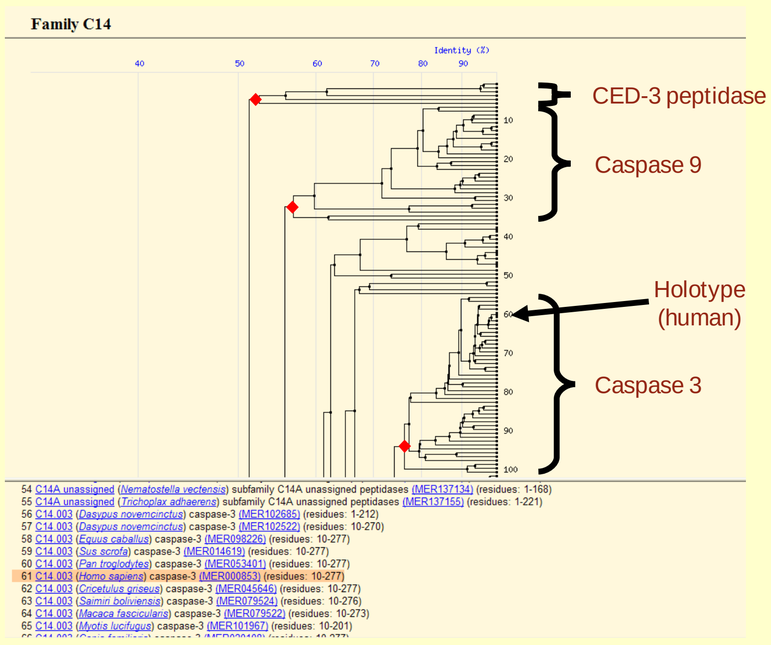
The *MEROPS* classification was developed by Rawlings & Barrett (1993) and is internationally recognized. The classification is based on protein domains and is hierarchical. For a peptidase the domain considered (the **PEPTIDASE UNIT**) is the one that carries the active site residues. For a protein inhibitor it is the domain carrying the reactive bond (or its equivalent) and it is common for one sequence to contain several inhibitor domains. The figure below shows the classification of the peptidase **CASPASE-3**, one of the executor caspases at the top of the apoptosis cascade. Homologues that are known or predicted to share similar tertiary structures are grouped into a **CLAN** (caspases are members of clan **CD**). Homologues that share significant sequence similarity are grouped into a **FAMILY** (the caspase family is **C14**). A member of the family that has been biochemically characterized is chosen as a **TYPE EXAMPLE**. Each different peptidase within the family is assigned to a ***MEROPS* IDENTIFIER** (the identifier for caspase-3 is **C14.001**) and one sequence is chosen as a **HOLOTYPE** (human caspase 3). An identifier is only established if there is published biochemical characterization, and many homologues remain unassigned. There are six different catalytic **TYPES** of peptidases (plus some where the catalytic type is unknown) and the initial letter of the clan and family names and the *MEROPS* identifier reflects this: A for aspartic, C for cysteine, G for glutamic, M for metallopeptidase, S for serine, T for threonine and U for unknown. Because inhibitors in the same family can inhibit peptidases of different catalytic types, inhibitor family name begins with I. Some clans of peptidases contain mixes of catalytic types and the name begins with P.



Families are included in the same clan when the fold is known to be similar but for several families the fold has not been solved for any member. A family can be included in a clan if the active site residues are in the same order in the sequence or are within conserved motifs. Below is shown the sequences around the active site dyad for type examples from families in clan CD.



