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

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

*To whom correspondence should be addressed. Tel: +34 985 104201; Fax: +34 985 103564; Email: clo@uniovi.es

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
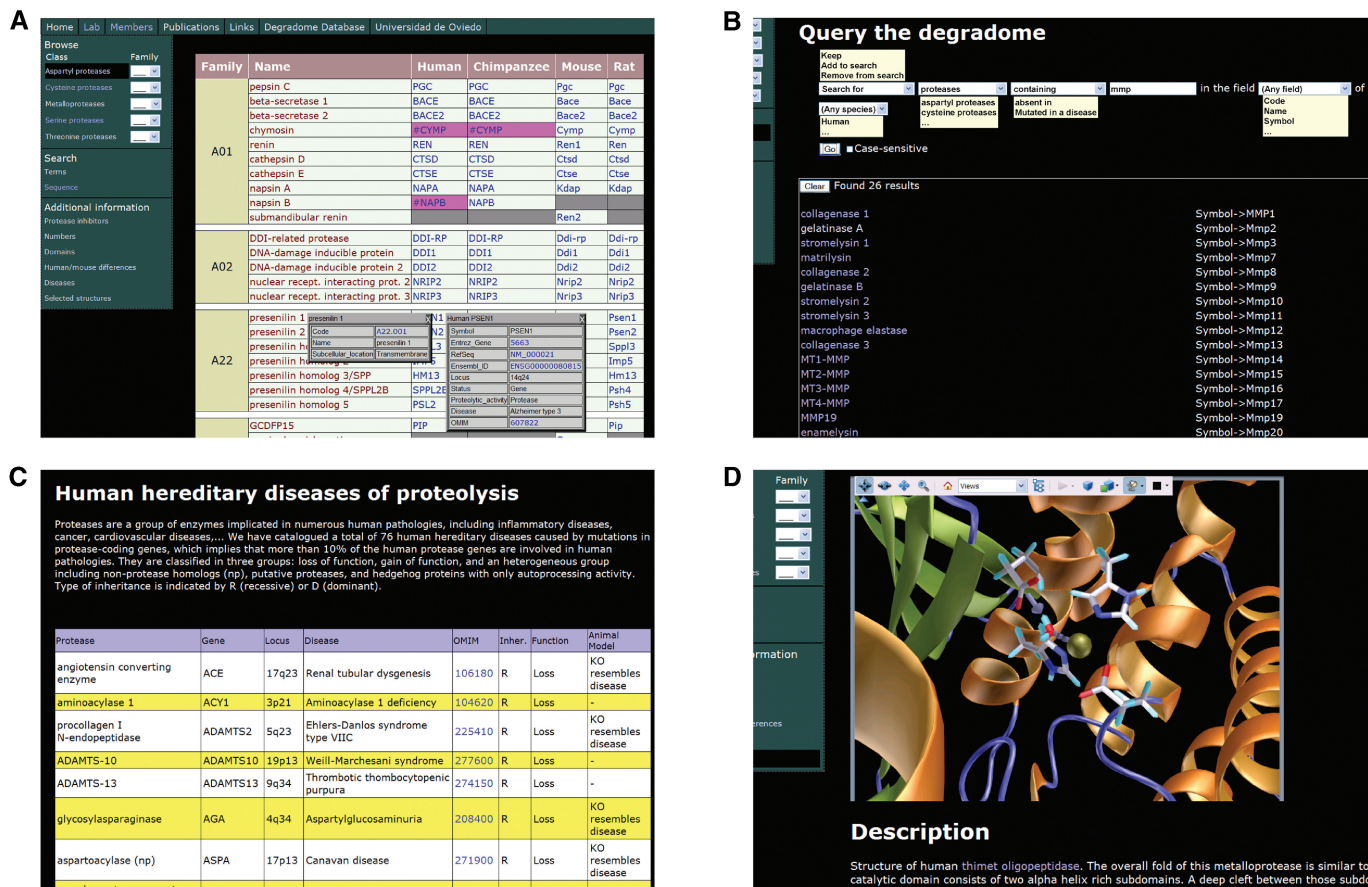
Figure 1. Main features of the Degradome database. (**A**) Individual annotations of aspartyl proteases. The first column contains the name of the family, with a hyperlink to a web page where the user can find related selected publications and structures. The second column contains the name of each protease, with a hyperlink which opens a 'popup' table with further general information—in this example, 'presenilin 1'. The rest of the columns show the symbol of each protease in each organism, with a hyperlink which opens a 'popup' table with species-specific information—in this example, 'human presenilin 1'. Pseudogenes are shown over a 'pink background'. Proteases absent in a species are shown as 'empty grey cells'. (**B**) Search interface. 'Dropdown boxes' are used to define the query. In this example, 26 proteases have been found to contain the term 'mmp' in any field of any species. The 'yellow boxes' have been added to show additional possibilities for each field of the query definition. The search is case-insensitive by default, but an 'option box' exists to make the search case-sensitive. Searches can be combined by 'keeping' results from a previous search, 'adding' results to a previous search or 'removing' results from a previous search. (**C**) Mutated proteases in human hereditary diseases. Each row contains information about gene locus, disease name, OMIM entry, mode of inheritance, pathologic protease alteration (activity gain or loss), and availability of animal models containing the same protease anomaly. (**D**) Interactive structure representation of 'thimet oligopeptidase', from the M03 metalloprotease family. The structure can be rotated, moved, and zoomed. A 'discontinuous bar' shows the coordination of the catalytic water molecule with the zinc ion ('golden sphere'). An 'arrow' shows the electrostatic interaction between the catalytic glutamic acid side chain and the catalytic water molecule.

