Nucleic Acids Research, 2002, Vol. 30, No. 1 343–346

MEROPS: the protease database

Neil D. Rawlings*, Emmet O'Brien and Alan J. Barrett

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

Received September 27, 2001; Accepted October 2, 2001

ABSTRACT

The *MEROPS* database (http://www.merops.ac.uk) has been redesigned to accommodate increased amounts of information still in pages of moderate size that load rapidly. The information on each PepCard, FamCard or ClanCard has been divided between several sub-pages that can be reached by use of navigation buttons in a frame at the top of the screen. Several important additions have also been made to the database. Amongst these are CGI searches that allow the user to find a peptidase by name, its *MEROPS* identifier or its human or mouse chromosome location. The user may also list all published tertiary structures for a peptidase clan or family, and search for peptidase specificity data by entering either a peptidase name, substrate or bond cleaved. The PepCards, FamCards and ClanCards now have literature pages listing about 10 000 key papers in total, mostly with links to MEDLINE. Many PepCards now include a protein sequence alignment and data table for matching human, mouse or rat expressed sequence tags. FamCards and ClanCards contain Structure pages showing diagrammatic representations of known secondary structures of member peptidases or family type examples, respectively. Many novel peptidases have been added to the database after being discovered in complete genomes, libraries of expressed sequence tags or data from high-throughput genomic sequencing, and we describe the methods by which these were found.

INTRODUCTION

The *MEROPS* database provides a catalogue and a 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 (1). A CGI search and an index of the peptidases by name or synonym gives access to a set of files termed PepCards (e.g. caspase-1) each of which provides information on classification and nomenclature for a single peptidase. Also provided is an interface to the relevant entries in databases for human genetics, protein and nucleic acid sequence data and tertiary structure data, and if the tertiary structure of the enzyme has been determined, a Richardson diagram (2) and a secondary structure schematic are shown. 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 because they have similar tertiary folds are grouped into clans. The *MEROPS* database provides sets of pages called FamCards (e.g. C14) and ClanCards (e.g. CD) describing the individual families and clans. Each FamCard page provides links to other databases of sequence motifs and secondary and tertiary structures, and shows the distribution of the family across the major kingdoms of living creatures, a sequence alignment and a cladogram.

DATABASE STRUCTURE

The structure of the database has been redesigned in order to accommodate more information of diverse types. Each ClanCard, FamCard and PepCard is now a two-frame page in which the top frame is a menu of the sub-pages available to be displayed in the lower frame. The ClanCard has 'Summary', 'Structure' and 'Literature' sub-pages. On the Summary page is presented a description of the clan, a list of member families and a table showing the distribution of peptidases in the clan across major kingdoms of organisms. The Structure page contains a diagrammatic representation of secondary structures for selected members of the clan, and the Literature page contains selected references with links to MEDLINE. Each FamCard similarly has Summary, Structure and Literature sub-pages, and also Alignment and Tree (cladogram) pages. Any PepCard may contain Summary, Sequences, Structure, Literature, Human ESTs and Mouse ESTs sub-pages. The Summary page shows nomenclature, classification, descriptions of physiological role and pharmacological relevance, and human and mouse genetics. The Sequences page includes links to public sequence database entries, and to our version of the sequence in FASTA format. The Structure page includes a Richardson diagram and links to entries in the Protein Databank. The Literature page contains a selection of key references with links to MEDLINE. The human and mouse EST pages are described below.

SEARCHES

Previously, navigation to PepCards has been solely through index files. We have now supplemented these with several CGI searches such that the user can reach a PepCard by entering all or part of the peptidase name, the MEROPS identifier or a

*To whom correspondence should be addressed. Tel: +44 1223 496649; Fax: +44 1223 496023; Email: neil.rawlings@bbsrc.ac.uk

public database accession number (GenBank/EMBL, SWISS-PROT, TrEMBL, PIR or PDB accession number, or GenBank/EMBL protein identifier). Other searches allow the user to retrieve all the peptidase homologues on a specified human or mouse chromosome, or to find all the peptidases in a given family or clan for which a tertiary structure has been published.

Peptidase specificity

Many users of the database have requested an algorithm that would predict cleavage sites in a given protein sequence. Unfortunately it is scarcely possible to provide this because the cleavage specificity of the majority of peptidases is extremely complex and poorly understood. Even when the specificity for cleavage of small peptides has been mapped in detail, the knowledge is commonly not transferable to proteins. Important amongst the reasons for this must be the three-dimensional complexity of a protein molecule. In this, many peptide bonds are buried and therefore inaccessible to proteolysis, and even those on the surface may occur in any of a variety of conformations and local environments. So generalisations cannot safely be made, but nevertheless there is a wealth of relevant empirical data, and it is this that we have tried to make accessible in the new searches in *MEROPS*. We have collected published experimental data for peptidase specificity as a basis for three searches in the *MEROPS* database. The first of these answers the question 'What peptidase can cleave this bond?'. The user is presented with a form in which to enter data for up to four residues either side of a scissile bond. The residues are selected from a pull-down list of amino acids and other elements such as the N- and C-terminal blocking groups commonly found in synthetic substrates. It is not necessary to fill all the boxes because defaults are provided. The table of data returned in response to such a query shows substrates known to be cleaved in the specified way as well as the names of the responsible peptidases and links to their PepCards. The second query answers the question 'How may this substrate be cleaved?'. The user is asked to select a substrate from a pull-down list, and a table is returned of cleavages that have been reported for this substrate, again with the name of each peptidase responsible and a link to the PepCard. The third query answers the question 'What cleavages does this peptidase make?', and the user is invited to select a peptidase from the pull-down menu. A list is returned of substrates that have been reported to be cleaved, each showing the cleavage position and residues in the P4 to P4′ subsites. The data underlying the specificity searches will never be complete, but more are being collected as data are published.

