D160–D164 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh071

MEROPS: the peptidase database

Neil D. Rawlings*, Dominic P. Tolle and Alan J. Barrett

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

Received September 12, 2003; Revised and Accepted September 30, 2003

ABSTRACT

Peptidases (proteolytic enzymes) are of great relevance to biology, medicine and biotechnology. This practical importance creates a need for an integrated source of information about them, and also about their natural inhibitors. The MEROPS database (http://merops.sanger.ac.uk) aims to fill this need. The organizational principle of the database is a hierarchical classification in which homologous sets of the proteins of interest are grouped in families and the homologous families are grouped in clans. Each peptidase, family and clan has a unique identifier. The database has recently been expanded to include the protein inhibitors of peptidases, and these are classified in much the same way as the peptidases. Forms of information recently added include new links to other databases, summary alignments for peptidase clans, displays to show the distribution of peptidases and inhibitors among organisms, substrate cleavage sites and indexes for expressed sequence tag libraries containing peptidases. A new way of making hyperlinks to the database has been devised and a BlastP search of our library of peptidase and inhibitor sequences has been added.

INTRODUCTION

Proteolytic enzymes, best termed peptidases, are important in many ways to human health and technology. Their more general biological importance is illustrated by the fact that ~2% of all genes encode peptidases or their homologues in all kinds of organism (1); there are ~500 human genes that encode peptidases or their homologues. The MEROPS database is a manually curated information resource devoted to this group of enzymes, together with the protein inhibitors that regulate their activities *in vivo*. The database was started in 1996 at the Babraham Institute, but in October 2002 moved to its present location at the Wellcome Trust Sanger Institute. It may be found at http://merops.sanger.ac.uk and is freely accessible to all.

The MEROPS classification was initially developed for peptidases (2) and is a hierarchical scheme. Homologues that are considered to be biochemically indistinguishable are given the same identifier, and homologous peptidases are grouped in a family. Families are grouped in a clan if there are indications, principally from tertiary structure comparisons, that there is a common ancestor. For each object in the classification a representative type example is nominated, and this makes the system more stable than it might otherwise be. There is an identifier for every peptidase, family and clan and every identifier starts with a capital letter that shows the catalytic type of the peptidases contained in the group. The letters used are 'A' (aspartic), 'C' (cysteine), 'M' (metallo), 'S' (serine), 'T' (threonine), 'U' (unknown type) or in certain clan names, 'P' (protein nucleophile: any of the C, S or T types). The identifier of a clan consists of two letters. The first, indicating the catalytic type, is followed by a serial letter. A family name consists of a letter indicating the catalytic type followed by a serial number of up to two digits. The identifier for an individual peptidase consists of the family name (padded with zeros if necessary to make it three characters long), a dot and a three-digit serial number. Thus, pepsin A is in clan AA and family A1 and has the code A01.001.

INCLUSION OF PEPTIDASE INHIBITORS

A major development has been the inclusion of peptidase inhibitors. The principles used in the classification of peptidases (3) have been applied to the inhibitors. A notional 'catalytic type' of I is applied to all inhibitors, so their identifiers start with this letter.

To deal with the problem that the molecules of many inhibitors are mosaic proteins containing domains that are not directly involved in activity, the term 'inhibitor unit' is used to refer to the sub-sequence that is responsible for reactivity, and it is the inhibitor unit that is used in sequence comparisons. However, the functional units are often much smaller than those of peptidases. It is scarcely possible to establish the statistical significance of similarities between very short sequences, and the MEROPS database only includes inhibitor units of 14 residues or more. This minimum size is that of the sunflower cyclic trypsin inhibitor (I12.002) and marinostatin C (I10.001). This size limit has the effect that all the inhibitors listed are encoded by genes.

The objects recognized at the inhibitor level are generally the complete, functional inhibitors, despite the fact that they are classified with reference to the amino acid sequence of the inhibitor unit. This is not possible, however, for compound inhibitors, i.e. those that contain more than one inhibitor unit in a single molecule. These compound inhibitors cannot be placed in the hierarchical classification. For these, it is the individual units that have the standard identifiers, and the compound molecules require a separate type of identifier. An example of such an identifier is 'LI01-001' for ovomucoid,

