# EVOLUTION IN THE STRUCTURE AND FUNCTION OF CARBOXYL PROTEASES

Jordan TANG

*Laboratory of Protein Studies, Oklahoma Medical Research Foundation, and Department of Biochemistry and Molecular Biology, University of Oklahoma Health Sciences Center, Oklahoma City, Oklahoma 73104*

(Received April 2, 1979)

## Summary

A model for the structure and function of extracellular carboxyl (acid) proteases can be established from three amino acid sequences and four crystal structures of these enzymes. The carboxyl proteases from gastric and fungal origins are very homologous in both primary and tertiary structures. The molecules consist of about 320 residues organized with a secondary structure which is primarily comprised of $\beta$-strands and very similar tertiary structures. An apparent binding cleft, which can accommodate a substrate with about eight amino acid residues, contains near its midpoint the active center residues Asp-215, Asp-32, and Ser-35. These three residues are hydrogen bonded to each other.

An intracellular carboxyl protease, cathepsin D, is very homologous to the extracellular enzymes in N-terminal amino acid sequence and primary structure location of active center residues. The tertiary structure of cathepsin D is probably similar, as well. However, cathepsin D contains a unique hydrophobic "tail" made up of about 100 residues added on the C-terminal side. Cathepsin D precursor is over 100,000 daltons in molecular weights, as contrasted to the gastric carboxyl protease zymogens, which are about 40,000 daltons.

Carboxyl proteases contain two lobes symmetrical in peptide chain conformations. Each of the lobes also consists of two homologous structural units. These structural characteristics suggest that the original gene was coded for only about eighty amino acid residues and that gene duplication and fusion has taken place twice to produce a single chain carboxyl protease with four basic structural units in two symmetrical lobes. The formation of the zymogens and the cathepsin D "tail" must have resulted from various gene fusions. Partial sequence comparisons also suggest that cathepsin D may be an evolutionary ancestral chain for gastric carboxyl proteases.

## Introduction

In the past few years a great deal of progress has been made in the understanding of structure and function of carboxyl proteases. Several reviews have appeared. They have covered either a special carboxyl protease-zymogen model (e.g., pepsin-pepsinogen)[1,2], or accounted for only the latest findings[3]. In this article, however, a broad view will be taken to encompass the structural evolution in carboxyl proteases and its relationship to the enzymic functions.

The term "carboxyl protease" describes a group of proteolytic enzymes, such as pepsin, which share an apparently unique catalytic apparatus and mechanism. For two main reasons the name "carboxyl protease" has become a more favorable term in recent years over the older name "acid protease"[1]. First, it is now apparent that the active sites of these enzymes contain two carboxyl groups; the name

TABLE I

Carboxyl (acid) proteases

| Source | Enzyme | Zymogen | Review or Original References |
|---|---|---|---|
| Stomachs | Pepsin | + | 4 |
| | Gastricsin | + | 4 |
| | Chymosin (rennin) | + | 4 |
| Lysosomes | Cathepsin D | + | 35 |
| | Cathepsin E | + | 35 |
| Plasma and urine | Uropepsin | + | 35 |
| Seminal plasma | Acid protease | + | 35 |
| Kidney | Renin | + | 13 |
| Microbial | | | 9 |
| *Penicillium janthinellum* | Penicillopepsin | - | |
| *Rhizopus chinensis* | Rhizopuspepsin | - | |
| *Endothia parasitica* | Endothiapepsin | - | |
| *Aspergillus saitoi*<br>*Aspergillus oryzae*<br>*Aspergillus niger* | Aspergillopepsin | - | |
| *Mucor pusillus*<br>*Mucor miehei* | Mucorchymosin<br>(Mucor renin) | - | 10 |
| Protozoan | | | |
| Tetrahemina | | | 11 |
| *Plasmodium berghei* | Acid protease | | 55 |
| *Plasmodium falciparum* | | | 56 |
| *Plasmodium knowlesi* | | | 56 |
| Plant | | | |
| *Sorghum vulgare* | Sorghum acid protease | | 8, 58 |

is thus consistent with other groups of proteases, i.e., sulfhydryl-, serine-, and metalloproteases. Second, some carboxyl proteases, notably renin, have pH optima near neutrality. The name "acid protease" seems inappropriate in such a case. Some investigators also have used the name "aspartyl protease" to describe this group of enzymes.

Carboxyl proteases are widely distributed in living organisms. Some of the examples are given in Table 1. The best known group is probably the gastric digestive proteases present in the stomach of all vertebrates. In high mammals only three structurally defined carboxyl proteases (pepsin, gastricsin, and chymosin) are found[4]. Even though the presence of other isozymes has been reported in man[5,6], these are likely to be different degradative products. All three gastric enzymes are secreted as zymogens. They are converted to active enzymes, as typified by the pepsinogen-pepsin conversion demonstrated by HERRIOTT[7] in the

1930's. A zymogen protease system very similar to the gastric enzyme is found in the seminal plasma of man.

Extracellular digestive carboxyl proteases are present in plants[8] and microorganisms[9]. Especially well studied are pepsin- and chymosin-like enzymes from fungi which are extensively used in the fermentation industry[10]. These microbial carboxyl proteases are secreted extracellularly. In contrast to the gastric enzymes no zymogen has been found for the extracellular microbial carboxyl proteases.

Intracellular carboxyl proteases are known to be located in the lysosomes and to be involved in the intracellular digestion of proteins. Like other lysosomal hydrolases, cathepsin D and E have optimal enzymic activity in a pH range of 4–5. The presence of these carboxyl proteases in higher living forms is well established. They may be broadly distributed in lysosomes of the lower forms of living organisms, since there is some evidence that tetrahemina lysosomes contain an "acid protease"[11]. A carboxyl protease, proteinase A, is known to be present in the vacuoles of yeast. This intracellular protease has been shown by HOLZER and co-workers to be involved in some important cellular regulatory functions[12].

Renin, another carboxyl protease, is of particular interest. This plasma enzyme is involved in a delicate regulation of blood pressure, and it also has an optimal pH for activity near neutrality. As will be discussed in the following sections, all the active-site directed reagents specific for carboxyl proteases inactivate renin, so it is considered a member of the carboxyl protease group. The enzyme is known to be made in the kidney and submaxillary gland, and large molecular weight enzyme forms, presumably precursors, have been identified[13].

All the enzymes mentioned above, and undoubtedly many others not covered or discovered, represent a group of enormous diversity in their functional locations, biological roles, enzymic properties, and evolutionary history. Yet it is increasingly clear that this group of enzymes shares common structural features, functional apparatus, and catalytic mechanisms. This has been established mainly from the studies of structure-function relationships of model enzymes, which are discussed in the following section.

## A Structure-Function Model for Extracellular Carboxyl Proteases

It is now known that the primary and three-dimensional structures of the gastric and microbial carboxyl proteases are very much alike. The primary structures have been revealed for three carboxyl proteases. Two enzymes are gastric in origin, pepsin[14] and chymosin[15], while a third, penicillopepsin[16], is secreted extracellularly by a fungus, *Penicillium janthinellum*. Figure 1 illustrates the alignment of three structures. Not only are the overall sizes similar, but the homology of amino acid sequences is apparent. Alignment of partial sequences from other gastric and microbial sources shows them to be homologous with three complete sequences[4].

The identification of catalytic-site residues was facilitated by the use of two affinity labeling reagents. Diazoacetyl-DL-norleucine methyl ester (DAN), developed in the laboratory of STEIN and MOORE[17], esterified a unique aspartyl residue[18], which was shown to be residue Asp-215 in pepsin[14]. Several similar diazo reagents have been tried on various carboxyl proteases. All reacted to the corresponding site at Asp-215[19]. A second active-site directed reagent, 1,2-epoxy-3-(p-nitrophenoxy)propane (EPNP), was used in our laboratory to specifically inactivate pepsin. This reagent esterifies a different aspartyl group[20,21], which is located at position 32 of the pepsin sequence[14]. EPNP inactivates all the gastric and microbial carboxyl proteases so far tested. The results of the use of these two reagents have been reviewed earlier[19,32].