SECONDARY STRUCTURE DIAGRAMS

In the evolution of proteins the protein fold (tertiary and secondary structure) is conserved for longer than statistically significant similarity in primary structure. Accordingly, data for protein fold are provided in the *MEROPS* database to support the assignment of peptidases to families and families to clans. The database contains over 150 Richardson diagrams, and we try to emphasise the similarities between homologous enzymes by showing them in comparable orientations. But this is rather subjective, and still it can be difficult to recognise the similarities between peptidases of different families within a single clan. For example, sometimes a non-peptidase domain or a carbohydrate moiety will obscure the peptidase unit in the three-dimensional representation. In a move towards overcoming these problems, and especially as an aid in assigning peptidase families to clans, we have developed a representation of secondary structure that shows structural relationships between peptidases in a simpler and more objective way. The peptidase sequence is plotted as a line on to which are superimposed blocks of red (for α-helices) and green (for β-strands) drawn to scale. Vertical bars indicate the positions of the active site residues. We find that related peptidases tend to show very similar structures around the catalytic residues, even when they are members of different families in the 'twilight zone' of sequence similarity. An example is clan CA, in which we see that the catalytic cysteine is always at the N-terminus of a helix, and the catalytic histidine is always at the N-terminus of a strand. Each FamCard includes a page of diagrams for the known secondary structures of members of the family, and each ClanCard has a page of such diagrams for the type examples of the families in the clan, where the structures are known.

EXPRESSED SEQUENCE TAGS (ESTs)

The *MEROPS* database now includes alignments and data tables for ESTs for human, mouse and rat peptidases. Our method for identifying and classifying EST sequences, known as ESTA, is as follows:

1. All publicly available human, mouse and rat EST sequences are downloaded once every 3 months and stored in separate libraries.
2. Sets of full-length human, mouse and rat query sequences are assembled, each containing protein sequences for all the known peptidases and peptidase homologues from the species, or failing that, from the most closely related mammalian species. Each query set also contains a sequence from another eukaryote if no mammalian homologue is known from the peptidase subfamily.
3. The entire EST library is searched with each query sequence for the relevant species by use of the TFASTX program (3), and the results are stored.
4. The results are analysed species by species and family by family. Each EST sequence 'hit' is assigned to the code of its closest peptidase homologue, recognised by the lowest 'expect' value. If the query sequence and the EST sequence were from the same species, and the derived protein sequence of the EST is ≥95% identical to the peptidase unit sequence of the query protein, then the assignment is category A. Alternatively, if the identity is <95%, but the EST sequence overlaps the peptidase unit by at least 30 residues, then the match is category B. The third category, C, is that in which the query sequence was from a different species to the EST.

We consider that an EST making a category A match almost certainly represents mRNA transcribed from the same gene as the query sequence, but it may show sequence errors in the published query sequence, residues affected by polymorphism, or splicing variations. A category B match may simply indicate an error-prone EST sequence, but it may also be an EST derived from a novel gene of the same family, and the distinction between these possibilities can be made with the help of the multiple alignment (see below). Finally, a category C match with a query sequence from another species probably

represents a so far undescribed orthologue of the query sequence in this family.

The results for each peptidase are presented as one or two multiple sequence alignments and tables. In the situation in which a query sequence of the same species was used, the PepCard will normally include an alignment and a table for the category A matches, and a second alignment and table for category B. When the peptidase was not yet known from the species, so a query sequence from another species had to be used, the ESTs matching in category C are shown in a single alignment and table.

Each multiple alignment shows the query sequence at the top, highlighted in green, and has the active site residues colour-coded, and the regions outside the peptidase unit shaded. Amino acid residues that differ from the query sequence are printed in red. Repeating patterns of the same red-printed residues are particularly helpful in revealing significant mismatches that are not the result of sequencing errors, and show errors in the query sequence, polymorphisms, alternative splicing or novel homologues.

In the data table associated with each EST alignment, the EST sequences are grouped according to the cell libraries from which they were derived. The table has columns for tissue, cell type, developmental stage, sex, disease state, EMBL/GenBank accession number with an active link, and a key to the UniGene cluster assignment.

