Nucleic Acids Research, 1999, Vol. 27, No. 1 325–331

MEROPS: the peptidase database

Neil D. Rawlings* and Alan J. Barrett

MRC Molecular Enzymology Laboratory, The Babraham Institute, Babraham, Cambridgeshire CB2 4AT, UK

Received August 31, 1998; Accepted September 22, 1998

ABSTRACT

The *MEROPS* database (http://www.bi.bbsrc.ac.uk/Merops/Merops.htm) provides a catalogue and structure-based classification of peptidases (i.e. all proteolytic enzymes). This is a large group of proteins (~2% of all gene products) that is of particular importance in medicine and biotechnology. An index of the peptidases by name or synonym gives access to a set of files termed PepCards each of which provides information on a single peptidase. Each card file contains information on classification and nomenclature, and hypertext links to the relevant entries in online databases for human genetics, protein and nucleic acid sequence data and tertiary structure. Another index provides access to the PepCards by organism name so that the user can retrieve all known peptidases from a particular species. The peptidases are classified into families on the basis of statistically significant similarities between the protein sequences in the part termed the 'peptidase unit' that is most directly responsible for activity. Families that are thought to have common evolutionary origins and are known or expected to have similar tertiary folds are grouped into clans. The *MEROPS* database provides sets of files called FamCards and ClanCards describing the individual families and clans. Each FamCard document provides links to other databases for sequence motifs and secondary and tertiary structures, and shows the distribution of the family across the major kingdoms of living creatures. Release 3.03 of *MEROPS* contains 758 peptidases, 153 families and 22 clans. We suggest that the *MEROPS* database provides a model for a way in which a system of classification for a functional group of proteins can be developed and used as an organizational framework around which to assemble a variety of related information.

INTRODUCTION

Peptidase is the term recommended by the International Union of Biochemistry and Molecular Biology for all proteolytic enzymes (1), and the peptidases represent a large group of proteins that are of exceptional importance in medicine and biotechnology. The great number of these enzymes is illustrated by an analysis of completely sequenced genomes and a major sequence database (Table 1); this shows that the peptidases represent ~2% (and possibly more) of all gene products. The importance of these enzymes is perhaps best illustrated by the rapid appearance of publications about them, which we have estimated at 8000 per annum (2). Obviously the wealth of data being produced on peptidases needs to be efficiently stored and retrieved but this is hindered by difficulties of nomenclature and classification. Most enzymes are named and classified on the basis of the reactions they catalyse but this has not proved possible for peptidases, because the specificities of enzymes hydrolysing proteins are typically almost impossible to determine rigorously or to describe in a simple name. For this reason, trivial names have to be used for most proteolytic enzymes, but these can easily lead to confusion if their use is not coordinated.

Table 1. About 2% of all gene products are peptidases

Source	ORFs or sequences	Peptidases	Peptidases (%)
Archaea			
Archaeoglobus fulgidus	2471	17	0.45
Methanobacterium thermoautotrophicum	1855	18	0.92
Methanococcus jannaschii	1692	17	0.71
Pyrococcus horikoshii	2061	16	0.78
Bacteria			
Aquifex aeolicus	1512	24	1.32
Bacillus subtilis	4000	94	2.35
Borrelia burgdorferi	863	23	1.27
Escherichia coli	4288	94	1.82
Haemophilus influenzae	1740	44	2.70
Helicobacter pylori	1590	25	1.32
Mycoplasma genitalium	470	10	2.34
Mycoplasma pneumoniae	677	11	1.62
Synechocystis sp.	3168	45	1.52
Treponema pallidum	1041	21	2.01
Eukaryota			
Saccharomyces cerevisiae	6034	93	1.49
Total for genomes	30 462	552	1.65
SWISS-PROT database	74 205	1944	2.62

Data for the numbers of predicted open reading frames (ORFs) in 15 completely sequenced genomes are from refs 35–38, and the numbers of peptidases are those to be found in the *MEROPS* database. Approximately 40% of the ORFs in these genomes are unidentified (35) and some of these are likely to encode additional as yet unrecognized peptidases. The percentage of peptidases in the SWISS-PROT database is based upon the total number of entries in Release 36 of SWISS-PROT plus updates to 12 August 1998 and the number of entries in the *MEROPS* database that have unique SWISS-PROT accession numbers (to account for RNA virus polyproteins that may include more than one peptidase).