The similarity in three-dimensional structures derived from X-ray crystallographic studies of four carboxyl proteases are even more striking. In addition to pepsin[23] and penicillopepsin[16], two fungal carboxyl proteases from *Rhizopus chimensis*[24] and *Endothia parasitica*[24] have all been solved at atomic resolution. For a direct comparison, the stereo pairs from pepsin and penicillopepsin are shown in Figure 2. As can be seen, the gastric and the microbial enzymes have very similar three-dimensional structures. This similarity is true also for the crystal structure of the two other microbial carboxyl proteases. Since the structures are all quite similar, the common features generated by the X-ray crystallographic studies are discussed

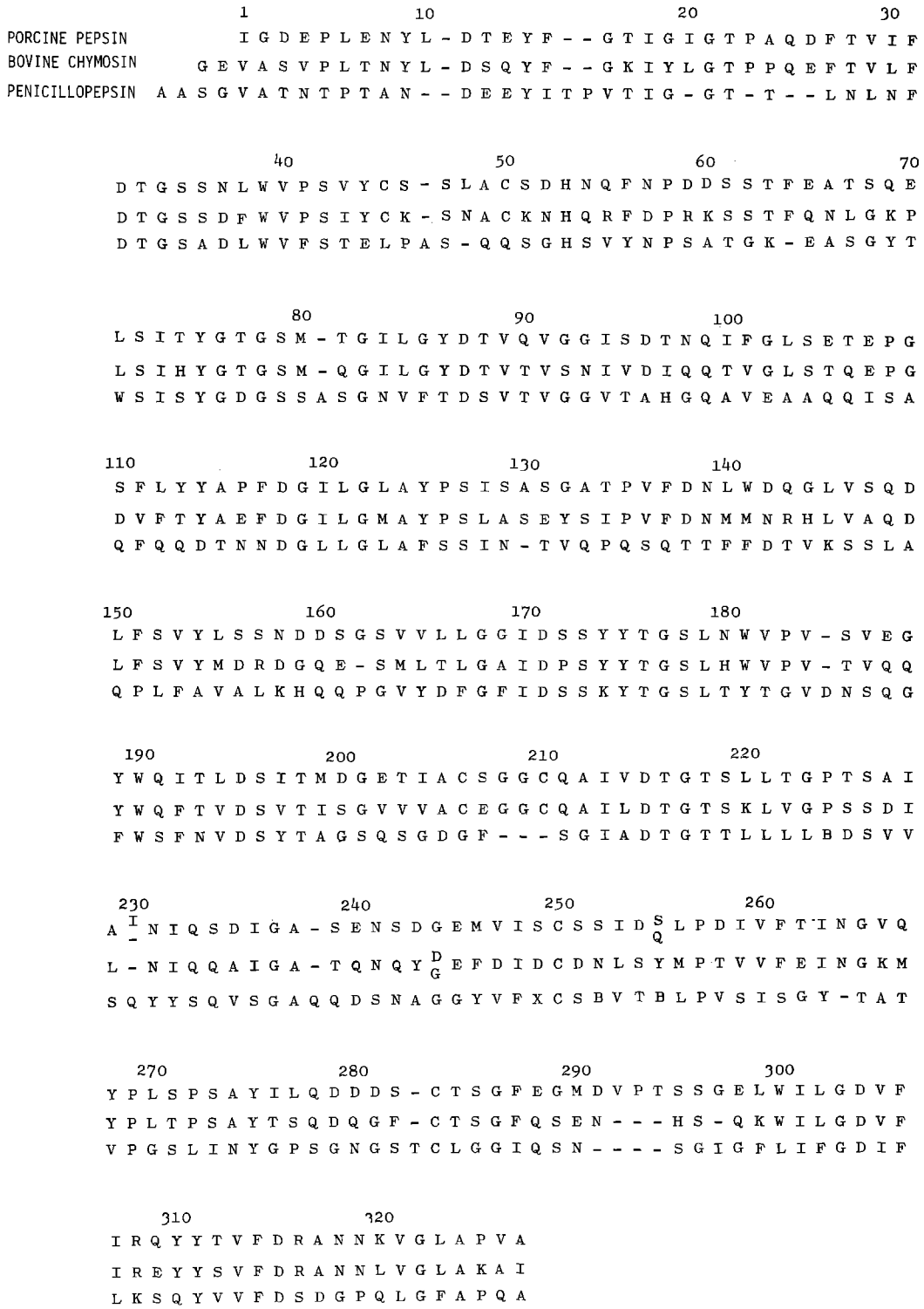


Fig. 1. The amino acid sequence of three carboxyl proteases, pepsin[14], chymosin[15], and penicillopepsin[16]. The numbering is based on pepsin sequence. The overall homology between the gastric and fungal enzymes is apparent. The similarity in sequences around active-center residues at positions 32, 35, 75, and 215 is particularly strong. The residues are in single-letter codes: A:Ala, B:Asx, C:Cys, D:Asp, E:Glu, F:Phe, G:Gly, H:His, I:Ile, K:Lys, L:Leu, M:Met, N:Asn, P:Pro, Q:Gln, R:Arg, S:Ser, T:Thr, V:Val, W:Trp, X:unknown, Y:Tyr, Z:Glx, –:gap. The disulfide linkages are between residues 45–50, 206–210, and 250–283.

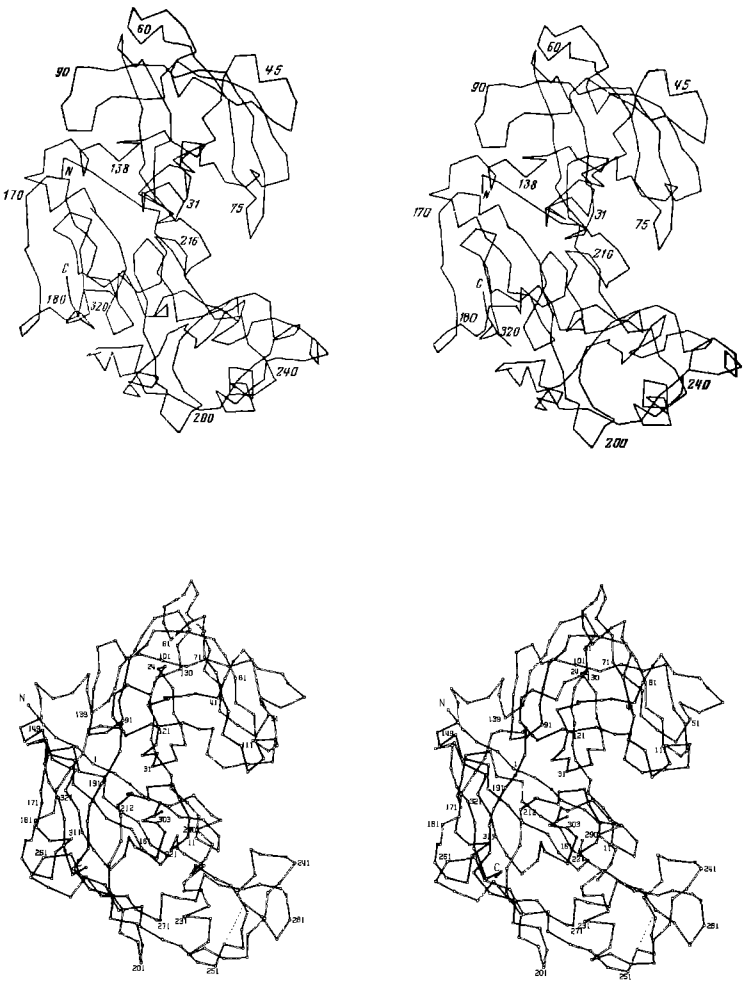


Fig. 2. Stereo drawing of the $\alpha$-carbon positions of two carboxyl proteases. The pepsin structure (top) is the work of ANDREEVA *et al.*[23]. The penicillopepsin structure was taken from the work of HSU *et al.*[16]. Strong similarity in overall shapes and polypeptide chain tracings are apparent.

