Evolution of Peptidase Diversity*1

Michael J. Page and Enrico Di Cera

From the Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110

A wide variety of peptidases associate with vital biological pathways, but the origin and evolution of their tremendous diversity are poorly defined. Application of the MEROPS classification to a comprehensive set of genomes yields a simple pattern of peptidase distribution and provides insight into the organization of proteolysis in all forms of life. Unexpectedly, a near ubiquitous core set of peptidases is shown to contain more types than those unique to higher multicellular organisms. From this core group, an array of eukaryote-specific peptidases evolved to yield well known intracellular and extracellular processes. The paucity of peptidase families unique to higher metazoa suggests a role for the core group in the shaping of proteolysis in all forms of life. The core group, an array of eukaryote-specific peptidases evolved to yield well known intracellular and extracellular processes. The paucity of peptidase families unique to higher metazoa suggests a role for the core group in the shaping of proteolysis in all forms of life. The core group, an array of eukaryote-specific peptidases evolved to yield well known intracellular and extracellular processes. The paucity of peptidase families unique to higher metazoa suggests a role for the core group in the shaping of proteolysis in all forms of life. The core group, an array of eukaryote-specific peptidases evolved to yield well known intracellular and extracellular processes. The paucity of peptidase families unique to higher metazoa suggests a role for the core group in the shaping of proteolysis in all forms of life. The core group, an array of eukaryote-specific peptidases evolved to yield well known intracellular and extracellular processes. The paucity of peptidase families unique to higher metazoa suggests a role for the core group in the shaping of proteolysis in all forms of life.

Proteolysis is a requirement for life. Some peptidases recycle polypeptides into their constituent amino acids, whereas others catalyze selective polypeptide cleavage for post-translational modification. Approximately 700 peptidases are present in the human genome, collectively termed the degradome (1), mediating diverse processes such as extracellular matrix remodeling, immunity, development, protein processing, cell signaling, and apoptosis (2). The MEROPS data base classifies peptidases into clans based on structural similarity or sequence features suggesting such homology (3, 4). Families divide clans by common ancestry. Subfamilies have common structure yet unclear ancestry. Individual peptidases are termed species. For example, clan PA peptidases contain eight families each bearing a classical two β-barrel architecture observed in chymotrypsin with either cysteine or serine acting as catalytic nucleophile (5). The blood clotting peptidase thrombin (6) is one species in the large S1A family of clan PA (7). Classification provides context to decipher peptidase function in the natural world.

Different classes of peptidases associate with vital biological pathways. Metallopeptidases dictate structural composition surrounding cells in multicellular organisms (8). Cysteine peptidases mediate regulated cell death (9). Serine peptidases enable our immune response (10). Threonine peptidases function as the central conduit for protein recycling in many organisms (11). Aspartic peptidases broker viral infection (12). However, each of these examples is a small subset within each catalytic type. To clarify peptidase diversity, we apply the MEROPS classification on a comprehensive set of genomes to visualize, compare, and contrast the proteolytic underpinnings of the biosphere. From this analysis, a map of peptidase distribution is drawn to provide a framework for decoding function, architecture, and evolution of proteolysis in vivo. Remarkably, a simple pattern is shown to underlie limited proteolysis in all forms of life.

EXPERIMENTAL PROCEDURES

Annotated protein sequences from whole genome sequence data were obtained from public ftp servers of the NCBI, DOE JGI, and select genome project homepages on the worldwide web. ClustalW was used to prepare sequence alignments from peptidase domains obtained from the MEROPS data base (3, 13). The HMMER package by Eddy (14) was used to build hidden Markov models (HMM) from the ClustalW prepared alignments and used to search protein sequence data using default cutoff values. HMMs were constructed for 154 peptidase families. Families were excluded from the census if they contained only a few sequences, typically less than 20, or if from viral origin. All of the data in the MEROPS data base were queried with each peptidase HMM and returned greater than 80% data base coverage indicating utility of the models (supplemental Fig. S1). An iterated search was applied on 128 genomes from all three kingdoms of life. The complete list of genomes examined, their taxonomy, and peptidase estimates are provided in supplemental Table S1. The HMMs applied capture peptidase diversity in the biosphere, yet the numbers of peptidases noted in each organism are only an estimate. Results were parsed with BioPerl (15). Cluster 3.0 and Treeview software was used to cluster census data after conversion into a binary matrix using the complete linkage algorithm and visualize the data as shown in Fig. 2 (16).