*To whom correspondence should be addressed. Tel: +44 1223 832312; Fax: +44 1223 837952; Email: neil.rawlings@bbsrc.ac.uk

The difficulty of defining and describing the specificities of proteolytic enzymes also affects their classification. This has been recognized in Enzyme Nomenclature (1) where the endopeptidases (the majority of all peptidases) are placed in just five sub-subclasses on the basis of their catalytic types and there is no provision for further subdivision of these. In an attempt to address these problems we have developed the *MEROPS* database. This provides a catalogue of the many names of peptidases and their synonyms as well as a comprehensive structure-based classification and a wealth of links to other databases. Users of the *MEROPS* database are asked to cite the present article.

DEFINITIONS

MEROPS is a database of peptidases that are classified in families, and the families are then grouped into clans. The terms *peptidase*, *family* and *clan* are used as follows.

Peptidase

A *peptidase* is a protein that causes the hydrolysis of peptide bonds (see Appendix, note 1). In any single species of organism there is normally a one-to-one relationship between peptidases and their genes, and the paralogous products of different genes in the same genome are treated as different peptidases (see Appendix, note 2). But a single peptidase may well include the products of the allelic variants of a single gene as well as products of alternative RNA splicing and molecular variants produced by post-translational modifications of the protein. Moreover, it may be expressed in different parts or developmental stages of the organism.

Any one peptidase is expected to occur in multiple species of organism, and the species variants are orthologues. The criteria that we use to recognize the orthologous forms of a single peptidase from different species are as follows:

(i) They have similar properties to enzymes, showing the same types or specificities of catalytic activity, pH optima and sensitivity to inhibitors (see Appendix, note 3). (When amino acid sequences are available but no data on specificity, we require that there are no differences in the sequences that would be predicted to result in differences of specificity from what is known of structure/activity relationships in the family.)

(ii) They have similar amino acid sequences throughout the length of the polypeptide encoded by the open reading frame.

(iii) An evolutionary tree for the peptidase units (see below) shows the sequences to have diverged at about the same time as the organisms in which they occur. A much earlier divergence time would imply that they are separate enzymes, not species variants of a single one.

Some peptidases are known only from nucleotide sequence data, and when the hypothetical protein is not close enough to be treated as a form of any peptidase that is already known experimentally then it is classified as a separate hypothetical peptidase.

Family

Each family of peptidases is formed around a founding member or *type example* such as cattle pancreatic trypsin in family S1. The family is then built up by adding peptidases that show statistically significant similarity in amino acid sequence either to the type example or to another existing member of the family. The relationships are therefore transitive, but we stipulate that they must be in the part of the molecule that is responsible for peptidase activity, which we term the *peptidase unit*. It is important to work with the peptidase unit because many peptidases are chimeric proteins containing additional, non-peptidase domains (including pre- and pro-peptides) that are shared with other groups of proteins. For each type example the extent of the peptidase unit has been determined. At most this is that part of the sequence that aligns with the sequence of the smallest mature peptidase molecule in the family. The supplementary domains that are present in the chimeric peptidases are most commonly attached at the N- or C-terminus, but they may also interrupt the peptidase unit as in gelatinase A (4). There are some peptidase families in which even the smallest mature peptidase can be seen to be a multidomain protein from the presence of a segment that is homologous to a known non-peptidase segment found in other proteins. An example is endopeptidase La (family S16) which has an N-terminal ATP-binding domain. Such a domain is excluded from the peptidase unit.

We have compared amino acid sequences by use of the RDF (5) and BLAST (6) programs. Strict statistical criteria are applied so that there is no realistic possibility of two peptidases being placed in the same family when they are not truly homologous [in the definition of Reeck *et al*. (7)]. Sequences matching with a *z* value >6.00 are considered to be homologous.

Subfamily. Some families contain two or more distinct groups of peptidases that differ greatly in primary structure. The structure of such a family may be graphically represented in a phylogenetic tree, and the groups that form the branches separated by deep divergences are recognized as subfamilies. The depth of the divergence between the subfamilies typically corresponds to 67% sequence difference, which is equivalent to ~150 accepted point mutations per hundred residues (8). At an average mutation rate of 0.60 substitutions per amino acid site per 10^9 years [the rate of evolution of the S1 family (9) which is typical], this represents a divergence time of 2.5×10^9 years ago, before the appearance of the earliest known eukaryotes. A subfamily is always founded with a known peptidase, not a hypothetical peptidase. In Release 3.03, 13 families are divided into subfamilies.