together in the following. An apparent substrate binding cleft can easily be seen to run across the whole length of the molecule. The size of the binding cleft is sufficient to accommodate 7–8 amino acid residues of a peptide substrate. This feature is predicted from earlier specificity studies, both in protein[25,26] and synthetic peptide substrates[27], which indicated that the specificity of pepsin lies in two primary and six secondary residues equally distributed on both sides of the hydrolyzing bond. The active site Asp-32 and Asp-215 mentioned above are located at the center of the binding cleft. In the crystal structure of penicillopepsin the positions and orientation of these two $\beta$-carboxyl groups are sufficiently clear to indicate that they are probably hydrogen bonded to each other. In addition, Asp-32 is hydrogen bonded to Ser-35, which appears to be conserved in the amino acid sequence of all carboxyl proteases studied[4]. The crystallographic findings in the active site structure of the other three carboxyl proteases are consistent with that for penicillopepsin. As will be discussed later, these three active-site residues comprise an important, and probably the main, component of the catalytic apparatus. The polypeptide foldings in the three-dimensional structure consist primarily of $\beta$-structure. Only three segments of short helical structure are present.

The chemical and three-dimensional framework described above provides a basis for mechanistic considerations of the catalytic function of carboxyl proteases. But as in the case of

other catalytic mechanisms, detailed knowledge of the electronic events is far from complete. Nevertheless, taking the evidence mainly from the chemical studies of pepsin and the crystallographic studies of penicillopepsin, the possible catalytic roles of the active-site residues are rather limited. It has been known for some time from the results of kinetic studies that the catalytic carboxyl groups have pKa value of about 2 and 5, respectively[22]. The low pKa (2.8) carboxyl group was assigned to Asp-32[28]. This down shift in pKa can be readily explained from the hydrogen bonding between Asp-32 and Ser-35. The high pKa (5) carboxyl group was assigned to Asp-215[17,29]. Thus, it is reasonably assumed that in the catalytic processes, Asp-32 should be ionized and involves either a direct or indirect (with H-bonded $H_2O$) nucleophilic attack of the carbonyl carbon atom of the scissile peptide bond (Fig. 3)[28,30]. This role for Asp-32 as a general base catalyst is supported by recent experiments of ANTONOV *et al.*[31] using $O^{18}$ exchange techniques. Whether Asp-32 carboxylate acts as a nucleophile directly has not been firmly established, although in the esterification reaction by EPNP, a substrate-like inactivator, Asp-32 must have been involved in a direct nucleophillic attack. However, the binding of the acyl component of the substrate and of EPNP may or may not be identical. The role of Asp-215 as proton donor has been proposed to involve either the carbonyl oxygen[30] or the amide nitrogen[28] of the substrate (Fig. 3). Tyr-75 has also been proposed to be the proton donor to the amide nitrogen[30]. From a chemical viewpoint Asp-215 with a pK of 5 is a better proton donor than a phenolic group which normally has a pK above 10. Additionally, the polarization of carbonyl of the substrate does not require the full transfer of protons. For example, in the serine proteases the carbonyl oxygen of the substrates is hydrogen bonded to two amide hydrogens from the polypeptide backbone of the enzymes[32]. These discussions clearly accentuate a lack of detailed knowledge in the carboxyl protease catalytic mechanism and a need for further studies. Some of these uncertainties will be clarified when the orientation of the substrates in the crystals is better known.

In addition to DAN and EPNP, pepstatin is a universal carboxyl protease inhibitor. The potency in its activity ($K_i$ about $10^{-10}$ M for pepsin) has been attributed to the structure of this heptapeptide being a transition state analog of carboxyl protease catalysis[33]. A slow developing inhibition by pepstatin, which is apparently stereo specific with respect to the inhibitor structure, has been observed by RICH using synthetic pepstatin analogs[34]. The reason for this unusual kinetic behavior is yet unclear.

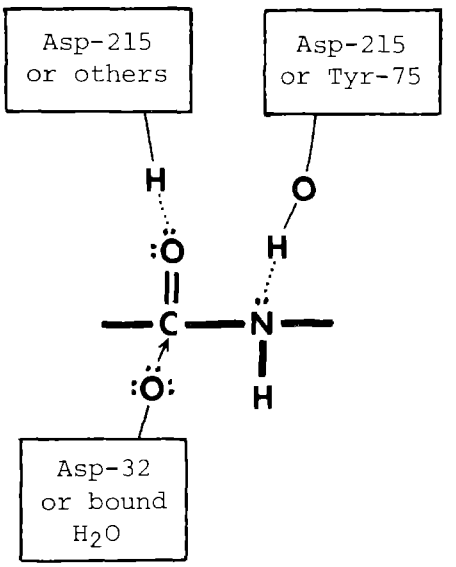


Fig. 3. The probable roles of active center residues of carboxyl proteases in the hydrolysis of a peptide substrate. For the explanations, see text.

The common features discussed above lead to a fairly clear conclusion that the structural features and the catalytic apparatus of extracellular digestive carboxyl proteases are very similar. These results have been obtained primarily from fungal and mammalian enzymes but most likely can be extended to the extracellular carboxyl proteases of other biological systems.

## Structure and Function of Cathepsin D – An Intracellular Carboxyl Protease

In contrast to the gastric and fungal carboxyl proteases, which digest protein extracellularly, cathepsin D is an intracellular enzyme. It is located in the lysosome and has an optimal pH for its function near that of intralysosomal pH – around 4. The role of cathepsin D as one of the major intracellular endopeptidases has long been recognized[35]. The level of this enzyme is greatly elevated when tissue resorptions

are taking place. It has also been implicated in the breakdown of tissues under the pathological conditions. More recently, cathepsin D has been shown to involve the normal turnover process of the intracellular proteins[36]. Therefore, cathepsin D seems ideally suited for a model of structure-function study of intracellular carboxyl proteases.

The carboxyl protease nature of cathepsin D is firmly established from its inhibition by specific inhibitors including DAN[37], EPNP[38], and pepstatin[39]. However, no significant structural study has been conducted until recently. In the following some recent results from our laboratory are described[40,41]. From porcine and bovine spleens a number of cathepsin D isozymes have been isolated in large enough quantities for structural comparisons. These isozymes, five from porcine[40] and two from bovine[41] spleens, differ primarily in their isoelectric points. The porcine isozymes have identical molecular weights of 50,000 daltons, while the bovine isozymes are 46,000 daltons. As illustrated in Figure 4, the isozymes are present either in single polypeptide chains or in two chains. The amino acid sequence at the $NH_2$-termini of the light and single chains are the same. This indicates that the two chain isozymes are derived from the single chain by limited proteolysis, which takes place near residue 100 in the original single chain isozyme. In bovine spleen the single chain cathepsin D comprises about 60% of the total, while in porcine spleen this percentage is less than 5. These two structural forms appear to exist *in vivo*. The two-chain species could not be produced *in vitro* through various proteolysis conditions, e.g., autolysis. In addition, the presence of protease inhibitors during the tissue homogenization did not significantly alter the ratios of single- to two-chain isozymes.

The $NH_2$-terminal sequence of porcine and bovine light chains are shown in Figure 5. A clear structural homology is seen in the comparison of these sequences to the $NH_2$-terminal sequences of extracellular carboxyl proteases. The evolutionary aspects of this comparison will be discussed in a separate section. From the structure-function viewpoint, it is important to note that the Asp-32 is most probably the EPNP reactive site in cathepsin D and that the active center structure in cathepsin D is very similar to that in pepsin and other extracellular carboxyl proteases. This assumption is supported by the DAN inhibition of cathepsin D. We have observed that one residue of DAN is incorporated into the heavy chain of cathepsin D, consistent with the location of Asp-215. A DAN reactive peptide was isolated by KEILOVA, who showed that the sequence was very similar to that around Asp-215 in pepsin[42].