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

(C26 family), serine-proteases (S10 family) and metallo-proteases (M03 family).

Additional contents

In addition to the Degradome database, the web site also offers several summaries of the characteristics of mammalian degradomes. Thus, a static table listing human, mouse and rat protease inhibitors can be found at http://degradome.uniovi.es/inhibitors.html. A count of proteases in these species, itemized by catalytic class, is shown at http://degradome.uniovi.es/numbers.html. These numbers should not be considered definitive and are likely to be expanded as novel catalytic classes are discovered and added to the Degradome database. Additionally, since most proteases display a series of non-proteolytic domains linked to the catalytic unit, we have also prepared a figure showing the different ancillary domains present in proteases (http://degradome. uniovi.es/domains.html). Finally, we have incorporated a figure summarizing the differences between human and mouse degradomes (http://degradome.uniovi.es/hmd. html). This figure is displayed as a static image and also as an interactive pdf file.

Implementation

The database with the annotations of individual proteases has been stored in five extensible markup language (XML) files, one for each catalytic class. These XML files are freely available upon request. Individual annotation tables access the corresponding XML files through Javascript. Therefore, if the browser lacks Javascript or Javascript is blocked, the user is offered a link to a static table displaying all of the information at once.

On the other hand, several interactive features—i.e. 'selected structures' and 'human/mouse degradome differences'—are offered as pdf files. Thus, the user does not need any *plugins* or specific software to manipulate these figures, only Adobe Reader v7.0 or higher. The web pages contain an external link to a web page where the user can download the last version of Adobe Reader. It must be noted that Microsoft Internet Explorer treats embedded pdf files as ActiveX content, which may be blocked in the browser. If this happens, the user can download the pdf file and access its contents locally.

CONCLUSION AND FUTURE DIRECTION

We have developed a database devoted to the degradome in several mammalian species, which is freely available through a web interface. Notably, this database contains information about the involvement of proteases in genetic diseases. The features provided by the degradome database are useful for researchers in the degradomics field, as well as for researchers working on individual proteases and protease families. It will also be of special interest to researchers working with animal models, as this database provides a highly curated repertoire of orthologous genes between human, mouse and rat. Likewise, the degradome database shows differences in protease genes between these organisms due to the selective expansion 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.

Conflict of interest statement. None declared.

REFERENCES

1. Lopez-Otin,C. and Bond,J.S. (Epub Ahead of print July 23, 2008) Proteases: Multifunctional enzymes in life and disease. *J. Biol. Chem*. doi:10.1074/jbc.R800035200.
2. Freije,J.M., Balbin,M., Pendas,A.M., Sanchez,L.M., Puente,X.S. and Lopez-Otin,C. (2003) Matrix metalloproteinases and tumor progression. *Adv. Exp. Med. Biol.*, **532**, 91–107.
3. Murphy,G. and Nagase,H. (2008) Reappraising metalloproteinases in rheumatoid arthritis and osteoarthritis: destruction or repair? *Nat. Clin. Pract. Rheumatol.*, **4**, 128–135.
4. Varela,I., Cadinanos,J., Pendas,A.M., Gutierrez-Fernandez,A., Folgueras,A.R., Sanchez,L.M., Zhou,Z., Rodriguez,F.J., Stewart,C.L., Vega,J.A. *et al.* (2005) Accelerated ageing in mice deficient in Zmpste24 protease is linked to p53 signalling activation. *Nature*, **437**, 564–568.
5. Nalivaeva,N.N., Fisk,L.R., Belyaev,N.D. and Turner,A.J. (2008) Amyloid-degrading enzymes as therapeutic targets in Alzheimer's disease. *Curr. Alzheimer Res.*, **5**, 212–224.
6. Dollery,C.M. and Libby,P. (2006) Atherosclerosis and proteinase activation. *Cardiovasc. Res.*, **69**, 625–635.
7. Lopez-Otin,C. and Overall,C.M. (2002) Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell. Biol.*, **3**, 509–519.
8. Do,J.H. and Choi,D.K. (2006) Computational approaches to gene prediction. *J. Microbiol.*, **44**, 137–144.
9. Puente,X.S., Sanchez,L.M., Overall,C.M. and Lopez-Otin,C. (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet.*, **4**, 544–558.
10. Puente,X.S. and Lopez-Otin,C. (2004) A genomic analysis of rat proteases and protease inhibitors. *Genome Res.*, **14**, 609–622.
11. Puente,X.S., Gutierrez-Fernandez,A., Ordonez,G.R., Hillier,L.W. and Lopez-Otin,C. (2005) Comparative genomic analysis of human and chimpanzee proteases. *Genomics*, **86**, 638–647.
12. Ordonez,G.R., LaDeana,W.H., Warren,W.C., Grützner,F., Lopez-Otin,C. and Puente,X.S. (2008) Loss of genes implicated in gastric function during platypus evolution. *Genome Biol.*, **9**, R81.
13. Varela,I., Pereira,S., Ugalde,A.P., Navarro,C.L., Suarez,M.F., Cau,P., Cadinanos,J., Osorio,F.G., Foray,N., Cobo,J. *et al.* (2008) Combined treatment with statins and aminobisphosphonates extends longevity in a mouse model of human premature aging. *Nat. Med.*, **14**, 767–772.
14. Gomperts,E.D., Astermark,J., Gringeri,A. and Teitel,J. (2008) From theory to practice: applying current clinical knowledge and treatment strategies to the care of hemophilia a patients with inhibitors. *Blood Rev.*, **22**(Suppl. 1), S1–11.