Clan

There is good reason to think that proteins derived from a single ancestral peptidase sequence have often diverged so far that they now represent several families in modern organisms. When we can recognise such a group of related families we place them together in a group termed a *clan*. The best evidence to support the formation of a clan is similarity in three-dimensional structures when the data are available, but the arrangement of catalytic residues in the polypeptide chains and limited similarities in amino acid sequence around the catalytic amino acids are also taken into account, so that the assignment of a family to a clan normally depends upon several lines of evidence. Sometimes the assignment has to be considered provisional.

MEROPS identifiers

The *MEROPS* system provides a simple identifier by which to refer to any clan, family or peptidase. This helps to deal with the

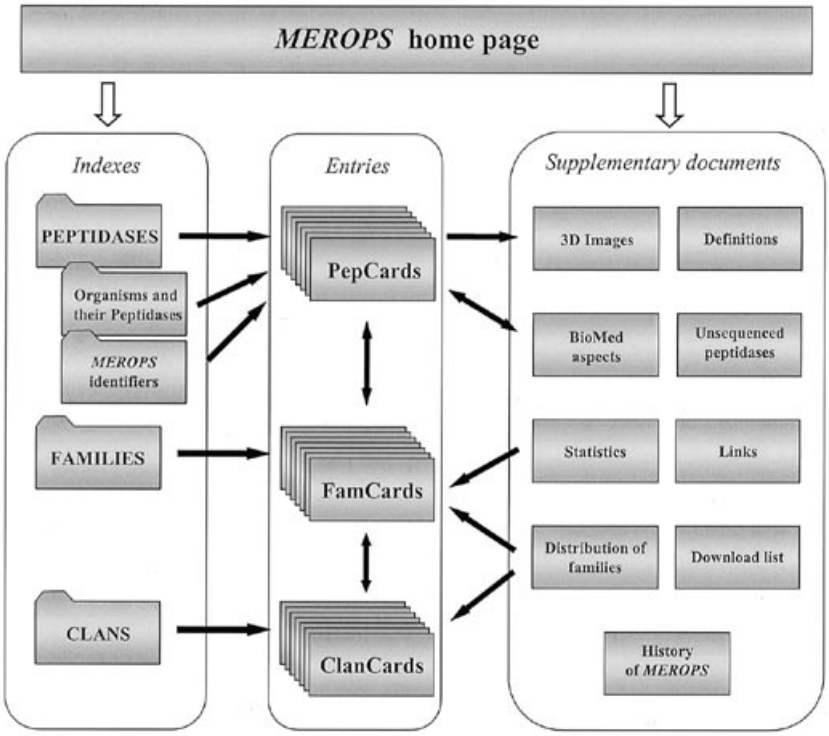

Figure 1. The structure of the *MEROPS* database. Open arrows indicate links from the home page to all the documents in the groups of Indexes and Supplementary documents. The solid arrows show links in one or both directions between individual documents.

problems of nomenclature that beset the study of proteolytic enzymes.

The identifiers of families and clans follow the system introduced by Rawlings and Barrett (10). The identifier of each family starts with a letter denoting the catalytic type of the peptidases it contains, i.e. A, C, M, S, T or U for aspartic, cysteine, metallo, serine, threonine or unknown type (1). The threonine type is the most recently discovered (11). The family identifier is completed with a number that is assigned sequentially as the families are recognized in the *MEROPS* classification. For example, the family of trypsin is called family S1, and that containing signal peptidase is S26. The numbers are simply accession numbers, and have no special significance. Subfamily names are derived from the family identifier with the addition of a letter also assigned sequentially, e.g. M10A, M10B and M10C for the subfamilies of M10.

The name of each clan also starts with a letter based on the catalytic type or types of the families it contains and is completed by a second letter of the alphabet assigned sequentially as the clan is recognised. A few clans contain families of more than one of the types C, S and T, and the letter P is used in forming the identifier of these. An example is clan PA that contains several families of serine peptidases but also contains families of cysteine peptidases from RNA viruses.

If a family or clan disappears, usually because it is merged with another, the identifier is not re-used, and for this reason there are interruptions in the sequence of identifiers that are of no current significance.

A *MEROPS* identifier for each peptidase is constructed from the family identifier padded to three characters when necessary, followed by a decimal point and a three-digit number. An example would be S01.151 for trypsin in family S1.