The sequence similarity of cathepsin D to pepsin also predicted, with good justification, that the three-dimensional polypeptide chain windings in two enzymes and other carboxyl proteases would be very similar. (In comparing the sequence and tertiary structural homology of evolutionary related proteins, the similarity in three-dimensional structures are always more extensive than in the primary structure[43]). However, a major intriguing point is evident: cathepsin D is about 120 residues longer than other carboxyl proteases. This extra, or "tail", region is most likely at the C-terminal region of cathepsin D (Fig. 4). Since cathepsin D is highly homologous to the gastric proteases, an approximate amino acid composition of the "tail" region was estimated from the difference between cathepsin D and chymosin. The calculation showed that the "tail" region contains high numbers of basic and hydrophobic residues. We tentatively speculate that the "tail" may function as a membrane binding region of cathepsin D in the lysosomes. A similar hydrophobic tail is known for cytochrome $b_5$[44], a membrane bound protein.

An interesting isozyme of cathepsin D isolated from porcine spleen is the high molecular weight isozyme. This single chain cathepsin D is about 100,000 daltons but contains less than 10% of the specific activity of the 50,000 dalton species. Upon denaturation with urea, the high molecular weight cathepsin D cross reacted with antiserum raised against cathepsin D heavy chain and showed complete identity. We feel that this 100,000 species may be a precursor to the single chain 50,000 dalton isozyme.

From the above discussed findings, it appears that in cathepsin D the catalytic apparatus and the tertiary structure are very much the same as in the secreted extracellular carboxyl proteases. This is undoubtedly a consequence of divergent evolutionary processes (see below). The extra piece of "tail" structure may serve the regulat-

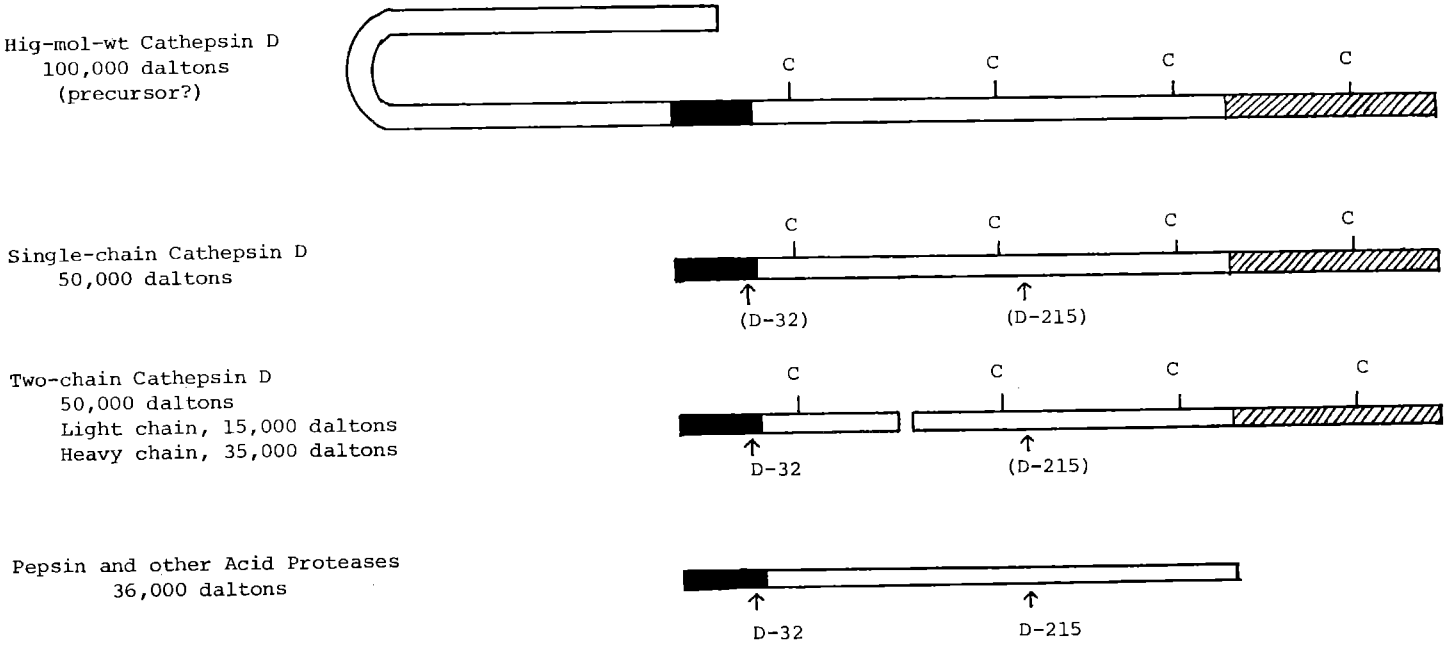


Fig. 4. Polypeptide chain arrangements in porcine spleen cathepsin D. The 50,000-dalton enzyme can be either single- or two-chain in structure. The dark areas represent the regions where the $NH_2$-terminal sequences are homologous to other carboxyl proteases (Fig. 5). The shaded regions illustrate the "tail", which are unique for cathepsin D. D-32 and D-215 are the active-center aspartyl residues. C indicates the approximate carbohydrate positions. The high-molecular weight isozyme is immunochemically indistinguishable from the heavy and single chain isozymes.

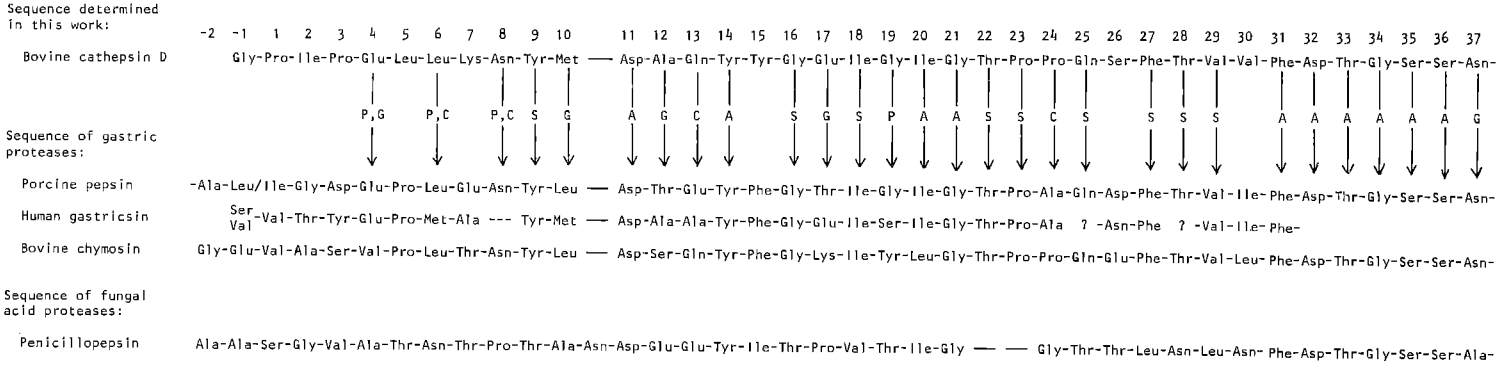


Fig. 5. The N-terminal sequence of bovine spleen cathepsin D[40] and its sequence homology with gastric and fungal proteases. The residue numbers on the top line are those for porcine pepsin[14]. The vertical arrows designate the positions where identical residues are found for A, all carboxyl proteases; S, stomach proteases, P, pepsin; G, gastricsin; and C, chymosin. Porcine cathepsin D differs from the bovine enzyme in this region only on position 5, which is Leu.

ory or specific recognition functions. Whether this is a general pattern for the intracellular carboxyl proteases remains to be seen. It is interesting to note, however, that two other intracellular carboxyl proteases of known molecular weights are both larger than pepsin. Yeast proteinase A is about 40,000 daltons[12] and lysosomal cathepsin E is more than 300,000 daltons[45].

