D270–D272 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj089

MEROPS: the peptidase database

Neil D. Rawlings*, Fraser R. Morton and Alan J. Barrett

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

Received September 15, 2005; Accepted October 13, 2005

ABSTRACT

Peptidases (proteolytic enzymes) and their natural, protein inhibitors are of great relevance to biology, medicine and biotechnology. The *MEROPS* database (http://merops.sanger.ac.uk) aims to fulfil the need for an integrated source of information about these proteins. The organizational principle of the database is a hierarchical classification in which homologous sets of proteins of interest are grouped into families and the homologous families are grouped in clans. The most important addition to the database has been newly written, concise text annotations for each peptidase family. Other forms of information recently added include highlighting of active site residues (or the replacements that render some homologues inactive) in the sequence displays and BlastP search results, dynamically generated alignments and trees at the peptidase or inhibitor level, and a curated list of human and mouse homologues that have been experimentally characterized as active. A new way to display information at taxonomic levels higher than species has been devised. In the Literature pages, references have been flagged to draw attention to particularly 'hot' topics.

INTRODUCTION

The *MEROPS* database is a manually curated information resource for proteolytic enzymes (best termed peptidases) and their protein inhibitors. The database has been in existence since 1996 and may be found at http://merops.sanger.ac.uk.

The importance of peptidases and their inhibitors is illustrated by the fact that 18% of sequences in the SwissProt database are annotated as undergoing proteolytic processing, ∼2% of all genes encode peptidases and their homologues in all kinds of organisms, and that there are over 550 active and putative peptidases in the human genome. The *MEROPS* classification of peptidases, founded in 1993 (1), is a classification at the protein domain level (we term the domain in question the peptidase unit) and is hierarchical. Homologues that are biochemically similar are given the same identifier, and for each identifier a holotype is nominated. Proteins with homologous peptidase units are grouped in a family. Every member of the family has to be shown to be related to a nominated type example. Families are grouped in a clan if there are indications, principally from tertiary structure similarities, that there was a common ancestor. The same principles have been used to classify the protein inhibitors of peptidases (2). The inhibitors are not as widely distributed as peptidases, with only one or two known from each bacterial or archaean genome. In the human genome, there are 105 known inhibitors and a further 176 homologues that are not known to be inhibitory.

Statistics from release 7.1 (July 2005) of *MEROPS* are shown in Table 1 and compared with release 6.3 from June 2003. Although the number of sequences classified as peptidase homologues has nearly doubled, these have mostly been additions to existing families, and only eight new peptidase families have been discovered since June 2003.

FAMILY SUMMARIES

We have added text summaries for all peptidases families. Each summary is structured under the following headings, with a brief description of the contents:

(i) Content of family: a description of the catalytic type and whether the peptidases in the family are endopeptidases or exopeptidases (aminopeptidases, carboxypeptidases, etc.) or a mixture.
(ii) History: when peptidases in the family were first discovered and other essential background information.

Table 1. Counts of identifiers, families and clans for peptidase and protein inhibitor homologues in the *MEROPS* database

	MEROPS 7.1 Peptidases	Inhibitors	MEROPS 6.3 Peptidases	Inhibitors
Sequences	30 909	3 690	18 076	2651
Identifiers	2 053	532	1 711	318
Families	180	53	172	48
Clans	39	32	33	25

Counts are shown for release 7.1 of the *MEROPS* database (July 2005) and for release 6.3 (June 2003).

*To whom correspondence should be addressed. Tel: +44 1223 495330; Fax: +44 1223 494919; Email: ndr@sanger.ac.uk

(iii) Active site: the residues that are important for catalysis, including metal ligands for metallopeptidases, and describing the variability of amino acids at each position.
(iv) Activities and specificities: reaction conditions and example substrates.
(v) Inhibitors: predominantly lists small molecule inhibitors that can be used to distinguish members of the family.
(vi) Molecular structure: if the tertiary structure has been determined for any member of the family, then the fold is described and compared with others. Also included here are domain organization and conservation of features, such as disulfide bridges and transmembrane regions.
(vii) Distribution of family: this is included when the distribution among organisms is unusual, e.g. C51 which is found only in bacteriophages that infect staphylococci.
(viii) Biological functions: examples of known physiological and pathological roles.
(ix) Pharmaceutical and biotech relevance: peptidases in the family that are drug targets or have industrial uses.

Summaries have also been written for families of protein inhibitors. The headings are content of family, history, reactive site, peptidase inhibited, molecular structure and distribution of family.

FLAGGING OF TOPICS IN LITERATURE PAGES

The literature on peptidases is large, and the Literature pages in *MEROPS* contain well over 20 000 references. So that it may be easier to spot a paper on a particular topic in a Literature page, we have added 'flags' for six important topics. Thus, 'E' indicates that the paper contains information on the recombinant Expression of a peptidase; 'I' shows that we found the article to be relevant to the design of Inhibitors for the enzyme; 'K' means that the paper deals with a gene Knockout or other artificial genetic manipulation; 'M' shows that the paper deals with a natural Mutation, allelic variant or polymorphism; 'R' indicates that the article includes information about an RNA splicing variant; 'S' means that the article deals with 3D Structure; and 'V' shows that the article is a review.