ORGANIZATION OF THE DATABASE

Figure 1 summarizes the structure of the database, showing the kinds of files that are included and the ways in which they are linked together. It can be seen in the center of the figure that there are three sets of entries, or 'card' files, corresponding to the three levels of classification in the database: peptidases (PepCards), families (FamCards) and clans (ClanCards). Shown to the left are the five indexes by which the card files may be accessed, and to the right is a set of supplementary documents some of which are linked to the card files.

Peptidase level

At the peptidase level of the database there are three indexes that provide access to the PepCard files dealing with individual peptidases.

PEPTIDASES index. In this file the peptidases are listed alphabetically by Peptidase name, and additional columns show Other Names and *MEROPS* Identifiers. The peptidase name is that recommended in the EC List of IUBMB (1) if the peptidase is one that is already included there. The other names are ones that

Caspase-1

MEROPS classification			
Clan: CD	Family: C14	Subfamily: -	MEROPS ID: C14.001
IUBMB classification			
3.4.22 Cysteine endopeptidases		EC Number: EC 3.4.22.36	
Nomenclature			
Recommended Name	caspase-1		
Other names	interleukin 1-beta-converting enzyme		
Other information			
For a review, see *Handbook of Proteolytic Enzymes* chapter 248			
Catalytic type	Cysteine		
Links:	Functional relevance	3D Image	

Human Genetics			
Gene names	*Chromosome location*	*GDB*	*OMIM*
IL1BC; IL1BCE; CASP1	11q22.2-q22.3	CASP1	147678

Protein Sequence Data		
Species (and isoform)	*SwissProt*	*PIR*
Gallus gallus	O42284	
Homo sapiens	P29466	A42677
		A56084
Mus musculus	P29452	A46495
		I48911
Rattus norvegicus	P43527	I53300

Nucleic Acid Sequence Data			
Comment	*EMBL*	*GenBank*	*CDS*
Gallus gallus			
cDNA	AF031351	AF031351	G2642241
Homo sapiens			
cDNA	M87507	M87507	G186286
cDNA	X65019	X65019	G33793
isoform beta	U13697	U13697	G717040
isoform delta	U13699	U13699	G717044
isoform epsilon	U13700	U13700	G717046
isoform gamma	U13698	U13698	G717042
sequence from patent; unannotated	I11817	I11817	
sequence from patent; unannotated	I17647	I17647	
unannotated	B39235	B39235	
Mus musculus			
cDNA	L03799	L03799	G198380
cDNA	L28095	L28095	G515630
complete gene	U04269	U04269	G476218
Rattus norvegicus			
cDNA	U14647	U14647	G555922
mid-section	S79676	S79676	G1172241
mid-section	U34621	U34621	G1002919

Tertiary Structure Data				
Comment	*Resolution*	*PDB*	*DSSP*	*RASMOL*
Homo sapiens				
mature peptidase	2.6 A	1ICE	1ICE	1ICE
complex with inhibitor ac-Trp- Glu- His- Asp	2.73 A	1IBC	1IBC	1IBC

Figure 2. A typical PepCard. The PepCard for caspase-1 is shown. Active links are shown as underlined text.

have been used in the literature, but it should be noted that many of these are obsolete and even ambiguous or misleading. Included in the index are the names of a number of peptidases for which sequence data are not yet available.

Organisms and their Peptidases index. This is an alphabetical list of the scientific binomial names of species of organisms, against each of which are given the names of the peptidases that have been sequenced from that species together with their *MEROPS* identifiers.

MEROPS Identifiers index. This index provides access to the PepCard from a known *MEROPS* identifier.

PepCard entries. A PepCard is provided for each unique peptidase. In addition, there may be two special PepCard files included in the set for a family that are titled 'Other peptidases' and 'Non-peptidase homologues'. Other peptidases is a collection of hypothetical peptidases that are not orthologues of known peptidases, but for which too little information is yet available to merit separate cards. The Non-peptidase homologues card lists a set of proteins that are homologues of the peptidases in the family but are unlikely to be active peptidases because there are mutations of one or more of the catalytic residues.

