The Degradome database: mammalian proteases and diseases of proteolysis

Víctor Quesada, Gonzalo R. Ordóñez, Luis M. Sánchez, Xose S. Puente and Carlos López-Otín*

Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, 33006-Oviedo, Spain

Received July 28, 2008; Accepted August 21, 2008

ABSTRACT

The degradome is defined as the complete set of proteases present in an organism. The recent availability of whole genomic sequences from multiple organisms has led us to predict the contents of the degradomes of several mammalian species. To ensure the fidelity of these predictions, our methods have included manual curation of individual sequences and, when necessary, direct cloning and sequencing experiments. The results of these studies in human, chimpanzee, mouse and rat have been incorporated into the Degradome database, which can be accessed through a web interface at http://degradome.uniovi.es. The annotations about each individual protease can be retrieved by browsing catalytic classes and families or by searching specific terms. This web site also provides detailed information about genetic diseases of proteolysis, a growing field of great importance for multiple users. Finally, the user can find additional information about protease structures, protease inhibitors, ancillary domains of proteases and differences between mammalian degradomes.

INTRODUCTION

Proteases are defined as hydrolytic enzymes acting on peptide bonds, in a process termed proteolysis. The biological significance of proteolysis has driven the evolutionary invention of multiple, extremely diverse classes and families of proteases. Thus, different proteases are known to play key roles in multiple biological processes, including cell cycle progression, differentiation and migration, morphogenesis and tissue remodelling, neuronal outgrowth, haemostasis, wound healing, immunity, angiogenesis and apoptosis (1). The importance of proteolysis is also apparent in the numerous pathological conditions related to alterations in proteases, including cancer, arthritis, progeria and neurodegenerative and cardiovascular diseases (1–6). The extensive biological and pathological implications of this large set of proteins with a common biochemical function led to the concept of proteases as a distinct subset of the proteome. Thus, the degradome of an organism was defined as the complete set of proteases in that organism (7).

The definition of degradome naturally led to the development of degradomics as a new experimental field which includes all genomic and proteomic approaches for the identification and characterization of proteases that are present in an organism. The completion of multiple genome projects has been instrumental in the advance of degradomics by allowing researchers to extend the degradomes of several species in silico from known protease sequences. While several computer programs allow the automatic prediction of genes based on similarity, a reliable prediction still requires manual curation by trained researchers (8). The reasons for this limitation include the difficulty to detect small or dissimilar exons as well as the occurrence of occasional sequencing errors. Indeed, the analysis of a large set of genes is likely to require manual inspection of sequencing traces and cloning and re-sequencing experiments for some of the genes. Additionally, orthology or paralogy assignment of protease genes between human and other animal models also requires the supervision of expert curators. We have used this manual procedure to predict the degradomes of human, mouse, rat, chimpanzee and platypus (9–12).

Furthermore, our continued effort in degradomics has led us to mine the literature and annotate known relationships between protease alterations and hereditary diseases. Since proteases make up promising drug development targets (13–16) and clinical markers (17–19), this compilation may prove very useful to researchers in different fields of human pathology. Likewise, this information on diseases of proteolysis represents a useful resource to determine the utility and limitations of diverse animal models to recapitulate certain human diseases.

Here we report the Degradome database, which contains the results of the manual annotation of every
protease gene in the genomes of human, chimpanzee, mouse and rat, along with relationships between protease alterations and hereditary diseases. This database complements existing databases devoted to proteases, by providing a different focus. Namely, the database ProLysED (20) is devoted to proteases in prokaryotes, whereas our target organisms are mammals. On the other hand, CutDB (21) focuses on annotation of individual proteolytic events, both actual and predicted, rather than on the proteases themselves. Finally, MEROPS (22) is a comprehensive and excellent database which relies on large-scale experiments and automatic annotation. However, a number of entries in this database correspond to pseudogenes or sequences derived from retroviral elements which do not code for any functional proteolytic enzyme. By contrast, our Degradome database, while less comprehensive in the number of species, relies on manual annotation and exhaustive curation of genes on an individual basis. In multiple cases, this informatic work is supported by direct cloning and sequencing experiments (23–27).

Furthermore, our emphasis in diseases adds important information on the pathological relevance of some proteases, which is not directly available in other databases.

The Degradome database is aimed at researchers looking for specific information about mammalian proteases and protease families. Additionally, we have incorporated features intended to help non-experienced users who want to learn about the degradome. These features include selected publications and interactive 3D structures that can be displayed and manipulated with Acrobat reader and thus do not need specialized software.