RESULTS

HMMs are useful indicators of peptidase content in a genome sequence (17, 18). As many peptidase families lack known fold, we rely upon MEROPS annotation for each domain. HMMs were constructed for 154 peptidase families using the HMMER software package to probe 128 eukaryotic,
Evolution of Peptidase Diversity

A landscape of genome-wide peptidase diversity is visualized from the census that enables a framework for understanding limited proteolysis in vivo. Serine peptidases are the most abundant class in the dataset followed by metalloproteases. Typically 100 peptidases are found in genomes of bacteria. In contrast, archaeal genomes present half that many. Explosive growth in multiple peptidase families is evident in eukaryotic organisms and the difference in degradome composition between eubacteria and eukarya is striking. Considerable work remains to detail how changes in polypeptide sequence associate with the functional gains in each of these families and further refine accuracy of the census. To derive a concise picture of peptidase diversity, we reduced dimensionality of the data.

Peptidase census results were converted into a binary matrix obviating the need for weighting functions and other advanced statistical measures. Clustering peptidase families from the binary dataset by the organisms that contain them greatly clarifies peptidase diversity (Fig. 2). The clustering algorithm, commonly applied to microarray data, segregates organisms into the three kingdoms of life and trends with accepted classification yet is entirely based upon degradome content (20–22). Despite presenting few peptidases in their genome, the archaea group is distinctively apart from the eubacteria. Inclusion of multiple genomes from parasitic organisms within the dataset complicates phylogenetic inference from the dataset due to significant gene loss (23, 24). Nevertheless, a clear pattern of peptidase distribution emerges accounting for all of the HMMs examined and nearly 90% of the families in the MEROPS data base of which 10% are viral peptidases not examined in this study.

A nearly ubiquitous core of 16 peptidase families encompasses proteolytic requirements of cellular life in a modern environment (Table 1). Given extensive distribution it is unlikely gene transfer between organisms explains this observation. Highly conserved peptidase families play central metabolic and protein processing roles such folate and glutathione metabolism, de novo purine biosynthesis, and signal peptide processing (25–28). Importance of intracellular protein homeostasis is evident by the presence of nine families with known digestive function (29). Rhomboid and FtsH peptidases cleave substrates on or in a phospholipid bilayer and are highly conserved (30, 31). The prolyl aminopeptidase family is associated with a variety of other chemistries outside of proteolysis including epoxide hydrolase and monoglyceride lipase activities (32).
autophagins are also critical for autophagy, a cellular process emerged through invention of six peptidase families (34). The gains in intra- and extracellular peptidase activity bring about four major cellular advances in eukaryotes (Table 1). First, ubiquitin- and SUMO-mediated protein turnover due to gene loss and/or horizontal gene transfer and limitations of the detection method. The nearly ubiquitous core of peptidase families can be identified in the genomes of all forms of cellular life. The known functions of these peptidases appear to encompass all the requirements for complex proteolysis capable of digestion and protein processing. Several organisms have lost one or more of these peptidase families. Of eukaryotic inventions only a handful are specific to higher metazoa.

