The C-terminal Sequence Encodes Function in Serine Proteases*

(Received for publication, June 25, 1999, and in revised form, July 21, 1999)
Maxwell M. Krem, Thierry Rose, and Enrico Di Cera‡

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

Serine proteases of the chymotrypsin family have maintained a common fold over an evolutionary span of more than one billion years. Notwithstanding modest changes in sequence, this class of enzymes has developed a wide variety of substrate specificities and important biological functions. Remarkably, the C-terminal portion of the sequence in the protease domain accounts fully for this functional diversity. This portion is often encoded by a single exon and contains most of the residues forming the contact surface in the active site for the P1–P3 residues of the substrate, as well as domains responsible for the modulation of catalytic activity. The evolution of serine proteases was therefore driven by optimization of contacts made with the unprimed sub-sites of the substrate and targeted a relatively short portion of the sequence toward the C-terminal end. The dominant role of the C-terminal sequence should facilitate the identification of function in newly discovered genes belonging to this class of enzymes.

Serine proteases of the chymotrypsin family feature a gamut of important physiological functions, ranging from digestive and degradative processes to blood clotting, cellular and humoral immunity, fibrinolysis, fertilization, and embryonic development (1). How this variety of functions emerged during evolution from a highly conserved three-dimensional fold (2) is incompletely understood. Enzymes involved in digestive and degradative processes, such as trypsin and chymotrypsin, are found from bacteria to human and are composed of a protease domain containing all epitopes for ligand recognition (3, 4). On the other hand, enzymes involved in more specialized functions such as blood coagulation, humoral immunity, and fibrinolysis are present almost exclusively in vertebrates and carry additional modules that confer more stringent specificity and localize the proteolytic function in space (5–7). A large repertoire of functions can be created by linking the protease domain to specialized functional modules, and most likely this strategy enabled the evolution of highly selective and specialized enzymes from primitive digestive proteases (8).

The mosaic organization of the primary structure of many serine proteases has fostered the notion that the protease domain alone would not bear reliable signatures of function. Previous sequence analyses and functional characterizations of serine proteases have relied heavily on the architecture of non-protease domains (6, 8, 9). Evolutionary trees for the hepatocyte growth factor (HGF),‡ HGF activator, kallikreins, mast cell protease, and complement lineages have been constructed in this manner (7, 10–13). However, Doolittle and Feng (5) were able to construct a reasonable evolutionary tree for clotting and fibrinolytic proteases from analysis of the protease domain only. This domain is the only structural component present in all members of the serine protease family. We therefore posed the question as to whether this domain would suffice to gain a predictive understanding of the function of the protease.

MATERIALS AND METHODS

Data Base—Amino acid and nucleotide sequences of 251 chymotrypsin-like serine proteases and homologous proteins were culled from GenBank. Of the 251 original sequences, 89 were selected for the construction of evolutionary trees to minimize duplication of nearly identical sequences of the same proteins. Serine protease domains were aligned from residues 16 to 245 of the chymotrypsin catalytic domain. Alignments were performed with CLUSTAL_W (14). Protein distance matrices were calculated using PROTDIST from the PHYLIP package (15). Unrooted evolutionary trees were constructed using the Fitch-Margoliash method (16) with the PHYLIP program FITCH.

Analysis of Crystal Structures—The contribution of 50-residue-long stretches to the surface contacting the substrate in the active site was calculated for thrombin (1PPB) (17), chymotrypsin (4CHA) (18), trypsin (1TLD) (19), tissue plasminogen activator (1RTF) (20), elastase (3EST) (21), and kallikrein A (2PKA) (22). The surface was calculated using the structure of thrombin inhibited with H-D-Phe-Pro-Arg-CH₂Cl (PPACK) at the active site as reference. The three residues of PPACK occupy the S3, S2, and S1 pockets, respectively. The surface area contacting the substrate was defined as being contributed by any residue within 4.5 Å of PPACK. Notably, the residues defining this area in thrombin were also found to define the same area in the other five proteases analyzed and enabled a consistent comparison of all structures. The area contacting the substrate was calculated for 50-residue-long stretches along the sequence, using the chymotrypsin(ogen) numbering and excluding all insertions relative to chymotrypsin. These insertions contributed no more than 5% to the S1–S3 pockets.