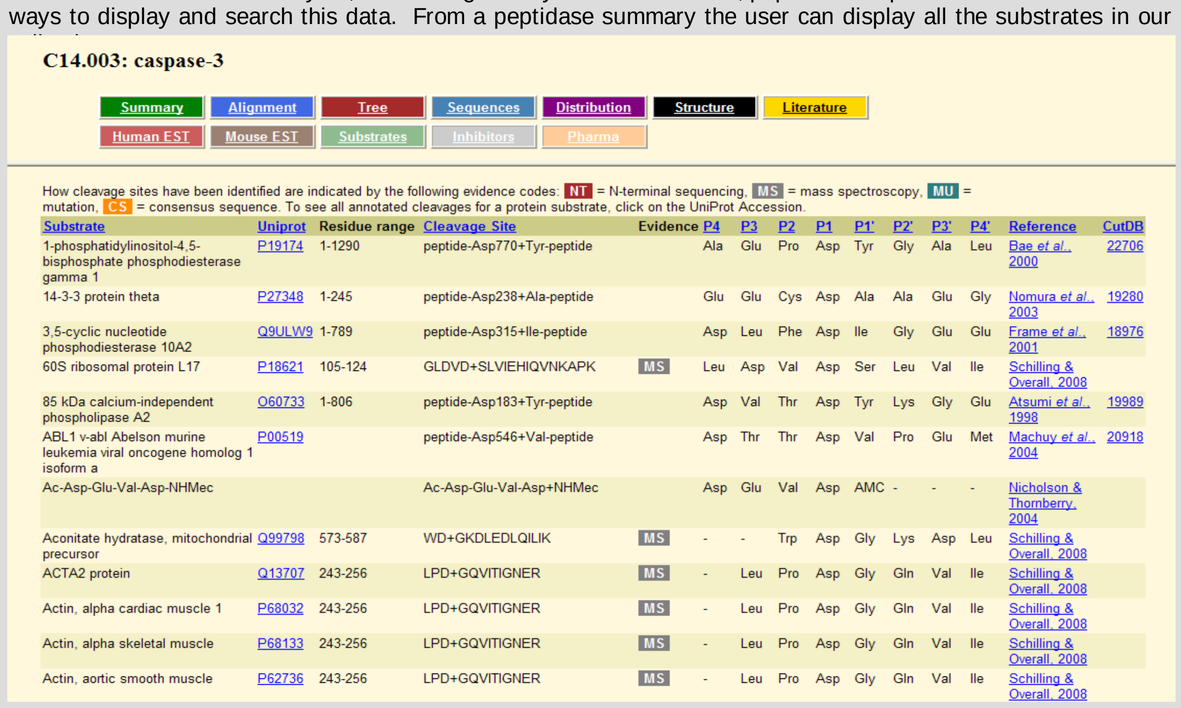
Assigning a biochemically uncharacterized sequence to a *MEROPS* identifier requires an alignment (built by MAFFT) and a phylogenetic tree derived from it (built by the UPGMA algorithm implemented in QuickTree). To be included in the same identifier, sequences must be derived from the same node on the tree as the holotype, and to have a similar protein architecture. There should no indication from the sequence of known substrate binding residues that specificity is going to be radically different from that of the holotype.



References:
Rawlings, N.D., Morton, F.R., Kok, C.Y., Kong, J. & Barrett, A.J. (2008) *MEROPS*: the peptidase database. *Nucleic Acids Res* 36, D320-D325.
Rawlings, N.D., Tolle, D.P. & Barrett, A.J. (2004) Evolutionary families of peptidase inhibitors. *Biochem J*. 378, 705-716.
Rawlings, N.D. & Barrett, A.J. (1993) Evolutionary families of peptidases. *Biochem J.* 290, 205-218.

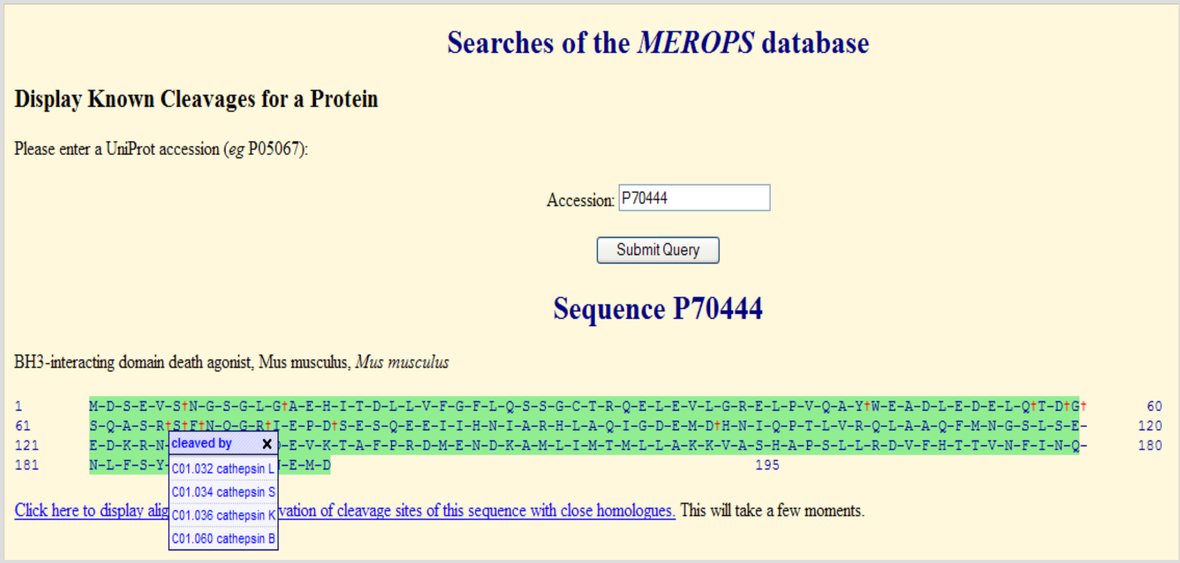
Peptidase and Substrate Interactions

The *MEROPS* database also includes data on peptidase/substrate interactions, which are short-lived but fundamental. The data have to be collected from the literature, which is a time-consuming process. Nevertheless, *MEROPS* now includes nearly 37,000 cleavages in synthetic substrates, peptides and proteins. There are various ways to display and search this data. From a peptidase summary the user can display all the substrates in our