## Zymogens of Carboxyl Proteases

The precursors of carboxyl proteases can be divided into two types by their activation mechanisms. As listed in Table 2, the acid activated zymogens, especially the gastric zymogens, have long been known. The activation mechanism of pepsinogen has been extensively studied in recent years and is well documented in recent reviews[2,46]. Chemically, the activation of pepsinogen to pepsin involves the removal of the $NH_2$-terminal 44 residues of the zymogen. Under the conditions similar to those expected in the stomach, the zymogen is activated by an intramolecularly catalyzed mechanism[47,48], which likely utilizes nascent catalytic apparatus of pepsin[49]. This mechanism appears to be a universal one for the acid activated zymogens.

For the second group in Table 2 the zymogens are apparently not converted to the enzymes upon acidification. At least two sizes of high molecular weight-apparent renin precursors

TABLE II

Precursors of carboxyl proteases

| Precursors | Source | Molecular Weights | | Review or Original References |
|---|---|---|---|---|
| | | Zymogen | Enzyme | |
| | | daltons | | |
| (A) Activated by acidification | | | | |
| Pepsinogen | Stomach | 40,000 | 35,000 | 4 |
| Gastricsinogen | Stomach of high mammals | ~40,000 | | 4 |
| Prochymosin | Stomach of | 40,000 | 35,000 | 4 |
| Uropepsinogen | Plasma and urine | ~40,000* | ~35,000* | 35 |
| Zymogen for seminal plasma acid protease | Seminal plasma | 40,000 | 35,000 | 35 |
| (B) Activated by other proteases | | | | |
| Prorenin | Kidney and submaxillary gland | 140,000 and 60,000 | 42,000 | 13 |
| Procathepsin D | Lysosomes | ~100,000 | 50,000 | 40 |

* The molecular weight of uropepsinogen and uropepsin have not been accurately determined. The figures shown in the table are based on the evidence that uropepsinogen is gastric in origin and therefore is very similar to pepsinogen.

have been observed. One has a molecular weight of about 55,000–60,000 (big renin), while the other is about 140,000 (big-big renin)[13]. Several proteases, including kalikrien and an unidentified sulfhydryl protease have been shown to activate the renin precursors[50,51]. It is interesting that a great many similarities are found in the prorenin and the possible procathepsin D. The high molecular weight species are similar in sizes (Table 2); each of these proenzymes contains only very low enzymic activity[13,40], and both enzymes are contained in intracellular granules, cathepsin D in lysosomes, and renin in presumably secreting granules[52]. These apparent similarities may indicate a common pathway in the genesis at the subcellular level.

## Carboxyl Proteases Involving Regulatory Cascades

Two regulatory cascades are known to involve carboxyl proteases. The system catalyzed by renin involving the release of angiotensin I, a decapeptide, from angiotensigogen in the plasma, is well studied. By the action of a carboxydipeptidase (converting enzyme), angiotensin I is converted to angiotensin II, which stimulates the synthesis of aldosterone in the adrenal cortex. The amount of renin in the plasma is one of the regulatory factors. Renin differs from other carboxyl proteases in two respects: the optimal pH of its catalytic activity is about 6, and its specificity is much more stringent. However, renin is inactivated by carboxyl protease inhibitors DAN, EPNP, and pepstatin[13]. The molecular weight is also very similar to pepsin. It seems reasonable to hypothesize that renin is homologous to pepsin in activity center structure and overall three-dimensional structure. The higher optimal pH for activity may be achieved by displacements of the pK's of catalytic residues in the active center. The stringent specificity requirement of about eight amino acid residues near the site of cleavage in angiotensinogen is compatible with the extended binding site of 7–8 residues in the extracellular carboxyl proteases discussed above.

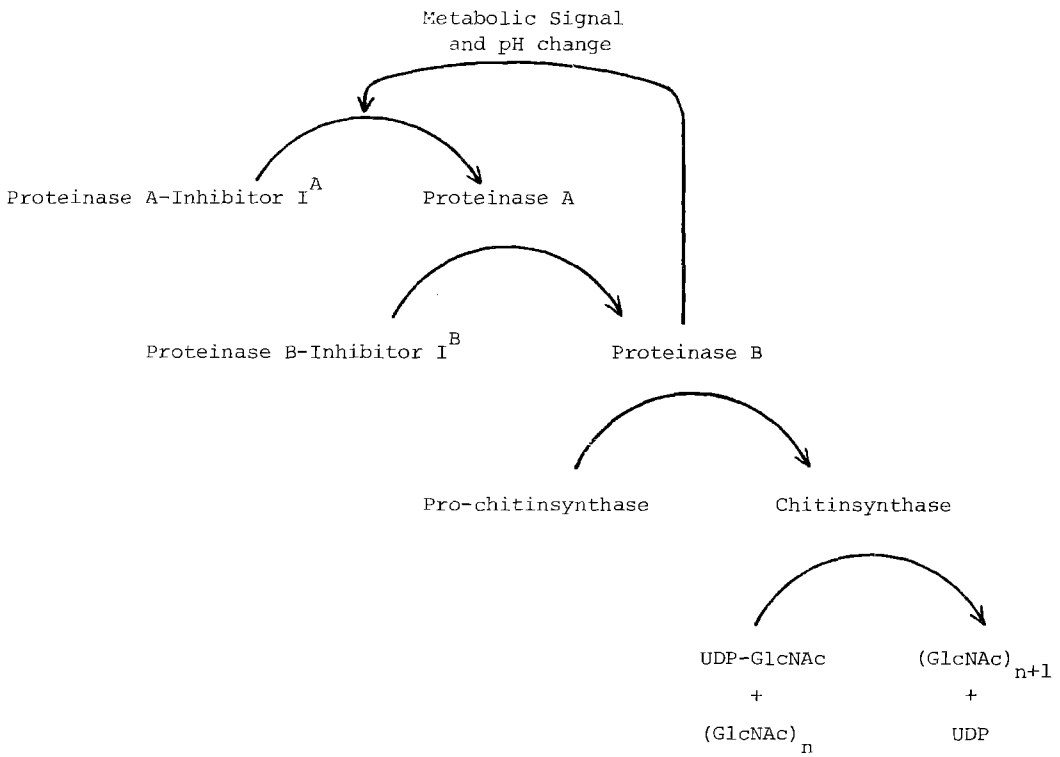


Fig. 6. Cascade mechanism in the activation of chitinsynthase in yeast, as proposed by HOLZER[12]. Proteinase A is a carboxyl protease. Proteinase B is a sulfhydryl protease.

A second carboxyl protease which may function in a regulatory cascade was proposed by HOLZER in the involvement of proteinase A of yeast in the activation of chitin synthetase[12]. As shown in Figure 6, proteinase A can remove a protein inhibitor, $I^B$, from a sulfhydryl requiring proteinase B, which in turn activates chitin-synthase from a precursor by a limited proteolysis. In the yeast cells proteinase A and B are located in the vacuoles and their respective protein inhibitors are located in the cytosol[12]. The precise *in vivo* mechanism of this cascade, which may involve more than one subcellular compartment, has not been worked out. However, it is interesting to note that possibly because of the compartmentation proteinase A does not have a stringent specificity characteristic for the enzymes involved in a regulatory cascade. In addition to yeast, a similar system may be present in *Phycomyces blakesleeanus*, a fungus[53].

## Evolution of Carboxyl Proteases

The chemical structures of carboxyl proteases determined so far indicate that they are all homologous in amino acid sequence. This suggests that they are derived from a common ancestral enzyme by divergent evolutionary processes. The closest remaining examples to such an ancestral enzyme are the fungal carboxyl proteases, e.g., penicillopepsin. As already described above, the mammalian gastric enzymes retain the same three-dimensional foldings as well as, to a somewhat lesser extent, sequence homology.

It is now quite certain that carboxyl proteases were originally derived from a gene duplication process. This was originally suspected from the high sequence similarity around the active center of Asp-32 and Asp-215 in pepsin[14]. The concrete evidence has come from the tertiary structures in which the nearly identical chain windings in the N- and C-terminal lobes were found[54]. Such two-fold symmetry is illustrated in Figure 7, in which a high degree of similarity between the two domains can be easily recognized. The relationships in the interacting $\beta$-strands are diagrammed in Figure 8. The chain windings and the relative active aspartyl positions in two halves are apparently the same.