RESULTS AND DISCUSSION

Analysis of nearly 90 non-redundant sequences of the protease domain, corresponding to residues 16–245 in the bovine chymotrypsin(ogen) numbering, was carried out using the Fitch-Margoliash algorithm (16) and gives the tree depicted in Fig. 1. The separation of functions is immediately recognizable in the tree, with the digestive and fibrinolytic proteases segregated from the complement and blood-clotting enzymes. The separation and the results derived from it are entirely independent of how the tree is rooted. We therefore chose the oldest eucaryotic organism as origin. With this choice, there is early divergence of several of the arthropod enzymes and of enzymes of cell-mediated immunity from the same ancestral lineage. This concurs with previous evidence of a unique origin of proteases of cell-mediated immunity (23). There is also early divergence of HGF-related enzymes and kallikreins. In contrast, there is late divergence of HGF activator-related enzymes and

‡ The abbreviations used are: HGF, hepatocyte growth factor; PPACK, H-D-Phe-Pro-Arg-CH₂Cl.
C-terminal Sequence Encodes Function in Serine Proteases

**Fig. 1. Phylogenetic tree of serine proteases constructed from residues 16-245 in the protease domain.** The tree is rationally, though arbitrarily, rooted so that enzymes from sea-based arthropods are the root group. However, the choice of the root group does not change the topographic structure of the tree. The tree depicts phylogenetic relationships among protease domains. Although temporal evolution of protease domains is a likely mechanism behind the tree structure, the tree should not be looked upon as an attempt to specify the precise sequence or dates of emergence of particular groups of proteases. Furthermore, the tree does not imply that the physiologic functions contributing to group names evolved concurrently with the proteases that participate in those functions (e.g. the tree does not suggest that the cell-mediated immune system developed concurrently with arthropod degradative enzymes but does suggest that the protease domains of the enzymes listed in the group are most closely related to the protease domains of arthropod degradative enzymes). Groups are named according to common functions shared by either a majority or plurality of the entities in a group. Proteases that differ in function from their corresponding group name are marked with annotations (see below). Proteases and proteins are from vertebrates unless a species is indicated. Most proteases or proteins are from humans or other mammals. Human proteases and proteins are used wherever possible. a, factor D is part of the alternative complement pathway and thus the humoral immune system. b, by aspartate and ancorcle cleave fibrinopeptides and thus act in fibrinolytic roles. c, plasma kallikrein and factor XI belong to the coagulation cascade. d, trypsinase is a mast cell protease and participates in the inflammatory response. e, plasmin is a fibrinolytic enzyme. f, apolipoprotein A is not a protease but contains a serine protease domain and participates in atherosclerotic degradation of blood vessels. g, factor I is part of the alternative complement pathway and thus the humoral immune system. h, caldrcin is closely related to elastase but acts to reduce serum calcium levels in a nonproteolytic way. i, factor XII is a clotting factor that also has activity similar to hepatocyte growth factor activator (HGFA). j, serine protease 24D is a digestive enzyme; its specific function is unknown. k, C. elegans S1 peptidase of unknown function (GenBank accession number 1572759). l, C. elegans MAP-1, factor Clr, CASP, C. elegans S1 peptidase of unknown function (GenBank accession number 1326386). m, haptoglobin is not a protease but participates in humoral immunity as an acute phase protein. n, trypsin-like enzyme of Streptococcus griseus; classified with S2a peptidase family. The grouping of this protease suggests that this enzyme has a eukaryotic origin (2). o, trypsin-like enzyme of the bread mold Fusarium oxysporum. p, C. elegans S1 peptidase of unknown function (GenBank accession number 2088818). q, hypodermins A and B interfere with the cell-mediated immunity of the host organism (cattle). r, allergens 1 and 2 interfere with the cell-mediated immunity of the host (cattle). NK, natural killer; PSA, prostate-specific antigen; SCCF, stratum corneum chymotryptic enzyme; t-PA, tissue plasminogen activator; u-PA, urokinase plasminogen activator; e-PA, salivary plasminogen activator; MASP, mannose-associated serine protease; CASP, calcium-activated serine protease.