DATABASE ACCESS

Annotation of individual proteases

The Degradome database contains information about 570 human, 568 chimpanzee, 651 mouse and 641 rat proteases. All of these proteases can be grouped into five catalytic classes, depending on the key residue for their catalytic mechanism. Accordingly, the database information is structured in five tables, containing aspartyl-, cysteine-, metallo-, serine- and threonine-proteases. Each table displays the name of the family using the MEROPS classification system, the name of each protease and the gene symbol for the protease in each species (Figure 1A). Orthologous genes for each individual organism are easily identified, as well as those pseudogenes for which a functional protease gene is present in at least one of the selected mammalian species. In addition to this summarized view of the tables, every field contains a link providing additional information. Thus, the name of the family leads to a different web page with selected publications intended as a primer for users who wish to gather specific information about that family of proteases. When available, this web page also contains a link to a description of the structural features of the family (see ‘selected structures’).

On the other hand, the protease-specific links open ‘popup’ tables with information about each protease (Figure 1A). Since this information is retrieved via a dynamic script, the user will be offered a link to a static table if Javascript is not enabled in the client browser. The table from the protease name field displays the MEROPS code with a hyperlink to the corresponding MEROPS web page and the predicted cellular localization of the protease. Finally, each of the species-specific cells provides external links to related entries in other databases, i.e. NCBI (28) and Ensembl (29)—genomic locus, gene status (gene/pseudogene in the indicated species) and proteolytic activity, whether the protein encoded by this gene is an active protease or an inactive protease homologue due to changes in key catalytic residues. In addition, if the protease is known to be mutated in a hereditary disease, this information is provided, along with a link to the OMIM database entry describing the disease (30).

Search engine

The Degradome database can also be queried using the specific search engine (http://degradome.uniovi.es/search.html). This option lets the user find proteases which meet user-defined criteria such as the presence or absence in selected species, the localization within a specific chromosomal locus, or the existence of mutations which lead to human hereditary diseases (Figure 1B). To make this process intuitive, all of the possibilities are listed in several ‘dropdown boxes’, so that every query is expressed as a meaningful simple sentence. Users can combine several searches to easily perform moderately complex queries. Thus, once a search has been finished, it can be refined with a second search by setting the first box to ‘keep’. This is equivalent to a logical ‘AND’ between both queries. The results of the second search can also be added to the results of the first search—logical ‘OR’—or removed from the results of the first search—logical ‘AND NOT’.

As an example, to find which human proteases are involved in a disease and located in human chromosome 12q, we can first set the query boxes to ‘Search for proteases containing 12q in the field Locus of Human’. This will retrieve 21 proteases. Then, we can rearrange the query boxes to ‘Keep proteases mutated in a disease’, which will narrow the results to a single hit.

Hereditary diseases of proteolysis

The information about mutated proteases in hereditary diseases has been compiled into a table (http://degradome.uniovi.es/diseases.html), so that users specifically interested in this subject do not need to browse or search the individual annotations. At its present form, the table of degradome-related genetic diseases contains 77 proteases, with information about gene locus, mode of inheritance, pathologic protease alteration (gain/loss of proteolytic activity) and availability of described animal models containing the same protease anomaly (Figure 1C). A link to related OMIM entries is also provided.

To our knowledge, this is the first summary of the relationships between degradomics and pathology. The large number of protease alterations related to diverse diseases highlights the significance of the degradome to maintain a correct physiological balance. It must be noted that this summary does not include the multiple examples of
non-hereditary diseases in which proteases are known to play an important role as a consequence of alterations in their spatio-temporal patterns of expression (2–6). Notably, the degradome database has also demonstrated its usefulness in the analysis of proteases associated with cancer (31).

Selected structures

Proteases represent important pharmacological targets for different human pathologies. Therefore, an important aspect of protease research has been to know the mechanism of action of these enzymes by determining the 3D structure of individual proteases. In this regard, we have prepared 22 web pages showing structural features of representative members of different protease families (Figure 1D). These web pages can be accessed from an index (http://degradome.uniovi.es/structures.html) or from the ‘family’ field in the tables of individual annotations. The figures show the secondary structure elements as ribbons, with catalytic side chains and inhibitors. The user can freely interact with the representations, rotating and moving the structure, zooming in or out, and hiding or showing parts of the protease. These capabilities are provided through portable document format (pdf) files, which are also freely available for download.

These structures have been selected to present a general view of the multiple folds that can be found in the degradome. Most of the structures include a specific inhibitor interacting with the catalytic residues. In contrast, three structures from different catalytic classes have been chosen which display proteases in their active form. In these structures, the user can display schematics explaining the putative catalytic mechanisms of cysteine-proteases.
of a series of protease coding genes in rodents, which might hamper the use of these animal models to study certain human proteases.

Our future plans include the extension of the database to other species. At this moment, we are in the process of adding the degradome of platypus. Other species we are currently studying include non-mammalian metazoa which may provide additional clues about the evolution of proteases. Finally, we plan to offer new features which will allow users to search for sequences and motifs in the degradome.

FUNDING

European Union (CancerDegradome-FP6 and FP7); Ministerio de Ciencia e Innovación-Spain; Fundación M Botín; Fundación Lilly; Obra Social Cajastur (to the Instituto Universitario de Oncologia). Funding for open access charge: Ministerio de Ciencia e Innovación-Spain.

Conflict of interest statement. None declared.

REFERENCES