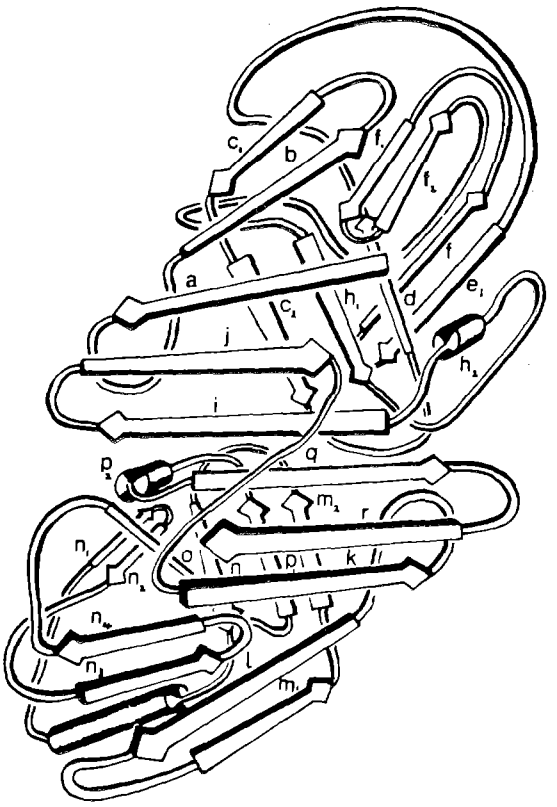

Fig. 7. A schematic drawing of the tertiary structure of a carboxyl protease from *Endothia parasitica*, viewed along the two-fold axis of symmetry[53].

Even though there are only 14 identical residues in the sequence alignment of the N- and C-terminal domains of pepsin (10 for penicillopepsin), 61 residues are at "topologically equivalent" positions in the tertiary-structural comparison[54].

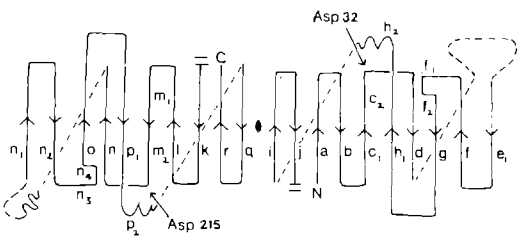


Fig. 8. A schematic drawing of the secondary structure found in the carboxyl proteases. The strands are the $\beta$-structures found in the crystal structures of carboxyl proteases, as shown in Figures 2 and 7. Strands j and k are connected by a short peptide (residues 72–76). The labeling of the strands is from N-terminal[53]. The N-terminal lobe consists of strands a–j and the C-terminal lobe consists of strands k–r. The symmetry of the two lobes can easily be seen.

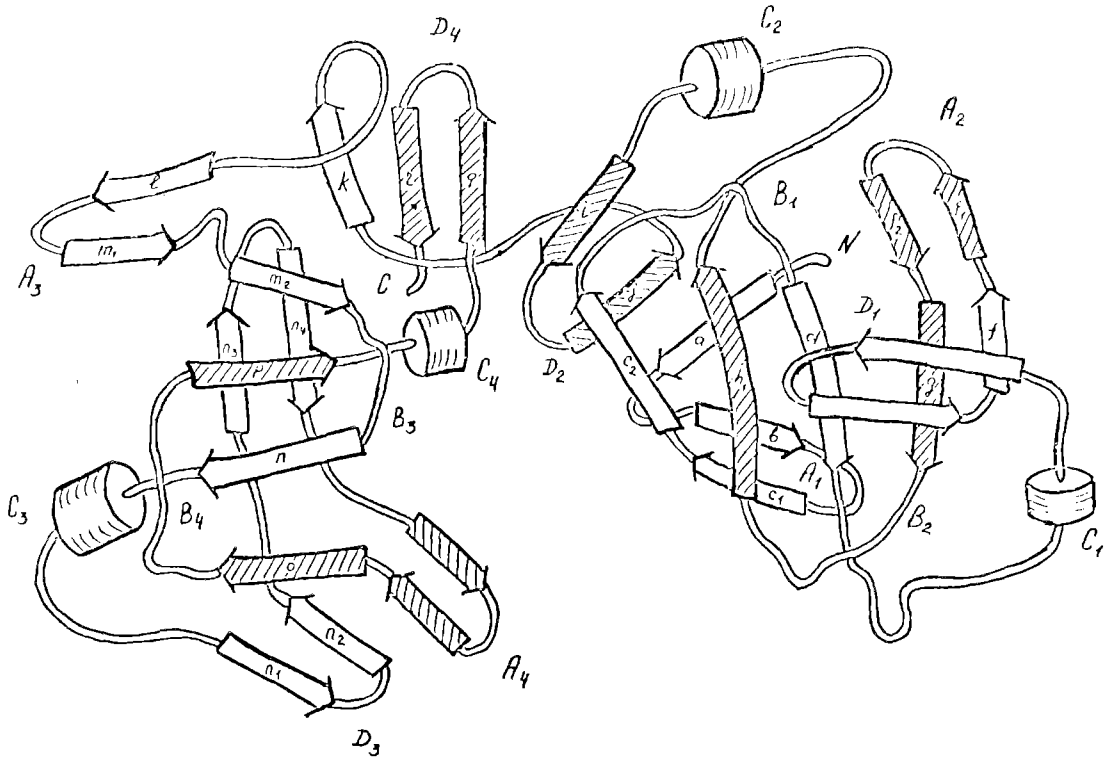


Fig. 9. A schematic drawing of four structural units found in pepsin. The drawing is taken from the work of ANDREEVA and GUSTCHINA[55]. The first structural unit consists of strands a–f; the second structural unit, strands f–j (shaded); the third structural unit, strands k–$n_4$; and the fourth structural unit, the rest of the C-terminal strands (shaded). The first two structural units comprised the N-terminal lobe (Fig. 7 and 8). The second and third structural units comprised the C-terminal lobe.

Recently, ANDREEVA and GUSTCHINA found an additional two-fold symmetry in polypeptide chain foldings within each of the two lobes in pepsin[55]. Each of the basic asymmetric units, approximately 80 residues or one-fourth of the pepsin molecule, consists of four structural elements: a $\beta$-structure hairpin, a wide loop, a helical section, and a second $\beta$-structure hairpin. Figure 9 shows a schematic drawing of these four interacting basic structural units in pepsin. The conformation among these four units is indeed very similar, especially the second unit (residues 86-184) and the third unit (residues 194–272), which ANDREEVA and GUSTCHINA showed to be nearly superimposeable. Such internal conformational homology would be an unlikely result of structural requirements of the protein; they must have also resulted from a gene duplication.

Based on these structural data, the overall genetic events in the emergence of carboxyl proteases are summarized in Figure 10. The ancestral gene must have coded for a protein of approximately 8500 daltons in molecular weight. The function of this protein cannot be certain now. (The complications in interwining between the two basic structural units in each lobe {Fig. 9} seems to prohibit the idea that they might have associated into a carboxyl protease with four subunits.) By means of gene duplication and fusion a new gene was formed which coded for a 17,000-polypeptide containing an internal two-fold symmetry. Two of these polypeptide subunits can readily associate to form a carboxyl protease. After further gene duplication and mutation the enzyme would have contained two non-identical subunits. Finally, through a second gene fusion a single-chain carboxyl protease emerged with two similar lobes in the molecule.

