Biochimie 122 (2016) 5–30

Contents lists available at ScienceDirect

Biochimie

journal homepage: www.elsevier.com/locate/biochi

Research paper

Peptidase specificity from the substrate cleavage collection in the *MEROPS* database and a tool to measure cleavage site conservation

Neil D. Rawlings

Wellcome Trust Sanger Institute and the EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK

ARTICLE INFO

Article history:
Received 26 June 2015
Accepted 5 October 2015
Available online 21 October 2015

Keywords:
Peptidase
Substrate
Specificity
Cleavage
Binding pocket
Scissile bond

ABSTRACT

One peptidase can usually be distinguished from another biochemically by its action on proteins, peptides and synthetic substrates. Since 1996, the *MEROPS* database (http://merops.sanger.ac.uk) has accumulated a collection of cleavages in substrates that now amounts to 66,615 cleavages. The total number of peptidases for which at least one cleavage is known is 1700 out of a total of 2457 different peptidases. This paper describes how the cleavages are obtained from the scientific literature, how they are annotated and how cleavages in peptides and proteins are cross-referenced to entries in the UniProt protein sequence database. The specificity profiles of 556 peptidases are shown for which ten or more substrate cleavages are known. However, it has been proposed that at least 40 cleavages in disparate proteins are required for specificity analysis to be meaningful, and only 163 peptidases (6.6%) fulfil this criterion. Also described are the various displays shown on the website to aid with the understanding of peptidase specificity, which are derived from the substrate cleavage collection. These displays include a logo, distribution matrix, and tables to summarize which amino acids or groups of amino acids are acceptable (or not acceptable) in each substrate binding pocket. For each protein substrate, there is a display to show how it is processed and degraded. Also described are tools on the website to help with the assessment of the physiological relevance of cleavages in a substrate. These tools rely on the hypothesis that a cleavage site that is conserved in orthologues is likely to be physiologically relevant, and alignments of substrate protein sequences are made utilizing the UniRef50 database, in which in each entry sequences are 50% or more identical. Conservation in this case means substitutions are permitted only if the amino acid is known to occupy the same substrate binding pocket from at least one other substrate cleaved by the same peptidase.

1. Introduction

In 2007, Barrett & Rawlings [1] proposed a list of criteria to distinguish one peptidase from another. To be considered different, any one of the following bioinformatics tests can be applied: the two peptidases have similar biochemical characteristics but unrelated sequences; the two peptidases have related sequences but different biochemical properties, different domain architectures or the domains are in a different order; or the two peptidases have greater than 50% sequence identity but are derived from nodes on a phylogenetic tree that are not adjacent. In addition, the following biochemical tests can be applied to distinguish two peptidases: the peptidases act under significantly different conditions; the peptidases have different post-translational modifications; the peptidases are sensitive to different inhibitors; the peptidases act on different substrates, or if they act on the same substrates then the cleavage positions are different. It is the last two criteria with which this paper is concerned.

A peptidase cleaves a substrate at the scissile bond, and substrate residues either side of this bond are known as P1 and P1′. Residues towards the N-terminus of the substrate are on the non-prime side, and are numbered P1, P2, P3, P4 and so on. Residues towards the C-terminus are on the prime side and are numbered P1′, P2′, P3′, P4′ and so on. A substrate binding pocket in the peptidase that accommodates a substrate residue is named according to the position the residue occupies in the substrate, except that the “P” is replaced by an “S”. So the S1 binding pocket accommodates the P1 residue, and the S4′ binding pocket accommodates the P4′ residue [2].

A collection of substrate cleavages has been assembled from the scientific literature, annotated, cross-referenced where applicable

E-mail addresses: ndr@ebi.ac.uk, ndr@sanger.ac.uk.

http://dx.doi.org/10.1016/j.biochi.2015.10.003

to the UniProt protein sequence database, and included within the MEROPS database. This collection was originally derived from the CD-ROM version of the first edition of the *Handbook of Proteolytic Enzymes* (1998) [3], which also included a search facility to find the peptidases able to cleave a substrate at a particular position. By knowing where in proteins, peptides or synthetic substrates cleavages occur, it is possible to postulate the specificity of a peptidase. By knowing which amino acids can occupy each substrate binding position, it is also possible to infer whether or not cleavage of a substrate at a particular position is likely to be physiologically relevant from an alignment of protein sequences of closely-related orthologues.

The MEROPS substrate cleavage collection has been widely used to predict cleavages in substrates (for a review see Song et al. (2011) [4]), and to predict what peptidase may be responsible for a known cleavage, for example PROSPER [5]. The MEROPS collection has also been used for the mapping of the human degradome and prediction of "cleavage entropy" as an overall measure of peptidase specificity [6], as well as in the development of the "protease web", the network of peptidase, substrate and inhibitor interactions [7].