A PepCard contains data for a single peptidase as defined above and thus includes data for isoforms of the enzyme and all species variants that appear to be orthologues, including otherwise

unknown homologues that we have detected by searching the sequence databases with sequences of the peptidase units of the type examples, using the programs TFASTA (5) and BLAST (6). A typical PepCard is shown in Figure 2. The first section concerns classification and nomenclature, with internal links to the relevant FamCard and ClanCard. There is also an external link to the IUBMB peptidase nomenclature document where enzymological information may be found. (The URLs for the home pages of all the external internet sites are given in Table 2.) There may also be a link to the supplementary BioMed document giving comments on biological, medical and biotechnological aspects of the peptidase, and to a supplementary document containing a colour molecular image of the peptidase displayed as a rendered Richardson plot (12) generated by the MOLSCRIPT (13) and RENDER (14,15) programs.

The second section concerns human genetics and contains data from the human genes database GDB (16) linking via the GeneCard (17), and to the Online Mendelian Inheritance in Man database (OMIM) (18).

The third section of the PepCard deals with protein sequence data, and lists the scientific binomials of the species from which the peptidase is derived and accession numbers for the SWISS-PROT/TREMBL (19) and PIR (20) databases. The species name is linked to the National Center for Biotechnology Information taxonomy database, and the sequence database accession numbers are linked directly to the relevant database entries. The links to the SWISS-PROT database are via the ExPASy server (21), whereas the links to the TREMBL and PIR databases are via the Sequence Retrieval System (SRS) (22) server at the European Biotechnology Institute (EBI) at Hinxton, Cambridgeshire, UK.

The fourth section deals with nucleic acid sequence data. Included in this section are the species of organism (with the same link to the taxonomy database as above), a comment, accession numbers for EMBL (23) and GenBank (24) databases, and the protein identification number (PID) for the protein sequence translation. The comment briefly describes the nature of the entry, for example whether it includes several open reading frames, whether the sequence was obtained from a patent, whether the sequence is a fragment, intron or exon of a peptidase gene, and the name of the strain or isolate if the sequence was derived from a virus. The database accession number links to the EMBL database entry via the SRS server at EBI, and to the GenBank database entry via the Entrez browser (25) at the National Center for Biotechnology Information. The PID number links to the relevant coding sequence within the EMBL entry. As a result of redundancy in the nucleic acid databases it is common for there to be several entries for the nucleic acid sequence of a peptidase from a particular organism. In many cases these nucleotide sequences differ to some extent and we treat the sequences as relating to the same orthologue if the derived protein sequences differ by less than 5% (or 10% if the sequences are from different strains or isolates of a virus, or if the sequence is obtained from a patent). The PepCard may include several EMBL entries for one orthologue, but the list is not comprehensive because we have not attempted to collate expressed sequence tag data or all variants of RNA virus polyprotein genes.

The fifth section concerns tertiary structure data, and contains the species (with the taxonomy link as above), a comment, the resolution of the structure in Ångstrom units, the database identifiers for the Protein Data Bank (PDB) (26) and the DSSP (27) secondary structure databases, and a link to download the PDB entry and launch the molecular visualization program Rasmol (28). The link to the PDB database is via the SRS server at EBI and the link to the DSSP database is via SRS at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany.

Family level

FAMILIES index. This index lists all families and subfamilies arranged by catalytic type. The family name, the type example and clan are listed, and there are links to the FamCard and ClanCard files.

Table 2. Internet resources to which the *MEROPS* database is linked

Resource	URL	Reference
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	(30)
Entrez browser	http://www.ncbi.nlm.nih.gov/Entrez/	(25)
ExPASy server	http://www.expasy.ch/	(21)
GeneCard	http://bioinfo.weizmann.ac.il/cards/	(17)
IUBMB Enzyme Nomenclature	http://www.chem.qmw.ac.uk/iubmb/enzyme/	(1)
National Center for Biotechnology Information taxonomy	http://www3.ncbi.nlm.nih.gov/Taxonomy/	(25)
Online Mendelian Inheritance in Man (OMIM)	http://www3.ncbi.nlm.nih.gov:80/Omim/	(18)
Protein Information Resource (PIR)	http://www-nbrf.georgetown.edu/pir/	(20)
Protein Data Bank	http://www.pdb.bnl.gov/)ar	(26)
PROTFAM	http://www.mips.biochem.mpg.de/	(32)
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/index.html	(29)
Sequence Retrieval System (SRS) servers	(EBI:) http://srs.ebi.ac.uk:5000/	(22)
	(EMBL:) http://www.embl-heidelberg.de:80/srs5/	

FamCard entries. There is a FamCard for each peptidase family, and this contains three sections. The first section concerns classification and gives the type example peptidase of the family with details of the limits of the peptidase unit and the SWISS-PROT/TREMBL accession number (as an active link). There are also external links to a number of other protein classification systems including two databases for the classification of tertiary structures, SCOP (29) and CATH (30) and two databases providing alignments for the family, HSSP (31), a database of protein structure-sequence alignments via a link to the SRS server at EMBL, and a link to the PROTFAM database at the Munich Information Centre for Protein Sequences (32). There is also a link to the PROSITE database of sequence motifs (33) maintained at the ExPASy server that takes the user to the descriptions of the motifs.