Table 1
The nearly ubiquitous core of peptidase families

| Family | Clan | Common name | Biological function |
|--------|------|-------------|--------------------|
| C26    | PC   | γ-Glutamyl hydrolase | Folate metabolism |
| C44    | PB   | Amidophosphoribosyltransferase | Adenosine biosynthesis |
| M1     | MA   | Aminopeptidase N | Protein turnover |
| M16B   | ME   | Pitrilysin | Signal peptide processing |
| M20A   | MH   | Glutamate carboxypeptidase | Protein turnover |
| M22    | MK   | Hsp70/73/Eukaryote homolog | Chaperone |
| M24A   | MG   | Methionyl aminopeptidase | Protein turnover |
| M24B   | MG   | Aminopeptidase P | Protein turnover |
| M41    | MA   | FtsH peptidase | Membrane protein processing |
| S14    | SK   | Clp peptidase | Protein turnover |
| S16    | SJ   | LonA peptidase | Protein turnover |
| S1B    | PA   | HtrA peptidase | Protein turnover |
| S33    | SC   | Prolyl aminopeptidase | Turnover and epoxide hydrolase |
| S54    | ST   | Rhomboid | Membrane protein processing |
| S8A    | SC   | Subtilisin | Protein turnover |
| T3     | PB   | γ-Glutamyltransferase | Glutathione metabolism |

Evolution of Peptidase Diversity

FIGURE 2. Linkage analysis of peptidase families and the organisms containing them. Six clusters result in recapitulating known classifications of organisms and simplify the tremendous diversity of peptidases (color coded). A common proteolytic core of 16 peptidase families occurs in all kingdoms of life. Higher metazoa clearly distinguish themselves within the data and include all metazoa, including the choanoflagellate Monosiga brevicollis, except yeast, which are a separate cluster. Interspersed in the result are several peptidase families with mixed distribution due to gene loss and/or horizontal gene transfer and limitations of the detection method.

A number of peptidase families involved in extracellular biology evolved in eukaryal to broker key regulatory processes and set the stage for further expansion (Table 2). Invention of novel selectivity mechanisms fostered gains in the complexity of extracellular biology. Selectivity advances include the removal of a single basic amino acid from the C terminus of a substrate (M14B family) (40), cleavage of substrates following a penultimate proline residue (S28 family) (41), and steric hindrance to a dipeptidyl peptidase-selective active site by an associated protein domain (S9B family) (42). The unusual selectivity of these novel peptidases promoted expansion of cellular communication through processing small bioactive peptides. Of eukaryotic inventions only a handful are specific to higher metazoa.

Remarkably, only 10 peptidase families are restricted to the presence in non-yeast metazoa implying late evolutionary discovery (Table 3). Of these, the C14A caspases are notable for their central role in apoptosis (43). The C14 family is the clearest example of phylogenetic distinction in a peptidase clan. Peptidases in the C14A subfamily are found in eukaryotes, whereas C14B peptidases are not. Therefore an ancestral C14B peptidase is the root of all C14A peptidases. Two additions to the ubiquitin-dependent protein recycling machinery are present in higher organisms, the Cezanne deubiquitinylating peptidase (44, 45). Diversity of function is embodied by the M12A family of astacins, which act in processes ranging from hormonal control to extracellular matrix remodeling (46). Early association with a disintegrin domain propelled masps following a penultimate proline residue (S28 family) (41), and steric hindrance to a dipeptidyl peptidase-selective active site by an associated protein domain (S9B family) (42). The unusual selectivity of these novel peptidases promoted expansion of cellular communication through processing small bioactive peptides. Of eukaryotic inventions only a handful are specific to higher metazoa.

Remarkably, only 10 peptidase families are restricted to the presence in non-yeast metazoa implying late evolutionary discovery (Table 3). Of these, the C14A caspases are notable for their central role in apoptosis (43). The C14 family is the clearest example of phylogenetic distinction in a peptidase clan. Peptidases in the C14A subfamily are found in eukaryotes, whereas C14B peptidases are not. Therefore an ancestral C14B peptidase is the root of all C14A peptidases. Two additions to the ubiquitin-dependent protein recycling machinery are present in higher organisms, the Cezanne deubiquitinylating peptidases and CyLD protein (44, 45). Diversity of function is embodied by the M12A family of astacins, which act in processes ranging from hormonal control to extracellular matrix remodeling (46). Early association with a disintegrin domain propelled masps following a penultimate proline residue (S28 family) (41), and steric hindrance to a dipeptidyl peptidase-selective active site by an associated protein domain (S9B family) (42). The unusual selectivity of these novel peptidases promoted expansion of cellular communication through processing small bioactive peptides. Of eukaryotic inventions only a handful are specific to higher metazoa.