HIGHLIGHTING IN SEQUENCE DISPLAYS, FAMILY PAGES AND BLASTP RESULTS

Active site residues tend to be highly conserved because of the restraints imposed by the catalytic function. For this reason, even unrelated peptidases may have similar or even superimposable active sites. For example, the serine, histidine, aspartate catalytic triad is found in trypsin (family S1, clan PA), subtilisin (family S8, clan SB) and carboxypeptidase Y (family S9, clan SC), although the order of the residues is different and the tertiary folds are completely different. Many peptidase families contain homologues that are not peptidases because one or more active site residues has been changed (though we stress that in some families there are enzymes other than peptidases that possess a matching set of active site residues and/or metal ligands, such as succinyl-diaminopimelate desuccinylase in family M20). In our annotated family alignments, we adopted the convention of showing active site residues as white text on a red background, metal ligands as white text on a blue background and replacements of either as white text on a black background. We have now extended the highlighting to our sequence displays, family summaries and BlastP search results.

In our display of each sequence in FastA format (3), not only do we highlight active site residues and metal ligands, but we also highlight the peptidase (or inhibitor) unit in red. We also show when a peptidase unit is interrupted by an unrelated ('nested') domain (for example the fibronectin-like repeats in gelatinase A). The residues are now numbered and the header line includes the range of the peptidase (or inhibitor) unit and residue numbers of catalytic residues.

We have enhanced the service whereby a user may submit a sequence and search our collection of peptidase sequences using the BlastP program (4) with new annotation of the BlastP results that shows active site residues according to our standard conventions. This allows the user to tell at a glance whether the query sequence is likely to be a peptidase or not.

Highlighting of reactive site residues is not appropriate for the protein inhibitors, because the reactive site residue is highly variable within a family.

In the label file for each family alignment and tree, we now highlight the holotypes as well as the family type example.

ALIGNMENTS AND TREES AT THE PEPTIDASE/INHIBITOR LEVEL

We have always provided alignments and trees derived from them for every family in *MEROPS*, but we now generate dynamically an alignment and tree for every peptidase and protein inhibitor. The family alignments show only the peptidase or inhibitor units, but the new option also allows the user to select the complete sequence to align. We expect every peptidase or inhibitor assigned to the same identifier to have sequence similarity from N-terminus to C-terminus. An alignment at the identifier level is generated by MUSCLE (5); this alignment is used to generate a Neighbor-joining tree (6) using QuickTree (7) which is displayed using the ClustalTree Java applet written by Rodrigo Lopez and Stephen Robinson at the European Bioinformatics Institute.

PEPTIDASES AND INHIBITORS AT THE ORGANISM LEVEL

The index of organism names now includes English common names and synonyms of scientific binomials. This helps the user to find the organism of interest.

For each organism from which a peptidase is known, we list an abbreviated taxonomy (superkingdom, kingdom, phylum, subphylum, superclass, class, subclass, superorder, order, suborder, superfamily, family, subfamily and genus) and all the known peptidase homologues. As an aid to examining the distribution of peptidases among organisms, we now provide a tool to display the peptidase (or protein inhibitor) data at all higher levels in the taxonomy table in each organism page. For example, on consulting the *Plasmodium falciparum* page, a user might find it useful to know how these peptidases are distributed among all *Plasmodium* species, and clicking on the genus level in the taxonomy table will display just that. Knowing that a peptidase is present in one pathogen but not another could indicate whether the peptidase is a potential drug target.

The Organism page has been made more interactive so that the user can sort the data by clan, family, peptidase or inhibitor, or gene name by simply clicking at the top of the appropriate column. By default, the list is sorted by family.

The user can now obtain all the peptidase (or inhibitor) unit sequences for each organism in FastA format by clicking a link at the top of each Organism page.

ACTIVITY STATUS OF HUMAN AND MOUSE PEPTIDASES

In general, we create an identifier for a peptidase when it is clear that it can be distinguished biochemically or functionally from other members in the family. However, we have created an identifier for each human and mouse peptidase and inhibitor homologue. Although the creation of an identifier for a *Drosophila melanogaster* peptidase immediately implies that the peptidase is active, this is not the case for a human or mouse peptidase homologue. Consequently, we have reviewed the literature for evidence of proteolytic activity for every human and mouse peptidase homologue and included references on each peptidase page. We state whether we currently regard the homologue as an 'active', 'putative' or 'inactive' peptidase homologue. If we are aware that a peptidase has been shown experimentally to be active, we give a reference, and if on the other hand we believe it to be inactive because one or more expected active site residues are replaced, this is indicated according to the convention described above. See Table 2 for examples.

In the human genome, there are 440 peptidase homologues that we believe are inactive against peptide substrates.

Table 2. Examples of annotation of activity status of human and mouse peptidase homologues

MEROPS identifier	Recommended name	Activity status (reference)
A01.007	Renin	Human: active (Suzuki *et al*., 2004) Mouse: active (Hansen *et al*., 2004)
S09.018	Dipeptidyl-peptidase 8	Human: active (Chen *et al*., 2004) Mouse: active (by similarity to human)
S09.973	Dipeptidylpeptidase homologue DPP6	Human: inactive; S D H Mouse: inactive; S D H

For non-peptidase homologues, conserved active site residues are shown as white text on a red background, and residues that have been replaced are shown as white text on a black background.

SUMMARY

The collection and analysis of peptidase and protein inhibitor sequences has now been largely automated, and so the thrust of the *MEROPS* database for the past 2 years has been towards increased annotation, and we intend to extend this in the future with summaries for inhibitor families, and clans of peptidases and inhibitors.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Medical Research Council (UK) and the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Rawlings,N.D. and Barrett,A.J. (1993) Evolutionary families of peptidases. *Biochem. J*., **290**, 205–218.
2. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) Evolutionary families of peptidase inhibitors. *Biochem. J*., **378**, 705–716.
3. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
4. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
5. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
6. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*., **4**, 406–425.
7. Howe,K., Bateman,A. and Durbin,R. (2002) QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.