The amino acid sequences of the gastric carboxyl proteases are very homologous, including the activation peptide (zymogen) portion of the structure[4]. This suggests that these enzymes were separated at a relatively recent evolutionary time. Apparently, the microbial carboxyl proteases have no zymogen[19]. The activation peptide of an ancestral gastric zymogen must

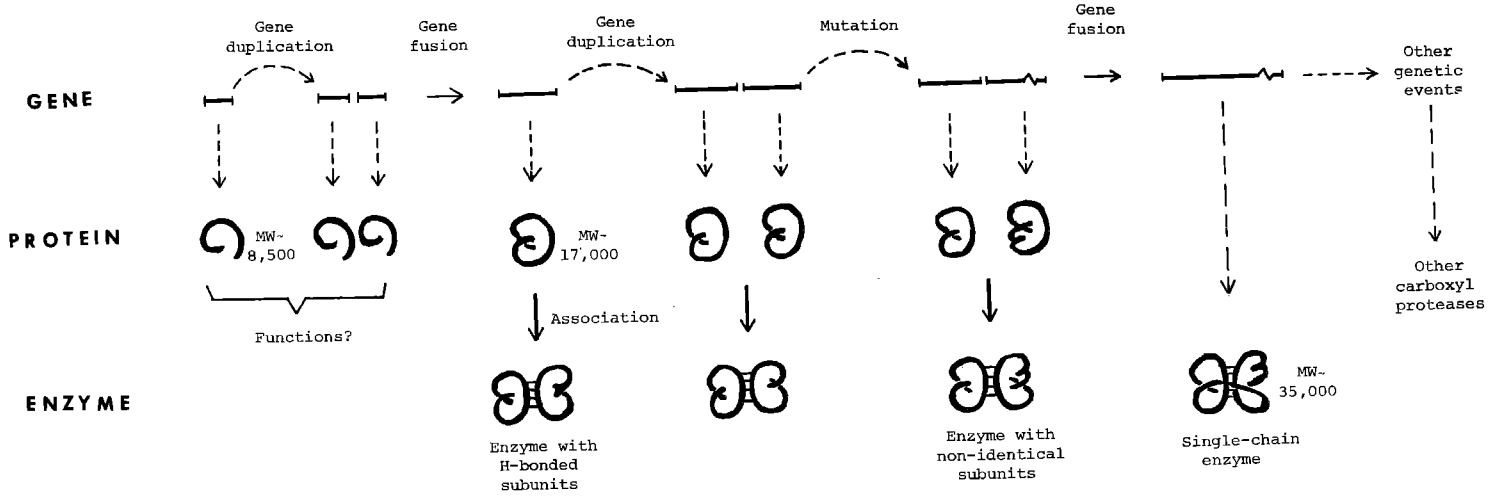


Fig. 10. A schematic presentation of genetic events in the evolution of carboxyl proteases. The ancestral gene was probably one-fourth of the current size. After the first gene duplication and fusion, the primitive carboxyl protease may have contained two subunits. A second gene duplication and fusion produced a single-chain enzyme. These hypothetical events are generated from the structural information which showed that the carboxyl proteases contain four similar structural units organized in two distinctly homologous lobes.

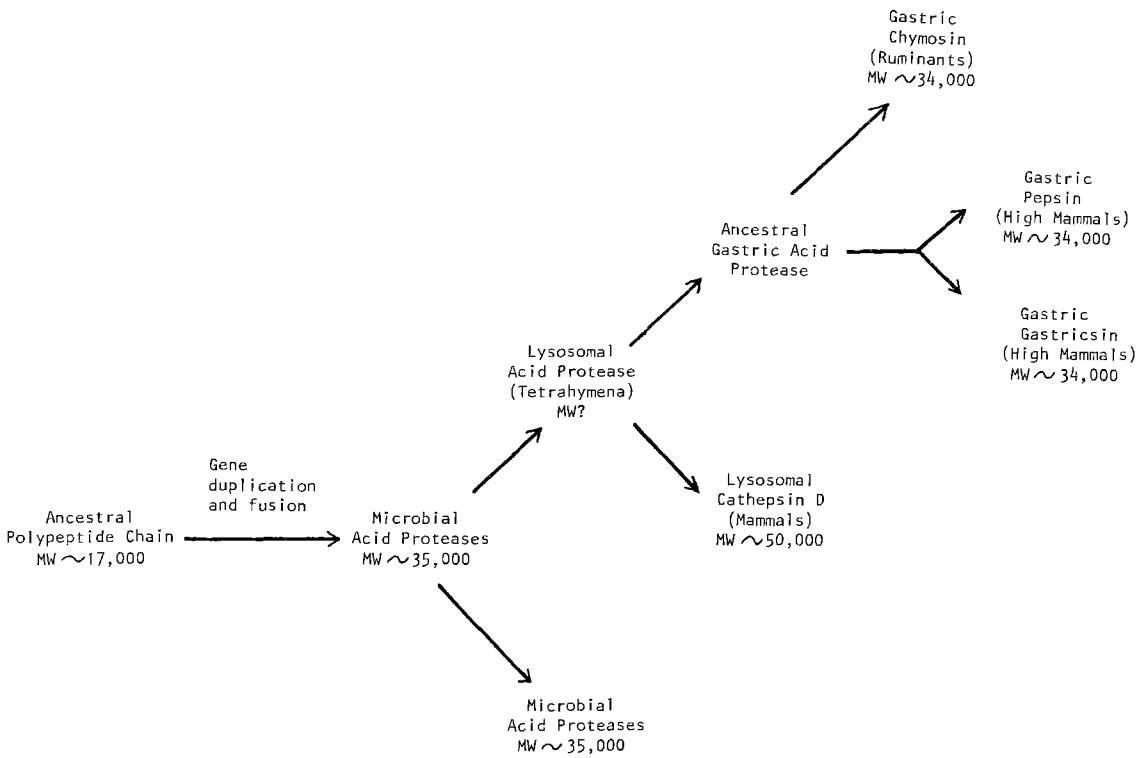


Fig. 11. The evolutionary relationships of carboxyl proteases. The hypothetical scheme is based on the structural information discussed in this article.

have been acquired by gene fusion with the emergence of gastric secretory processes. This may have occurred independently for prorenin and procathepsin D.

A close comparison of the N-terminal sequences reveals that cathepsin D possesses the characteristics of an ancestral chain of gastric carboxyl proteases. As illustrated in Figure 5, cathepsin D shares some unique residues from each of the three gastric proteases. These include residues 6, 8, and 19 for pepsin, residues 10, 12, and 17 for human gastricsin, and residues 6, 8, 13, and 24 for chymosin. The sequence homology between the gastric carboxyl proteases and cathepsin D appears to be stronger than between the microbial carboxyl proteases and the lysosomal enzyme. These observations suggest that divergence of the gastric proteases from lysosomal carboxyl proteases may have been a relatively late event in evolution. Since lysosomes and the lysosomal carboxyl proteases are found in lower forms of life, such as *Tetrahymena pyriformis*[11], it seems reasonable to suggest that the origin of the gastric carboxyl proteases was rooted in the lysosomal cathepsin D. An overall scheme of the possible evolutionary pathways is shown in Figure 11. A cathepsin D-like carboxyl protease in a primitive lysosome is depicted to be the hypothetical ancestral chain of both the gastric enzymes and cathepsin D. The latter can be expected to have fewer structural changes from the common ancestral chain because of an unaltered function in lysosomes. Following this line of reasoning, one might expect that proteinase A of yeast, a vacuole enzyme, would be an evolutionary ancestor of mammalian cathepsin D.

## Acknowledgments

The studies cited from our laboratory were supported in part by Research Grants AM-01107 and GM-20212 from the National Institute of Health. The author wishes to thank Professor Natalia Andreeva for making the results available to me prior to their publication, and to Dr. Jean Hartsuck for constructive criticism.