The second section is an alphabetical list of the peptidases in the family with the *MEROPS* identifiers and links to the relevant PepCards. The third table maps the distibution of peptidases in the family across the different kingdoms of organisms.

Clan level

CLANS index. The index that provides access to the individual ClanCards lists the clans alphabetically and also includes a brief description of the characteristics of each clan and gives an indication of whether a tertiary structure has been solved for any member of the clan.

ClanCard entries. There is a ClanCard for each peptidase clan. This includes a brief description of the clan, a list of the families and subfamilies in numerical order with the type example for each, an indication of whether a tertiary structure has been solved for any member of each family, and links to the relevant FamCards. There is also a table mapping the distribution of peptidases in the clan across the different kingdoms of organisms that has implications for the evolution of the clan.

Other documents

A number of other documents are included at the *MEROPS* site. These include definitions of terms, descriptions of the databases to which links are made, and a history of the previous versions of the *MEROPS* database. There is a document that gives brief descriptions of known physiological functions and biomedical and biotechnological relevance of peptidases. This is arranged in alphabetical order of recommended peptidase name, and is linked to and from PepCard files. A further document presents summary counts of peptidases for each family and catalytic type. Counts shown are numbers of distinctive peptidases, the numbers of peptidases in the IUBMB list, the number of sequences, and the number of peptidases for which tertiary structure data have been released in the Protein Data Bank. A third document maps the distribution of peptidases within each peptidase family across the major groups of organisms. The final document presents a non-redundant list of peptidase sequences from the protein sequence databases SWISS-PROT, TREMBL and PIR. This list can be saved to an ASCII file and used directly in the GCG package (34) for sequence comparison purposes.

CONTENT OF THE DATABASE

Release 3.03 of the *MEROPS* database contains identifiers for 758 individual peptidases from 934 species of organism. The peptidases are grouped into 153 families, 13 of which are further divided into subfamilies. Of the total families, 102 are placed in 22 clans; the remainder cannot yet be assigned. The database contains 1944 identifiers for peptidases from the SWISS-PROT database and 8911 from the EMBL database. The downloadable list of peptidase sequences provided for use with the GCG (34) package contains 4552 accession numbers.

DISCUSSION

The classification of proteins by comparison of amino acid sequences is attractive in principle, but it is greatly complicated in practice by the chimeric nature of many proteins, which have separate domains that show quite different relationships. We have found that focusing on a particular functional group of proteins and classifying these strictly with reference to the part of each protein that is primarily responsible for the function has allowed a coherent system of classification to be developed for the peptidases. We suggest that this method could be applied to other functional groups of proteins.

The *MEROPS* database now provides an organizational framework to which it is relatively easy to add further information on peptidases. We plan to include an annotated alignment and an evolutionary tree for each family. There is also scope for adding data for the specificity, biochemical characteristics, inhibitor profile and a concise bibliography for each peptidase.

APPENDIX

1. The terms ‘enzyme’ and ‘catalyst’ are used loosely in reference to peptidases. Some peptidases are not catalysts in the strict sense because the peptide bond hydrolysed is in the peptidase itself, and the peptidase is inactivated by the cleavage (3).
2. The equivalence between genes and peptidases is inevitably less simple for the minority of peptidases that are hetero-oligomeric proteins composed of more than one gene product, such as the proteasome.
3. The differences in catalytic activities that are sufficient to distinguish different peptidases are roughly similar to those that would apply in the enzyme nomenclature of the International Union of Biochemistry and Molecular Biology (1).