Blood-clotting enzymes. The latter are always interspersed with complement enzymes, and both groups diverge in close relation to late diverging arthropod enzymes, especially dorsal-ventral determinants from Drosophila melanogaster (24-26). Additionally, fibrinolytic enzymes show late divergence, in association with degradative proteases that include digestive proteases, and branch from the tree at various evolutionary times. The most striking trends appear to be the previously documented shared ancestry between complement and blood-clotting enzymes (27), the close relationship between these and arthropod enzymes, and the close association between arthropod enzymes and the proteases of cell-mediated immunity.

Next, we asked whether the entire sequence of the protease domain was also necessary, other than sufficient, to provide the documented separation of function. Trees such as the one in Fig. 1 were constructed from 50-residue-long fragments of the sequence in the protease domain. The distance from the ideal tree approximated from that constructed from the complete sequence was computed in each case (Fig. 2). The minimum limit of sequence necessary to produce an acceptable tree that maps serine proteases into functional groups as in Fig. 1 varied according to the position along the sequence. The tree constructed from the C-terminal segment yielded coherent separation of functional groups, corresponding closely to the tree constructed from the entire protease domain. The two trees shared general relationships and branch points between major functional groups. Farther than 75 residues from the C terminus, using a 50-residue fragment to construct trees, resulted in the breakup of major functional groups, with the general exception of proteases of cell-mediated immunity. That such trees represent a less satisfactory result is indicated by increased deviations between observed distances in the tree and the expected deviations based on the protein distance matrix input (Fig. 2). Remarkably, the tree generated from the last 50 residues alone deviated from the complete tree less than a tree constructed from the first 175 residues of the protease domain. This makes the last 50 residues necessary, other than sufficient, to reliably assign function.

The dominant role of the C-terminal sequence in specifying the function of a serine protease has a direct structural link, insofar as this portion of the protease domain makes up most of the surface of the S1–S3 specificity sites that interact with the substrate (Fig. 2). The contribution to substrate recognition by 50-residue-long stretches of the sequence was calculated for...
C-terminal Sequence Encodes Function in Serine Proteases

The C-terminal sequence of serine proteases contains most of the structural determinants important for direct substrate recognition at the S1–S3 sites. It should be pointed out that residues not contained within the C-terminal domain also shape some of these sites. For example, the S2 and S3 sites of clotting and fibrinolytic proteases are partially shaped by insertions within the 60 loop and residues 99 and 174 (33). However, the overall contribution of the last 50 residues in the sequence to the surface area contacting the bound substrate far exceeds that of any other segment of the structure. In addition, the C-terminal sequence contains domains such as the 220 loop that shapes portions of the water channel embedding the primary specificity pocket and the conserved C-terminal helix that stabilizes the fold of the C-terminal domain of the protease and the environment of the catalytic Asp192.