## References


1. Tang, J., 1976. Trends in Biochemical Sciences, 1, 205–208.
2. Hartsuck, J. A., and Tang, J., 1978. In Regulatory Proteolytic Enzymes and Their Inhibitors (Magnusson, S., Otteson, M., Foltmann, B., Danø, K., and Neurath, H., editors), pp. 35–36, Pergamon Press, Oxford.
3. Tang, J., 1977. Nature, 266, 119–120.
4. Foltmann, B., and Pedersen, V. B., 1977. In Acid Proteases, Structure, Function, and Biology (Tang, J., editor), pp. 3–22, Plenum Press, New York.
5. Kushnev, I., Rapp, W., and Burtin, P., 1964. J. Clin. Invest., 43, 1983–1993.
6. Samloff, I. M., and Townes, P. L., 1970. Science, 168, 144–145.
7. Herriott, R. M., 1938. J. Gen. Physiol., 21, 501–540.
8. Garg, G. K., and Virupaksha, T. K., 1970. Eur. J. Biochem., 17, 4–12.
9. Morihara, K., 1974. In Adv. Enzymol., 41, 179–243.
10. Ottesen, M., and Rickert, W., 1970. Compt. Rend. Trav. Lab. Carlsberg, 37, 301–325.
11. Dickie, N., and Liener, I. E., 1962. Biochim. Biophys. Acta., 64, 41–51.
12. Holzer, H., Bünning, P., and Meussdoerffer, F., 1977. In Acid Proteases, Structure, Function, and Biology (Tang, J., editor), pp. 271–290, Plenum Press, New York.
13. Inagami, T., Murakami, K., Misono, K., Workman, R. J., Cohen, S., and Suketa, Y., 1977. In Acid Proteases, Structure, Function, and Biology (Tang, J., editor), pp. 225–248, Plenum Press, New York.
14. Tang, J., Sepulveda, P., Marciniszyn, J., Jr., Chen, K. C. S., Huang, W.-Y., Tao, N., Liu, D., and Lanier, J. P., 1973. Proc. Nat. Acad. Sci. USA., 70, 3437–3439.
15. Foltmann, B., Pedersen, V. B., Jacobsen, H., Kauffman, D., and Wybrandt, G., 1977. Proc. Nat. Acad. Sci., USA, 74, 2321–2324.
16. Hsu, I.-N., Delbaere, L. T. J., James, M. N. G., and Hofmann, T., 1977. Nature, 266, 140–145.
17. Rajagopolan, T. G., Stein, W. H., and Moore, S., 1966. J. Biol. Chem., 241, 4295–4297.
18. Bayliss, R. S., Knowles, V. R., and Wybrandt, G. B., 1966. Biochem. J., 113, 377–386.
19. Hofmann, T., 1974. In Adv. in Chem., 139, 146–185, Am. Chem. Soc.
20. Tang, J., 1971. J. Biol. Chem., 246, 4510–4517.
21. Chen, K. C. S., and Tang, J., 1972. J. Biol. Chem., 247, 2566–2574.
22. Clement, G. E., 1973. In Prog. in Bioorg. Chem., 2, 177–238.
23. Andreeva, N., Fedorov, A., Gustchina, A., Riskulov, R., Safro, M., and Shutzkever, N., 1978. J. Mol. Biol. (Russian), 12, 922–927.
24. Subramanian, E., Swan, I. D. A., Liu, M., Davies, D. R., Jenkins, J. A., Tickle, I. J., and Blundell, T. L., 1977. Proc. Nat. Acad. Sci. USA, 74, 556–559.
25. Tang, J., 1963. Nature, 199, 1094–1095.
26. Powers, J. C., Harley, A. D., and Myers, D. V., 1977. In Acid Proteases, Structure, Function, and Biology (Tang, J., editor), pp. 141–157, Plenum Press, New York.
27. Fruton, J. S., 1976. In Adv. Enzymol., 44, 1–36.
28. Hartsuck, J. A., and Tang, J., 1972. J. Biol. Chem., 242, 2575–2580.
29. Lunblad, R. L., and Stein, W. H., 1966. J. Biol. Chem., 244, 154–160.
30. James, M. N. G., Hsu, I.-N., and Delbaere, T. J., 1977. Nature, 267, 808–813.
31. Antonov, V. K., Ginodman, L. M., Kapitannikov, Yu. V., Barshevskaya, T. N., Gurova, A. G., and Rumsh, L. D., 1978. FEBS Letters, 88, 87–90.
32. Huber, R., and Bode, W., 1978. In Regulatory Proteolytic Enzymes and Their Inhibitors (Magnusson, S., Ottesen, M., Foltmann, B., Danø, K., and Neurath, H., editors), pp. 15–34, Pergamon Press, Oxford.
33. Marciniszyn, J., Jr., Hartsuck, J. A., and Tang, J., 1976. J. Biol. Chem., 251, 7088–7094.
34. Rich, D. H., and Sun, E. T., 1977. In Peptides Proc. Fifth Amer. Peptide Symp. (Goodman, M., and Meinhofer, J., editors), pp. 209, J. Wiley and Sons, New York.
35. Barrett, A. J., 1977. In Proteinases in Mammalian Cells and Tissues (Barrett, A. J., editor), pp. 209–233, North Holland, Amsterdam.
36. Dean, R. T., 1975. Nature, 257, 414–416.
37. Smith, G. D., Murray, M. A., Nichol, L. W., and Trikojus, V. M., 1969. Biochim. Biophys. Acta., 171, 288–298.
38. Cunningham, M., and Tang, J., 1976. J. Biol. Chem., 251, 4528–4536.
39. Aoyagi, T., Morishima, H., Nishizawa, R., Kunimoto, S., Takeuchi, T., and Umezawa, H., 1972. J. Antibiotics, 25, 689–694.
40. Huang, J. S., Huang, S. S., and Tang, J., 1979. (Submitted for publication.)
41. Huang, J. S., Huang, S. S., and Tang, J., 1979. Proc. of FEBS. Symposium on Proteolytic Enzymes, FEBS Special Meeting on Enzymes, Dubrovnik-Cavtat, 1979. Pergamon (in Press)
42. Keilova, H., 1976. In Intracellular Protein Catabolism (Hanson, H., and Bohley, P., editors), pp. 237–251, J. A. Barth, Leipzig.
43. Liljas, A., and Rossman, M. G., 1974. Ann. Rev. Biochem., 43, 475–507.
44. Spatz, L., and Strittmatter, P., 1973. J. Biol. Chem., 248, 793–799.
45. Lapresle, C., 1971. In Tissue Proteinases (Barrett, A. J., and Dingle, J. T., editors), pp. 135–155, North Holland, Amsterdam.
46. Hartsuck, J. A., Marciniszyn, J., Jr., Huang, J. S., and Tang, J., 1977. In Acid Proteases, Structure, Function, and Biology (Tang, J., editor), pp. 85–102, Plenum Press, New York.
47. Al-Janabi, J., Hartsuck, J. A., and Tang, J., 1972. J. Biol. Chem., 247, 4628–4632.
48. Sunny, C. G., Hartsuck, J. A., and Tang, J., 1975. J. Biol. Chem., 250, 2635–2639.
49. Marciniszyn, J., Jr., Huang, J. S., Hartsuck, J. A., and Tang, J., 1976. J. Biol. Chem., 251, 7095–7102.
50. Inagami, T., Hirose, S., Matoba, T., Murakami, K., and Okamoto, K., 1977. Fed. Proc., 36, 2372.
51. Inagami, T., Takahashi, N., Yokosawa, N., and Takii, Y., 1979. In Intern. Symp. on Kinins (Suzuki, T., and Moriya, H., editors), Plenum Press, New York.

52. Gross, D. M. and Barajas, L., 1975. J. Lab. Clin. Med., 85, 467–477.
53. Fischer, E.-P., and Thomson, K. S., 1979. J. Biol. Chem., 254, 50–56.
54. Tang, J., James, M. N. G., Hsu, I.-N., Jenkins, J. A., and Blundell, T. L., 1978. Nature, 271, 618–621.
55. Andreeva, N. S., and Gustchina, A. E., 1979. Biochem. Biophys. Res. Commun., 87, 32–42.
56. Levy, M. R., and Chou, S. C., 1974. Biochim. Biophys. Acta., 334, 423–430.
57. Levy, M. R., Siddiqui, W. A., and Chou, S. C., 1974. Nature, 247, 546–549.
58. Virupaksha, T. K., and Wallenfels, K., 1974. FEBS Letters, 40, 287–289.