This paper describes the MEROPS substrate cleavage collection and the various displays present on the MEROPS website (http://merops.sanger.ac.uk) which aid in understanding peptidase specificity and the processing and degradation of a protein substrate. In order to help determine whether or not a cleavage is physiologically relevant, a service is also described where a user can upload substrate cleavages and receive by E-mail an analysis to show how well conserved, in terms of peptidase binding, each cleavage is.

2. Materials and methods

2.1. Identification of peptidases, homology searching, sequence alignment and phylogenetic tree generation

A peptidase species was defined according to the principles established in Barrett & Rawlings (2007) [1]. The methods for homology searching, family building, and generation of protein sequence alignments and phylogenetic trees are the same as those described in Rawlings et al. (2014) [8]. In brief, the following methods were used. Only the peptidase domain was used for sequence searching and sequence alignment. For each family a *type example* was chosen and for each peptidase species a *holotype* was chosen. The type example and holotype were usually the sequence of the best characterized peptidase in the family or protein species, respectively. A BlastP search [9] of the NCBI non-redundant protein sequence database was performed, using the family type example sequence. Sequences retrieved with an E value of 0.01 or less were considered homologues and included in the family. To find more distant homologues, a HMMER search [10] was performed using a ClustalW alignment [11] of a selection of sequences from the family that included an example from every phylum for which there was a representative. Sequence alignments were built using MAFFT [12]. Phylogenetic trees were built from the family sequence alignment using QuickTree [13].

2.2. Manual substrate cleavage curation

The scientific literature was searched manually for substrate cleavage sites by peptidases. Data were acquired from over 7280 references. The following data were collected, transformed as required and stored in a MySQL database. From the name of the peptidase as given by the authors of the publication, a *MEROPS* identifier and, if possible, a MERNUM indicating the source organism, were assigned. From the name of the substrate and its source, a UniProt accession was assigned where possible, and the name recommended by UniProt was stored in the MySQL database, unless the substrate was a peptide or was processed, in which case a peptide name or a name to indicate that processing had occurred was stored (for example, "Met-enkephalin" would be stored in preference to "pro-opiomelanocortin" if the substrate was just the peptide). Where more than one UniProt entry existed, the annotated SwissProt accession, name and sequence where used in preference. Where isoforms derived from alternative initiation and alternative splicing were indicated in the UniProt database entry, the sequence chosen as the representative sequence by UniProt was selected unless the original publication indicated that a particular isoform had been used. There was no attempt to map a cleavage to all isoforms on the presumption that a change in sequence could lead to a change in cleavage position. The cleavage position (the position of the P1 residue in the substrate) was converted to the equivalent residue number from the respective UniProt entry. Up to four residues either side of the scissile bond (residues P4 to P4′) were stored for each cleavage. The residue range of the substrate used compared to the sequence in the UniProt entry was also stored. This allowed for annotation of peptide substrates derived from full-length proteins and processing events, such as removal of signal and transit peptides and precursor sequences. The CDC checksum for the UniProt entry was also stored so that any changes to the sequence could be identified subsequently. Kinetic data (K_m, K_{cat}, and/or K_m/K_{cat}) were stored where available. Annotations to indicate how the peptidase and cleavage position were identified were also stored. The initials of the curator and the date the cleavage was collected were also stored. The reference was stored in a Reference Manager database (Thomson Reuters) and the PubMed accession was obtained and stored where possible. Any additional data that affected where cleavage occurred, such as reactions conditions, where stored as a comment in the MySQL database.

To ensure that curation was consistent, a Perl program was written to aid cleavage data collection and storage. The user (either the author or a summer student) was asked to enter his or her initials; the UniProt accession of the protein substrate in question; the cleavage position; the residue range of the substrate sequence compared to the UniProt entry; the codes for how the cleavage was identified and how the peptidase was identified; whether the cleavage was physiological, non-physiological, pathological or theoretical; whether the substrate was denatured; the reference and its PubMed identifier; and any comment.

Collection of cleavage data from the literature was also outsourced to Molecular Connections, Bangalore, India. Data were returned to the author as an Excel spreadsheet and a pipeline developed to extract data from the spreadsheet and import it into the MySQL database. Existing substrate cleavage collections were also imported into the MEROPS collection. These included data from the CutDB database [14] and the CASBAH database of caspase substrates [15].

A Perl program to check that the P4–P4′ residues around the cleavage position matched the sequence in the UniProt entry was written as a quality control measure and to identify any subsequent changes in the UniProt sequence.

Cleavage data were also stored for the cleavage of synthetic substrates. These were manually entered into the database. For a synthetic substrate it was not possible to map the sequence to a UniProt database entry. The P4–P4′ positions around the scissile bond were stored where possible (many synthetic substrates do not have residues beyond P1′ or P3), including a unique identifier for each N- or C-terminal blocking or reporter group occurring within that range.

In certain cases, it was not possible to map a cleavage to a single enzyme. This most frequently occurred when cleavage was performed by an enzyme complex, such as the proteasome or