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

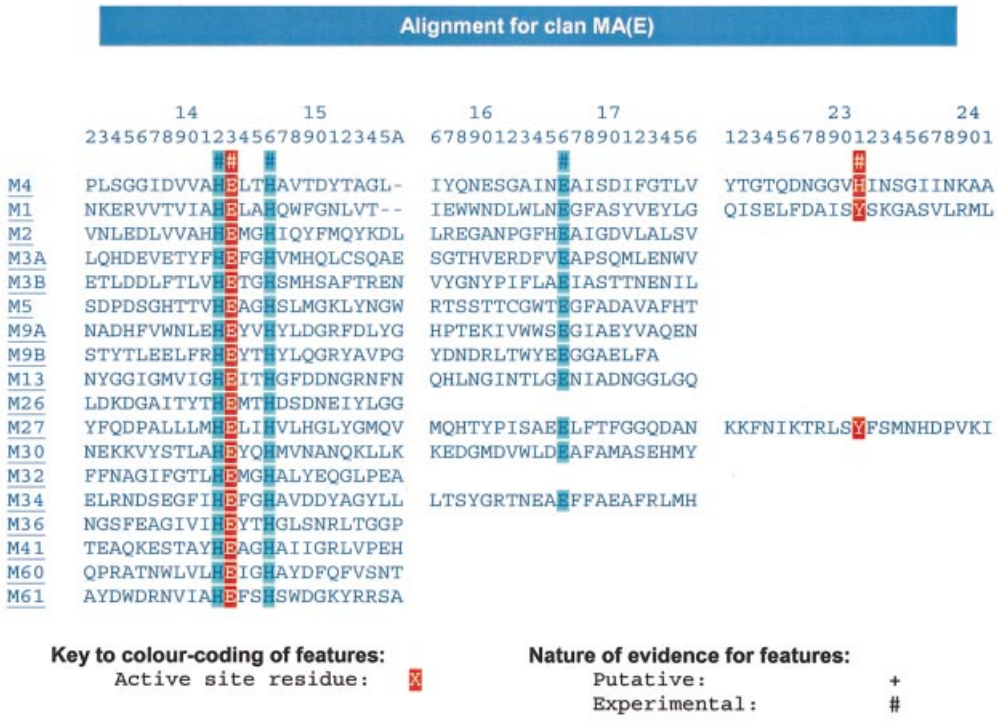

Figure 1. An alignment around the catalytic residues and zinc ligands in sub-clan MA(E). Ten residues are shown N-terminal and C-terminal to each catalytic residue or zinc ligand for the type example for each family in the subclan.

which contains three inhibitory units. The initial 'L' is for any compound inhibitor. The next three characters give the family name of the inhibitor units (padded to three characters). Following the dash, there is a three-digit serial number. A few compound inhibitors contain inhibitor units from more than one family and special identifiers are used for these.

Inhibitor families and clans are assembled in much the same way as peptidase families and clans. However, there are many more clans of inhibitors that contain only one family. Inhibitors are also not as numerous as peptidases, though 250 human genes encode inhibitors or their homologues.

USER INTERFACE

The general appearance of the MEROPS database has been described previously, as have many of its features (1–5). Those descriptions will not be repeated here, and attention will be reserved for the newer aspects. MEROPS is now effectively two databases, one for peptidases and one for inhibitors, and the user can switch between one and the other by selecting the first option in the sidebar.

LINKS TO OTHER RESOURCES

New connections to other databases have been added. From the family summary pages there are now links to the HOMSTRAD database of structural alignments (6), and the Pfam (7) and InterPro (8) databases of protein domains. From the peptidase and inhibitor summary pages we have added links to the Ensembl database (9), connecting the user to the Gene Report page for the human and mouse genes. There are also links to the Pfam database from the sequences pages that connect the user to the graphical display of the Swiss-Prot entry showing the domain architecture for each protein.

CLAN ALIGNMENTS

Being able to compare sequences around active site residues has proved a useful tool for assigning peptidase families to clans, and so we now include alignments at the clan level. These graphically illustrate relationships between families within a clan, and assist users in identifying active site residues and in the making of active site motifs. It is not useful to make full alignments of sequences from different families, and to include all the sequences in a clan would produce inconveniently large alignments, so we show only the type example from each family, and only 10 residues either side of an active site residue. An example of such a clan alignment is shown in Figure 1.

SEQUENCE COLLECTIONS

We have assembled a file containing the sequences of all the human or mouse peptidase or inhibitor units for each family. These are in FastA format with a single-line header giving the MEROPS recommended name, source organism, MEROPS identifier, sequence source (i.e. from which primary database the sequence is derived) and the limits of the peptidase unit. Each file can be downloaded and used as a library for BlastP or FastA searches. The files are accessible from the family summary page.