Remarkably, only 10 peptidase families are restricted to the presence in non-yeast metazoa implying late evolutionary discovery (Table 3). Of these, the C14A caspases are notable for their central role in apoptosis (43). The C14 family is the clearest example of phylogenetic distinction in a peptidase clan. Peptidases in the C14A subfamily are found in eukaryotes, whereas C14B peptidases are not. Therefore an ancestral C14B peptidase is the root of all C14A peptidases. Two additions to the ubiquitin-dependent protein recycling machinery are present in higher organisms, the Cezanne deubiquitinylating peptidases and CyLD protein (44, 45). Diversity of function is embodied by the M12A family of astacins, which act in processes ranging from hormonal control to extracellular matrix remodeling (46). Early association with a disintegrin domain propelled masps following a penultimate proline residue (S28 family) (41), and steric hindrance to a dipeptidyl peptidase-selective active site by an associated protein domain (S9B family) (42). The unusual selectivity of these novel peptidases promoted expansion of cellular communication through processing small bioactive peptides. Of eukaryotic inventions only a handful are specific to higher metazoa.
reasonable that the majority of peptidase families arise from convergent evolution. It is likely that the same ancestral peptidase family has independently evolved the ability to process different peptide substrates, as is the case with the cysteine proteases. However, this does not mean that these peptidases have originated from a common ancestor. Indeed, the vast majority of peptidase families are unique to yeasts. Together, these observations suggest that the conserved degradome among eukaryotes required few gains and many losses.

Peptidase families with mixed distribution are spread throughout the clustered data. Dispersal of these 32 peptidase families result from the diversity of both peptidases and genomes examined, gene transfer between organisms, and misidentification. Both free-living and parasitic lifestyles are also encompassed by the dataset and further complicate interpretation of the results. Of mixed peptidase families, the tricorn and M12B peptidase families have diverse functions and undergone rapid expansion to enable complex multicellularity.

Although most peptidase families can be grouped through organism-level description, several cannot due to sporadic distribution, poor identification by the HMMs applied, and rare occurrence. In certain instances, the MEROPS classification may require adjustment. For example, several larger peptidase clans with many families are likely distantly related. The common ancestor of these families lived over a billion years ago when organism-level description, poor identification by the HMMs applied, and rare occurrence. In certain instances, the MEROPS classification may require adjustment. For example, several larger peptidase clans with many families are likely distantly related. The common ancestor of these families lived over a billion years ago. However, phylogenetic reconstruction is challenging and demands both additional sequence information and close attention to reconcile convergent and divergent evolution.

Identification of a near ubiquitous core of peptidase families provides a working definition of a basal degradome and a starting point to appreciate evolutionary expansion of proteolytic
Evolution of Peptidase Diversity

networks. Near ubiquitous peptidase families do not increase the copy number per genome over time and do not expand into different functional niches. Non-expanding, highly conserved peptidase families broker intracellular processes and absence in certain organisms suggests they can be replaced. Therefore, although the core set of peptidases fulfill vital roles they are neither malleable to the forces of molecular evolution nor location independent. Such observations are all the more remarkable given exposure to billions of years of natural selection. Degradome expansion demanded novel mechanisms of limited proteolysis particularly in extracellular environments. Eukaryotes met this requirement through a variety of protein innovations yet a surprising paucity of these novel families is unique to higher metazoas.

Limited peptidase inventions in higher organisms suggest network organization, refining properties of select peptidases and enhanced spatial and temporal regulation were sufficient to generate the wide diversity of proteolysis in life. Trypsin, the S1A family, are the most abundant member in the human degradome, such as those of all non-yeast metazoa and many lower eukaryotes. Emergence of the S1A peptidase family pre-dates major speciation events in multicellular eukaryotes, yet the origin of this family remains unclear as with several other peptidase families due to mixed distribution. Visualization of degradome diversity provides a roadmap to develop a timeline for the evolution of proteolysis at the systems level.