DETECTION OF PREVIOUSLY UNDESCRIBED PEPTIDASE HOMOLOGUES

The novel (previously unidentified) peptidase homologues that are shown in each release of *MEROPS* may have been found in any of three different data sources: completed prokaryotic genomes, mammalian EST libraries or eukaryotic data from high-throughput genomic (HTG) sequencing. The methods we use to search these are as follows.

Prokaryote genomes

First, a set of query sequences is assembled. This is a complete set of the protein sequences of the peptidase units of the type examples of the peptidase subfamilies recognised in *MEROPS*. If available, a library of all the protein sequences for the organism is downloaded from the FTP site at the National Center for Biotechnology Information (NCBI). A series of FASTA (4) searches is then run with each query sequence against the microbial library. Hits with an expect value of 0.01 or less are saved, and the complete protein sequence is extracted from the microbial library. A second series of FASTA searches is then performed against a library of all peptidase unit sequences. This enables us to find exact matches and to determine closest homologues. Any new peptidase homologue is catalogued and the peptidase unit sequence added to the library.

In the second stage of the analysis, the complete chromosomal and/or plasmid DNA sequence is downloaded in FASTA format from the FTP site at NCBI and saved as a sequence library. A series of TFASTX searches is then run with each query sequence against the microbial DNA. Again, hits with an expect value of 0.01 or less are saved, and the protein sequence is extracted from the TFASTX alignment. These are also run against the library of peptidase units, and this time sequences that do not match exactly are taken to represent peptidase homologues not recognised by the depositors of the microbial genome. Because microbial genomes do not have the complication of introns, we assume the range of the newly discovered gene to be from the first initiation codon following a stop codon to the next stop codon. If the newly discovered homologue appears to be that of a pseudogene then only the sequence detected in the TFASTX search is stored.

EST sequences

The protein sequence of a previously unidentified peptidase homologue is often easy to extract from the alignment of EST sequences in the ESTA system. In a category B or C alignment, the predicted protein sequence for an EST tends to overlap others, and it is often possible to assemble the complete sequence of the peptidase unit by copying and pasting. Because EST sequences frequently contain errors, this method is used only when there are sufficient EST sequences that a convincing consensus sequence can be assembled.

HTG sequences

A set of peptidase unit protein sequences from the type example of each peptidase subfamily is assembled. Each is used as a query sequence to search the non-redundant nucleic acid sequence library at NCBI in a TBLASTN search (5). For each TBLASTN results file, all hits having an expect value of 0.01 or less are saved, but the list is further filtered to exclude sequences with identifiers that have already been collected and any HTG sequences that have not yet been finished (phases 0, 1 and 2). Each results file is further analysed to collect new deposits of known gene products. Each of the remaining hits is processed so as to obtain as complete a protein sequence as possible. Hits that relate to HTG sequences are analysed by our EGRET system (Genomic Recognition of Exons) in the following way.

1. The protein sequence as determined by TBLASTN is extracted from the results file.
2. The corresponding nucleic acid sequence is downloaded from NCBI.
3. The TBLASTN-derived protein sequence is compared to our library of peptidase unit protein sequences (by use of FASTA) to determine the closest homologue.
4. The full protein sequence of the closest homologue is extracted and cut into segments of 50 residues with 10 residue overlaps. Each segment is stored in FASTA format.
5. A series of TFASTX searches is run with each segment against the nucleic acid sequence of the hit.
6. A map is produced for the hit that shows the nucleic acid sequence with the protein sequence translation below. Where significant similarity was found against the segment of the closest homologue in the TFASTX search, this protein sequence is highlighted in the map. All potential exon/intron junctions are also highlighted on the nucleic acid sequence.
7. Potential exons may be identified as regions of highlighted amino acid sequence with the 3′ end 5′ ends indicated by potential intron/exon junctions. A full or partial coding sequence may then be assembled by matching the phases of successive intron/exon junctions.

STATISTICS

Release 5.6 of the database includes 34 clans, 170 families, 1549 peptidases (of which 207 are as yet unsequenced), 8469 sequences, 25 000 HTML pages and 256 000 external links.

An addition to the Statistics page lists all the organisms for which genome sequencing is complete and presents the number of peptidase homologues in each as a percentage of the total number of protein coding genes.

ACCESS TO THE DATABASE

The *MEROPS* database is made freely available to the academic community, whereas scientists in the commercial sector are invited to purchase licenses to a special edition, *MEROPS-PRO*. The free academic access to *MEROPS* is provided to any computer with an IP address that is registered to an academic Internet domain name, and no username or password is then required. If our server is unable to determine that the IP address is registered to an academic domain name, then the user is asked for a username and password as though he or she were a commercial subscriber. The user is also given the opportunity to request that we register the domain name as academic. Unfortunately, we are not able to provide academic access to any computer connected through a commercial Internet service provider.

REFERENCES

1. Rawlings,N.D. and Barrett,A.J. (1999) *MEROPS*: the peptidase database. *Nucleic Acids Res.*, **27**, 325–331.
2. Richardson,J.S. (1985) Schematic drawings of protein structures. *Methods Enzymol.*, **115**, 359–380.
3. Pearson,W.R., Wood,T., Zhang,Z. and Miller,W. (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.
4. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.