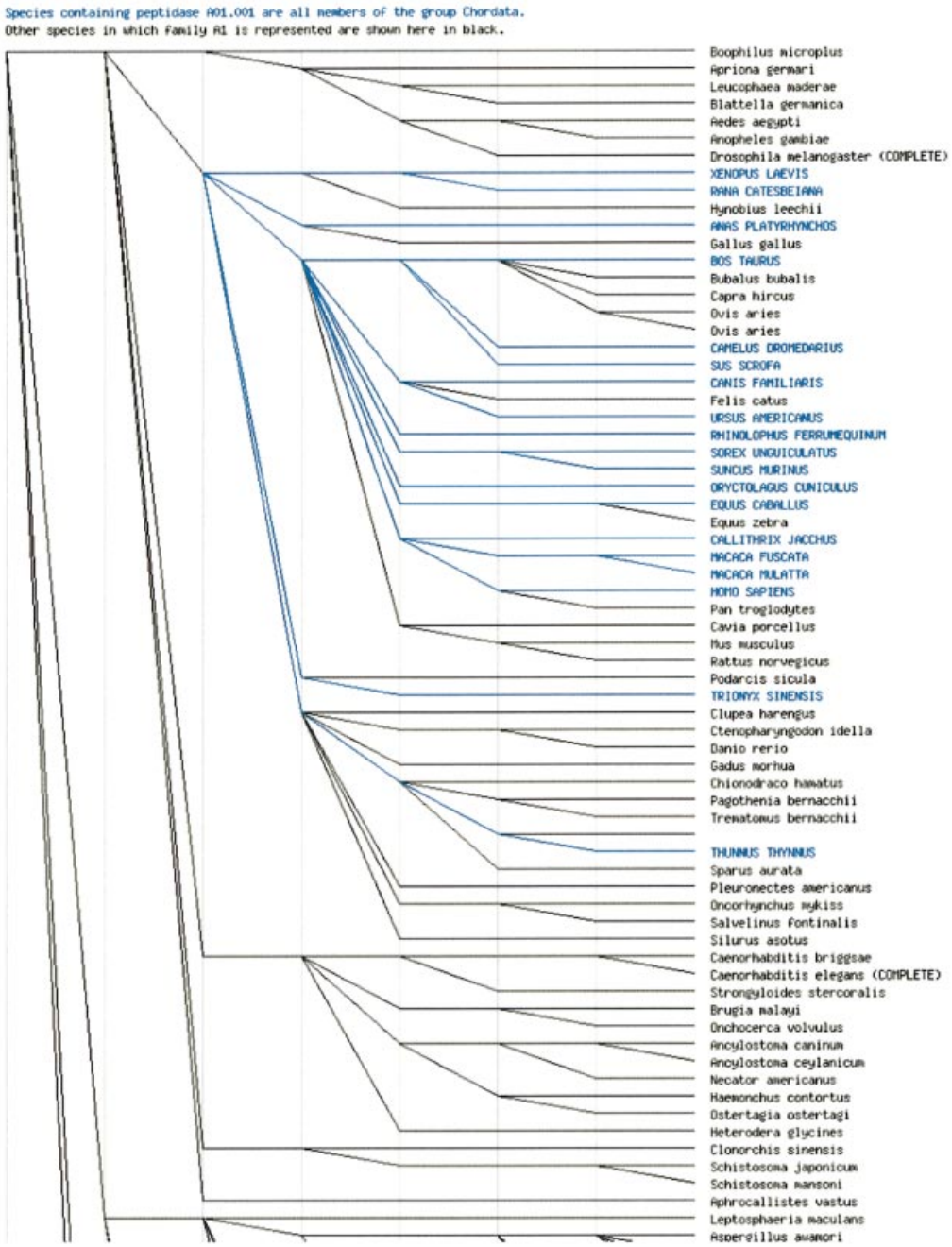

Figure 2. Distribution tree for pepsin A. All organisms with genes encoding proteins in family A1 are included but only a portion of the tree is shown. The relationship between the organisms is shown graphically, with each tip representing a species and each node representing (from right to left) genus, family, order, class, phylum, kingdom and domain (or superkingdom). An organism name is highlighted in blue if one of its homologues is considered to be pepsin A.