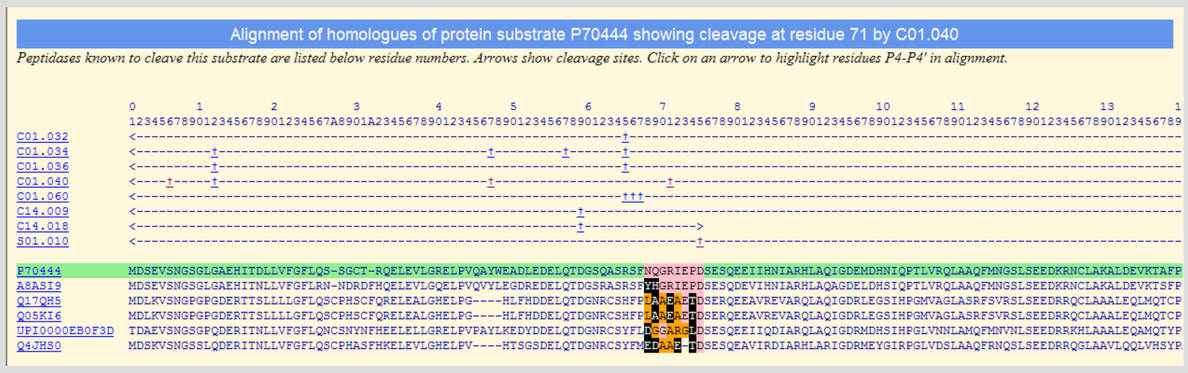


Substrate	Uniprot	Residue range	Cleavage Site	Evidence	P4	P3	P2	P1	P1'	P2'	P3'	P4'	Reference	CutDB
1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase gamma 1	P19174	1-1290	peptide-Asp770+Tyr-peptide		Ala	Glu	Pro	Asp	Tyr	Gly	Ala	Leu	Bae *et al.*, 2000	22706
14-3-3 protein theta	P27348	1-245	peptide-Asp238+Ala-peptide		Glu	Glu	Cys	Asp	Ala	Ala	Glu	Gly	Nomura *et al.*, 2003	19280
3,5-cyclic nucleotide phosphodiesterase 10A2	Q9ULW9	1-789	peptide-Asp315+Ile-peptide		Asp	Leu	Phe	Asp	Ile	Gly	Glu	Glu	Frame *et al.*, 2001	18976
60S ribosomal protein L17	P18621	105-124	GLDVD+SLVIEHIQVNKAPK	MS	Leu	Asp	Val	Asp	Ser	Leu	Val	Ile	Schilling & Overall, 2008	
85 kDa calcium-independent phospholipase A2	O60733	1-806	peptide-Asp183+Tyr-peptide		Asp	Val	Thr	Asp	Tyr	Lys	Gly	Glu	Atsumi *et al.*, 1998	19989
ABL1 v-abl Abelson murine leukemia viral oncogene homolog 1 isoform a	P00519		peptide-Asp546+Val-peptide		Asp	Thr	Thr	Asp	Val	Pro	Glu	Met	Machuy *et al.*, 2004	20918
Ac-Asp-Glu-Val-Asp-NHMec			Ac-Asp-Glu-Val-Asp+NHMec		Asp	Glu	Val	Asp	AMC	-	-	-	Nicholson & Thornberry, 2004	
Aconitate hydratase, mitochondrial precursor	Q99798	573-587	WD+GKDLEDLQILIK	MS	-	-	Trp	Asp	Gly	Lys	Asp	Leu	Schilling & Overall, 2008	
ACTA2 protein	Q13707	243-256	LPD+GQVITIGNER	MS	-	Leu	Pro	Asp	Gly	Gln	Val	Ile	Schilling & Overall, 2008	
Actin, alpha cardiac muscle 1	P68032	243-256	LPD+GQVITIGNER	MS	-	Leu	Pro	Asp	Gly	Gln	Val	Ile	Schilling & Overall, 2008	
Actin, alpha skeletal muscle	P68133	243-256	LPD+GQVITIGNER	MS	-	Leu	Pro	Asp	Gly	Gln	Val	Ile	Schilling & Overall, 2008	
Actin, aortic smooth muscle	P62736	243-256	LPD+GQVITIGNER	MS	-	Leu	Pro	Asp	Gly	Gln	Val	Ile	Schilling & Overall, 2008	

By clicking on the UniProt identifier, the user is taken to a second display showing the protein sequence in question with the mature protein highlighted and known cleavage sites indicated by red cross symbols. The mouse-over (white box) shows the peptidases responsible for the cleavages, with the non-physiological cleavages in italics.