REFERENCES

1 NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology) (1992) Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. Academic Press, Orlando.
2 Barrett,A.J., Rawlings,N.D. and Woessner,J.F. (Eds) (1998) In *Handbook of Proteolytic Enzymes*, Academic Press, London, pp. xxiii–xxiv.
3 Ringe,D. and Petsko,G.A. (1991) *Nature*, **354**, 22–23.
4 Rawlings,N.D. and Barrett,A.J. (1995) *Methods Enzymol.*, **248**, 183–228.
5 Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1435–1441.
6 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
7 Reeck,G.R., de Haën,C., Teller,D.C., Doolittle,R.F., Fitch,W.M., Dickerson,R.E., Chambon,P., Mclachlan,A.D., Margoliash,E., Jukes,T.H. and Zuckerkandl,E. (1987) *Cell*, **50**, 667–667.

8 George,D.G., Barker,W.C. and Hunt,L.T. (1990) *Methods Enzymol*., **183**, 333–351.
9 Nei,M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.
10 Rawlings,N.D. and Barrett,A.J. (1993) *Biochem. J*., **290**, 205–218.
11 Seemüller,E., Lupas,A., Stock,D., Löwe,J., Huber,R. and Baumeister,W. (1995) *Science*, **268**, 579–582.
12 Richardson,J.S. (1985) *Methods Enzymol*., **115**, 359–380.
13 Kraulis,P.J. (1991) *J. Appl. Cryst*., **24**, 946–950.
14 Bacon,D. and Anderson,W.F. (1988) *J. Mol. Graphics*, **6**, 219–220.
15 Merritt,E.A. and Murphy,M.P. (1994) *Acta Crystallogr. Sect. D-Biol. Cryst*., **50**, 869–873.
16 Letovsky,S.I., Cottingham,R.W., Porter,C.J. and Li,P.W.D. (1998) *Nucleic Acids Res*., **26**, 94–99.
17 Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1997) *Trends Genet*., **13**, 163.
18 Brenner,S.E., Chothia,C., Hubbard,T.J.P. and Murzin,A.G. (1996) *Methods Enzymol*., **266**, 635–643.
19 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res*., **26**, 38–42.
20 Barker,W.C., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S.L., Ledley,R.S., Mewes,H.W., Pfeiffer,F. and Tsugita,A. (1998) *Nucleic Acids Res*., **26**, 27–32.
21 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci*., **19**, 258–260.
22 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol*., **266**, 114–128.
23 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res*., **26**, 8–15.
24 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res*., **26**, 1–7.
25 Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) *Methods Enzymol*., **266**, 141–162.
26 Abola,E.E., Sussman,J.L., Prilusky,J. and Manning,N.O. (1997) *Methods Enzymol*., **277**, 556–571.
27 Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
28 Sayle,R.A. and Milner-White,E.J. (1995) *Trends Biochem. Sci*., **20**, 374.
29 Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol*., **247**, 536–540.
30 Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
31 Schneider,R., De Daruvar,A. and Sander,C. (1997) *Nucleic Acids Res*., **25**, 226–230.
32 Mewes,H.W., Hani,J., Pfeiffer,F. and Frishman,D. (1998) *Nucleic Acids Res*., **26**, 33–37.
33 Bairoch,A. and Bucher,P. (1994) *Nucleic Acids Res*., **22**, 3583–3589.
34 Genetics Computer Group (1994) *Program Manual for the Wisconsin Package*, *Version*, ***8***, *September 1994*. University of Madison, Wisconsin.
35 Pennisi,E. (1997) *Science*, **277**, 1432–1434.
36 Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M., Huber,R., Feldman,R.A., Short,J.M., Olsen,G.J. and Swanson,R.V. (1998) *Nature*, **392**, 353–354.
37 Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A., Nagai,Y., Sakai,M., Ogura,K., Otsuka,R., Nakazawa,H., Takamiya,M., Ohfuku,Y., Funahashi,T., Tanaka,T., Kudoh,Y., Yamazaki,J., Kushida,N., Oguchi,A., Aoki,K. and Kikuchi,H. (1998) *DNA Res*., **5**, 55–76.
38 Fraser,C.M., Norris,S.J., Weinstock,G.M., White,O., Sutton,G.G., Dodson,R., Gwinn,M., Hickey,E.K., Clayton,R., Ketchum,K.A., Sodergren,E., Hardham,J.M., McLeod,M.P., Salzberg,S., Peterson,J., Khalak,H., Richardson,D., Howell,J.K., Chidambaram,M., Utterback,T., McDonald,L., Artiach,P., Bowman,C., Cotton,M.D. and Venter,J.C. (1998) *Science*, **281**, 375–388.