DISTRIBUTION TREES

A button from the peptidase or inhibitor summary page gives access to a 'distribution tree'. This shows a graphical display of the relationships between all the organisms from which a member of the peptidase or inhibitor family is known. The organism name is highlighted in blue if it has been reported to contain the particular peptidase or inhibitor in question. Figure 2 shows the distribution tree for pepsin A (A01.001). To be included in the same identifier, we require that there be no important difference in activity. Puente *et al*. (10) have recently shown that human pepsin A and mouse pepsin F are encoded by syntenic genes. Pepsin A is a gastric, digestive enzyme whereas pepsin F is predominantly expressed in the placenta and only occurs in the stomach of embryos and neonates. These differences justify assignments to different identifiers in MEROPS because they are likely to be significant to the users of the database. Hence no rodent names are

Some known substrate cleavages for A01.001

Substrate	Cleavage Site	P4	P3	P2	P1	P1'	P2'	P3'	P4'	Reference
blocked synthetic dipeptide	Glu + Tyr	-	-	NBk	Gly	Tyr	CBk	-	-	Tang, 1998
insulin B chain (oxidized)	Phe-Val-Asn-Gln-His-Leu-Cya-Gly-Ser-His-Leu + Val-Glu-Ala-Leu-Tyr-Leu-Val-Cya-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Ala	Gly	Ser	His	Leu	Val	Glu	Ala	Leu	Murakami-Murofushi, 1998
insulin B chain (oxidized)	Phe-Val-Asn-Gln-His-Leu-Cya-Gly-Ser-His-Leu-Val-Glu-Ala-Leu + Tyr-Leu-Val-Cya-Gly-Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Ala	Val	Glu	Ala	Leu	Tyr	Leu	Val	Cya	Murakami-Murofushi, 1998
insulin B chain (oxidized)	Phe-Val-Asn-Gln-His-Leu-Cya-Gly-Ser-His-Leu-Val-Glu-Ala-Leu-Tyr-Leu-Val-Cya-Gly-Glu-Arg-Gly-Phe + Phe-Tyr-Thr-Pro-Lys-Ala	Glu	Arg	Gly	Phe	Phe	Tyr	Thr	Pro	Murakami-Murofushi, 1998
Lys-Pro-Ala-Glu-Phe-NPh-Arg-Leu	Lys-Pro-Ala-Glu-Phe + NPh-Arg-Leu	Pro	Ala	Glu	Phe	Nph	Arg	Leu	-	Dunn *et al.*, 1986
Lys-Pro-Arg-Arg-Pro-Tyr-Ile-Leu-Lys-Arg-Gly-Ser-Tyr-Tyr-Tyr	Lys-Pro-Arg-Arg-Pro-Tyr-Ile-Leu + Lys-Arg-Gly-Ser-Tyr-Tyr-Tyr	Pro	Tyr	Ile	Leu	Lys	Arg	Gly	Ser	Carraway *et al.*, 1992
Nph-Ala-Ala-NH2	Nph + Ala-Ala-NH2	-	-	-	Nph	Ala	Ala	NH2	-	Hofmann, 1998
renin substrate tetradecapeptide	Asp-Arg-Val + Tyr-Ile-His-Pro-Phe-His-Leu-Leu-Val-Tyr-Ser	-	Asp	Arg	Val	Tyr	Ile	His	Pro	Murao, 1998
renin substrate tetradecapeptide	Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu + Leu-Val-Tyr-Ser	Pro	Phe	His	Leu	Leu	Val	Tyr	Ser	Murao, 1998
Ser-Gln-Asn-Phe-Pro-Ile-Val-Gln	Ser-Gln-Asn-Phe + Pro-Ile-Val-Gln	Ser	Gln	Asn	Phe	Pro	Ile	Val	Gln	
synthetic pepsin peptide substrate	Lys-Pro-Xaa-Glu-Phe + Nph-Arg-Leu	Pro	Xaa	Gln	Phe	Nph	Arg	Leu	-	Tang, 1998
Z-Gly-Gly-Phe-Phe-4-pyridinium alkoxy	Z-Gly-Gly-Phe + Phe-OP4P	Z	Gly	Gly	Phe	Phe	P4P	-	-	Tang, 1998
Z-Phe-Phe-4-pyridinium alkoxy	Z-Phe + Phe-OP4P	-	-	Z	Phe	Phe	P4P	-	-	Tang, 1998

Figure 3. Substrate card for pepsin A.

highlighted in Figure 2 because pepsin F has been assigned the identifier A01.026.

GENOME TREES

We have added a feature similar to the distribution trees for each peptidase or inhibitor family. The MEROPS team analyses completed genomes as they are released, and the distribution of a peptidase or inhibitor family within a genome can be informative for the evolution of that family. The absence of a gene is as important for this as its presence. Every family summary page now includes a button that links to a 'genome tree'. This display includes a graphical overview of the classification of all the organisms for which the genome has been completely sequenced, and the names of the organisms encoding at least one member of the family are highlighted in blue. For example, in the genome tree for family S16, which is present in nearly every genome, there is an intriguing absence of the family from the primitive eukaryote *Encephalitozoon cuniculi*. This may indicate that the *lon* gene reached higher eukaryotes via a lateral gene transfer from a symbiotic bacterial precursor of the mitochondrion.