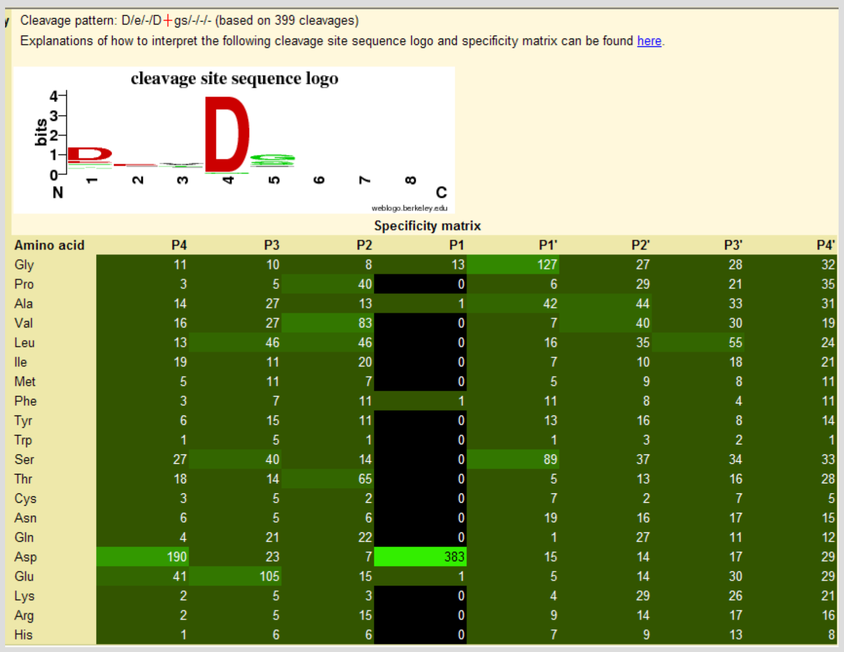


On clicking the link at the bottom of this display a dynamic alignment is generated showing all close homologues of this sequence (from UniRef50, *i.e.* sequences with 50% identity or more). The sequence where the cleavages are known is highlighted in green. Peptidases known to cleave this protein are shown along with the cleavage position and the range of the peptide used in the study. In this case, cleavage of the BID protein by cathepsin H is shown. Cleavage of BID is known to instigate apoptosis. Residues highlighted in pink are completely conserved, residues highlighted in orange are acceptible substitutions known from other substrates of cathepsin H, and residues highlighted in black are unacceptible replacements. It is clear that this cleavage position is not conserved, implying that cathepsin H cleavage here is either not physiological, or may be pathological in humans.



Specificity

The specificity of caspase-3 has been calculated from 399 known cleavages and is shown as a **LOGO**. The logo is derived from the four residues either side of each cleavage (known as P4-P4'). Cleavage occurs after residue 4 (P1). This accords very well with the known absolute requirement for cleavage of aspartyl bonds and a preference for Asp in P4. The second display is a **HEAT MAP** derived from the same data. The brighter the cell the greater the preference for an amino acid at this position. A black cell indicates that the amino acid is unknown at a position.



Amino acid	P4	P3	P2	P1	P1'	P2'	P3'	P4'
Gly	11	10	8	13	127	27	28	32
Pro	3	5	40	0	6	29	21	35
Ala	14	27	13	1	42	44	33	31
Val	16	27	83	0	7	40	30	19
Leu	13	46	46	0	16	35	55	24
Ile	19	11	20	0	7	10	18	21
Met	5	11	7	0	5	9	8	11
Phe	3	7	11	1	11	8	4	11
Tyr	6	15	11	0	13	16	8	14
Trp	1	5	1	0	1	3	2	1
Ser	27	40	14	0	89	37	34	33
Thr	18	14	65	0	5	13	16	28
Cys	3	5	2	0	7	2	7	5
Asn	6	5	6	0	19	16	17	15
Gln	4	21	22	0	1	27	11	12
Asp	190	23	7	383	15	14	17	29
Glu	41	105	15	1	5	14	30	29
Lys	2	5	3	0	4	29	26	21
Arg	2	5	15	0	9	14	17	16
His	1	6	6	0	7	9	13	8

Nature Precedings : doi:10.1038/npre.2009.3103.1 : Posted 20 Apr 2009