Fig. 2. Sum of squares of deviations versus position along the chymotrypsin(ogen) sequence for trees calculated from segments 50 residues in length (© and interpolating spline; each point maps the center of the 50-residue-long sequence). Deviations were generated by comparing the interprotein distances from the tree program and those specified in the distance matrix, which serves as the input file for the tree program. Deviations become markedly lower as the C terminus is approached, indicating that the C-terminal portion of the sequence specifies a phylogenetic tree that corresponds more closely to the actual sequence differences (see Fig. 1). Also shown is the contribution from segments of 50 residues to the solvent-accessible surface area (Å²) contacting the substrate into the active site, derived from the crystal structures of chymotrypsin, elastase, kallikrein, thrombin, tissue plasminogen activator, and trypsin. Each point represents the average computed from the six structures, and the dispersion around the point is the standard deviation of the values. The area increases significantly as the C-terminal end is approached, mirroring the opposite trend observed in the deviation from the tree constructed from alignment of the complete sequence. The horizontal bar at the top of the figure displays the boundaries of the last exon, which contains the critical residues in the C-terminal sequence.

Fig. 3. Contribution of the C-terminal residues 189–245 (white) to the solvent-accessible surface of chymotrypsin, elastase, kallikrein, thrombin, tissue plasminogen activator (tPA) and trypsin, projected on the surface of the enzymes (blue). The active site inhibitor H-n-Phe-Pro-Arg-CH₂Cl of thrombin is displayed as a stick (green). The catalytic His57 is in red. All molecules are displayed in the same orientation relative to the position of the catalytic triad. The C-terminal residues define most of the accessible surface in the specificity sites S1–S3 of the enzyme. In addition, these residues define the 220 loop that shapes portions of the water channel embedding the primary specificity pocket and the conserved C-terminal helix that stabilizes the fold of the C-terminal domain of the protease and the environment of the catalytic Asp192.
ing Asp^{102} via the critical residue Trp^{237}, thereby affording control and modulation of the catalytic activity of the enzyme.

The foregoing arguments lend support to the hypothesis that the C-terminal segment of the protease domain of serine proteases exerts an overwhelmingly disproportionate influence, compared with other sequence segments, on the recognition of substrate and the control of catalytic activity within the serine protease family. Consequently, the P1–P3 residues of the substrate are expected to dictate binding in a dominant manner, with the primed residues adding to the specificity of recognition to a lesser extent. Mutagenesis data from a number of systems support this conclusion, because mutations in the unprimed sites of the substrate are typically far more deleterious to binding than those involving the primed sites (38).

Consistent with its dominant functional role in substrate recognition, the C-terminal segment of the protease domain of serine proteases seems to have been involved in all of the evolutionary decisions pertaining to this class of enzymes. The exon structure of serine proteases may underlie this phenomenon (39). Chymotrypsin, which possesses the largest number of exons (seven) in the catalytic domain (40), is encoded from the C terminus by one exon that corresponds closely to the boundaries of the 50-residue sequence that produces a coherent functional tree (Fig. 2). Several proteases split this terminal chymotrypsin exon (23, 39), but the location of the intron/exon boundary near Gly^{193} varies little. Therefore, the C-terminal exon of serine proteases coincides with what appears to be the function-determining domain of the protein sequence.

The special role of the C-terminal sequence in serine proteases may facilitate the identification of function in newly discovered genes, even when information is available only on limited stretches of the sequence. The consensus repeat GDGSGG around the active site Ser^{195} is usually diagnostic of a serine protease. The results discussed here demonstrate that this sequence and the 30–40 residues immediately downstream from it are both necessary and sufficient to reliably assign function. We therefore postulate possible functions for three recently discovered serine proteases from the nematode worm Caenorhabditis elegans (41). The protease with GenBank TMD accession number 2088818 diverges just prior to nucleot, stubble, and snake in the tree determined from the 16–245 sequence, and with complement and clotting proteases in the tree derived from the C-terminal 191–245 sequence. We predict that it has either a developmental or a primordial humoral defense role. The protease with accession number 1572759 diverges with complement and clotting proteases in the 16–245 tree and with degradative and cell-mediated cytotoxic proteases in the 191–245 tree. We therefore propose that it is a degradative, perhaps cytotoxic enzyme involved in humoral immunity, as nematodes are unlikely to have cell-mediated immune responses. The protease with accession number 1326386 diverges with complement and clotting in both trees and should be involved in humoral immunity.