SUBSTRATE CARDS

An important property of any peptidase is its substrate specificity. MEROPS has for some time provided data for substrate specificity through CGI searches, but we have now added a button to the peptidase summary page that links the user to a page of known substrate cleavages for the peptidase in question. Each page lists substrates alphabetically and shows up to four residues either side of the scissile bond. The information may help in the design of test substrates and inhibitors, and in distinguishing the peptidase from others. An example substrate page is shown in Figure 3.

PEPLIST

We believe that some general enzyme databases may find it helpful to have a list of peptidases that could be included in their coverage, and we have compiled the Peptidase List, normally abbreviated PepList. In release 6.4 of MEROPS it contains 424 well-characterized peptidases many of which are the type examples of families in MEROPS and/or have published 3D structures.

EST CELL LINES

MEROPS contains alignments of thousands of expressed sequence tag (EST) sequences for mouse and human peptidases, and the alignments and data lists are available from the peptidase summary pages. However, a user may wish to know which peptidases are expressed in a particular tissue or in a particular disease state, perhaps to identify targets for drugs. The new 'EST cell lines' item on the menu bar links to indexes for EMBL library number, tissue, developmental stage and disease. For each library there is a page detailing species (human or mouse), tissue, cell type, developmental stage, sex and disease state, and listing all the peptidases and homologues for which ESTs have been isolated.

LINKING

The MEROPS database makes use of frames, and this has proved to be a problem for users wishing to link from another database to a particular page. To overcome this problem, we have put in place a CGI script that allows a user to link anywhere in the database. The CGI script takes two arguments, the first being a MEROPS identifier for a peptidase, inhibitor, family or clan, and the second being an 'action', which determines which kind of page to display. Valid options for the 'action' parameter include 'summary', 'alignment' and 'tree', linking to the summary page, sequence alignment or the evolutionary tree, respectively. A full list of all 'action' parameters is presented in the FAQ file. The following are example links:

http://www.merops.ac.uk/cgi-bin/merops.cgi?id=CD;action=alignment

http://www.merops.ac.uk/cgi-bin/merops.cgi?id=C14;action=tree

http://www.merops.ac.uk/cgi-bin/merops.cgi?id=C14.001;action=summary

The first of these goes to a clan alignment, the second to a family phylogenetic tree and the third to the summary page for a peptidase.

BLASTP SEARCH

We have implemented a BlastP search (11) that allows a user to search the MEROPS sequence database with an unknown query sequence. The results will show whether the sequence is known to MEROPS and if so, how we classify it. The MEROPS library contains many sequences that have not yet been deposited in the primary sequence databases.

STATISTICS

The MEROPS database release 6.4 includes 34 clans, 179 families and 1748 codes for peptidases; 25 clans, 48 families and 318 codes for peptidase inhibitors. There are 222 diagrams showing protein tertiary structures and over 2000 reference lists. Data are included from 115 whole genome sequences.

ACKNOWLEDGEMENTS

The MEROPS database depends upon financial support from the UK Medical Research Council and the Wellcome Trust.

REFERENCES

1. Rawlings,N.D. and Barrett,A.J. (1999) MEROPS: the peptidase database. *Nucleic Acids Res.*, **27**, 325–331.
2. Rawlings,N.D. and Barrett,A.J. (1993) Evolutionary families of peptidases. *Biochem. J.*, **290**, 205–218.
3. Barrett,A.J., Rawlings,N.D. and O'Brien,E.A. (2001) The MEROPS database as a protease information system. *J. Struct. Biol.*, **134**, 95–102.
4. Rawlings,N.D., O'Brien,E. and Barrett,A.J. (2002) MEROPS: the protease database. *Nucleic Acids Res.*, **30**, 343–346.
5. Rawlings,N.D. and Barrett,A.J. (2000) MEROPS: the peptidase database. *Nucleic Acids Res.*, **28**, 323–325.
6. deBakker,P.I., Bateman,A., Burke,D.F., Miguel,R.N., Mizuguchi,K., Shi,J., Shirai,H. and Blundell,T.L. (2001) HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families. *Bioinformatics*, **17**, 748–749.
7. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L., Studholme,D.J., Yeats,C. and Eddy,S.R. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
8. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
9. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
10. Puente,X.S., Sanchez,L.M., Overall,C.M. and Lopez-Otin,C. (2003) Human and mouse proteases: a comparative genomic approach. *Nature Rev. Genet.*, **4**, 544–558.
11. 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.