REFERENCES

1. Lopez-Otin, C., and Overall, C. M. (2002) Nat. Rev. 3, 509–519
2. Puente, X. S., Sanchez, L. M., Overall, C. M., and Lopez-Otin, C. (2003) Nat. Rev. Genet. 4, 544–558
3. Rawlings, N. D., Morton, F. R., Kok, C. Y., Kong, J., and Barrett, A. J. (2008) Nucleic Acids Res. 36, D320–D325
4. Barrett, A. J., and Rawlings, N. D. (2007) Biol. Chem. 388, 1151–1157
5. Page, M. I., and Di Cera, E. (2008) Cell Mol. Life Sci. 65, 1220–1236
6. Di Cera, E. (2008) Mol. Aspects Med. 29, 203–254
7. Krem, M. M., and Di Cera, E. (2001) EMBO J. 20, 3036–3045
8. Mott, J. D., and Werb, Z. (2004) Curr. Opin. Cell Biol. 845–859
9. Kolaenko, A., Chable-Bessia, C., Cantarella, G., Israel, A., Wallach, D., and Courtois, G. (2003) Nature 424, 801–805
10. Evans, P. C., Smith, T. M., Lai, M. G., Burke, D. F., Heyn, K., Kreike, M. M., Beyaert, R., Blundell, T. L., and Kilshaw, P. J. (2003) J. Biol. Chem. 278, 23180–23186
11. Stocker, W., Grams, F., Baumann, U., Reiner, P., Gomis-Ruth, F. X., McKay, D. B., and Bode, W. (1995) Protein Sci. 4, 823–840
12. Black, R. A., Rauch, C. T., Zupan, C., Kossiakoff, A. A., Slack, J. L., Wolfson, H. O., Castner, B. J., Stocking, K. L., Reddy, P., R. Srinivasan, S., Nelson, N., Boiani, N., Schoole, K. A., Gerhart, M., Davis, R. F., Fitzenrider, J. N., Johnson, R. S., Paxton, R. J., March, C. J., and Cerretti, D. P. (1997) Nature 385, 729–733
13. Boldt, H. B., Overgaard, M. T., Laursen, L. S., Weyer, K., Sotrup-Jensen, L., and Oxvig, C. (2001) Biochem. J. 358, 359–367
14. Feng, D. F., Cho, G., and Doolittle, R. F. (1997) Proc. Natl. Acad. Sci. U. S. A. 94, 13028–13033
15. Shen, Y., Joachimak, A., Rosner, M. R., and Tang, W. J. (2006) Nature 443, 870–874
16. Taylor, A. B., Smith, B. S., Kitada, S., Kojima, K., Miyaura, H., Otwinowski, Z., Ito, A., and Deisenhofer, J. (2001) Structure 9, 615–625
17. Johnson, K. A., Bhushan, S., Stahl, A. H., Hallberg, B. M., Frohn, A., Glaser, E., and Eneqvist, T. (2006) EMBO J. 25, 1977–1986
18. W. F., Zhou, Y., and Gottesman, S. (1999) J. Bacteriol. 181, 3681–3687
19. Glas, R., Bogoy, M., McMaster, J. S., Gaczynska, M., and Ploegh, H. L. (1998) Nature 392, 618–622
20. Devos, D. and Valencia, A. (2001) Trends Genet. 17, 429–431
21. Puente, X. S., Sanchez, L. M., Gutierrez-Fernandez, A., Velasco, G., and Lopez-Otin, C. (2005) Biochem. Soc. Trans. 33, 331–334
22. Puente, X. S., Gutierrez-Fernandez, A., Ordonez, G. R., Hillier, L. W., and Lopez-Otin, C. (2005) Genomics 86, 638–647
23. Lopez-Otin, C., and Bond, J. S. (July 23, 2008) J. Biol. Chem., DOI 10.1074/jbc.R800035200