15. Chauhan,D., Hideshima,T. and Anderson,K.C. (2008) Targeting proteasomes as therapy in multiple myeloma. *Adv. Exp. Med. Biol.*, **615**, 251–260.

16. Turk,B. (2006) Targeting proteases: successes, failures and future prospects. *Nat. Rev. Drug Discov.*, **5**, 785–799.

17. Schoenberger,J., Bauer,J., Moosbauer,J., Eilles,C. and Grimm,D. (2008) Innovative strategies in in vivo apoptosis imaging. *Curr. Med. Chem.*, **15**, 187–194.

18. Sloane,B.F., Sameni,M., Podgorski,I., Cavallo-Medved,D. and Moin,K. (2006) Functional imaging of tumor proteolysis. *Annu. Rev. Pharmacol. Toxicol.*, **46**, 301–315.

19. Paliouras,M., Borgono,C. and Diamandis,E.P. (2007) Human tissue kallikreins: the cancer biomarker family. *Cancer Lett.*, **249**, 61–79.

20. Firdaus Raih,M., Ahmad,H.A., Sharum,M.Y., Azizi,N. and Mohamed,R. (2005) ProLysED: an integrated database and meta-server of bacterial protease systems. *Appl. Bioinformatics*, **4**, 147–150.

21. Igarashi,Y., Eroshkin,A., Gramatikova,S., Gramatikoff,K., Zhang,Y., Smith,J.W., Osterman,A.L. and Godzik,A. (2007) CutDB: a proteolytic event database. *Nucleic Acids Res.*, **35**, D546–549.

22. Rawlings,N.D., Morton,F.R., Kok,C.Y., Kong,J. and Barrett,A.J. (2008) MEROPS: the peptidase database. *Nucleic Acids Res.*, **36**, D320–325.

23. Quesada,V., Diaz-Perales,A., Gutierrez-Fernandez,A., Garabaya,C., Cal,S. and Lopez-Otin,C. (2004) Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases. *Biochem. Biophys. Res. Commun.*, **314**, 54–62.

24. Marino,G., Uria,J.A., Puente,X.S., Quesada,V., Bordallo,J. and Lopez-Otin,C. (2003) Human autophagins, a family of cysteine proteinases potentially implicated in cell degradation by autophagy. *J. Biol. Chem.*, **278**, 3671–3678.

25. Cal,S., Quesada,V., Garabaya,C. and Lopez-Otin,C. (2003) Polyserase-I, a human polyprotease with the ability to generate independent serine protease domains from a single translation product. *Proc. Natl Acad. Sci. USA*, **100**, 9185–9190.

26. Diaz-Perales,A., Quesada,V., Peinado,J.R., Ugalde,A.P., Alvarez,J., Suarez,M.F., Gomis-Ruth,F.X. and Lopez-Otin,C. (2005) Identification and characterization of human archaemetzincin-1 and -2, two novel members of a family of metalloproteases widely distributed in Archaea. *J. Biol. Chem.*, **280**, 30367–30375.

27. Cal,S., Obaya,A.J., Llamazares,M., Garabaya,C., Quesada,V. and Lopez-Otin,C. (2002) Cloning, expression analysis and structural characterization of seven novel human ADAMTSs, a family of metalloproteinases with disintegrin and thrombospondin-1 domains. *Gene*, **283**, 49–62.

28. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–21.

29. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.

30. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–517.

31. Lopez-Otin,C. and Matrisian,L.M. (2007) Emerging roles of proteases in tumour suppression. *Nat. Rev. Cancer*, **7**, 800–808.