Protein J (2008) 27:343–354
DOI 10.1007/s10930-008-9143-2

Structural Organization of Precursors of Thermolysin-like Proteinases

Ilya V. Demidyuk · Eugene V. Gasanov · Dina R. Safina · Sergey V. Kostrov

Published online: 27 June 2008

Abstract The primary structures of the full-length precursors of thermolysin-like proteinases (TLPs) were systemically analyzed. Structural comparison of the precursor amino-terminal regions (ATRs) removed during maturation allowed us to divide the family into two groups: peptidases with short (about 50 amino acids) and long (about 200 amino acids) ATRs. The accumulation of mutations in the ATRs of both types proved to correlate with that in the catalytic domains. No classical signal peptides were identified in the short ATRs, but they contained a conserved PPL-motif near the initiation methionine. The functional role of the short ATRs and PPL-motif is currently unclear. The C-terminal regions (CTRs) of TLP precursors, which are often removed during maturation, too, are found in about a half of precursors with long ATRs, but occur more rarely in precursors with short ATRs. CTRs in TLP precursors contain previously identified conserved domains typical for many other proteins and likely underlie the interaction with high molecular weight substrates.

Keywords Classification · Propeptide · Protein motif · Protein precursor · Thermolysin-like proteinase

Abbreviations

TLP	Thermolysin-like proteinase
ATR	Amino-terminal region
CTR	Carboxy-terminal region

I. V. Demidyuk (✉) · E. V. Gasanov · D. R. Safina · S. V. Kostrov
Laboratory of Protein Engineering, Institute of Molecular Genetics, Russian Academy of Sciences, Kurchatov sq. 2, Moscow 123182, Russia
e-mail: duk@img.ras.ru

1 Introduction

Thermolysin-like proteinases (TLPs), or peptidases of the M4 family [1], are a group of metalloendopeptidases found in dozens of gram-positive and gram-negative bacterial species. In addition, TLPs were identified in fungi and the archaeon *Methanosarcina acetivorans* [2]. TLPs contain a zinc ion in the catalytic site [3] and prefer to hydrolyze substrates with large hydrophobic amino acids (Leu or Phe) at the P1′ position [4–6] (using the numbering system of Schechter and Berger [7]).

Similar to many other proteolytic enzymes, TLPs are synthesized as precursors carrying structural elements that are absent in mature molecules. Such additional sequences can be localized in both N- and C-terminal regions of the precursor. Many amino-terminal extensions of the precursors contain a signal peptide, providing for extracellular secretion of the protein, and a prosequence. The prosequences act as an intramolecular chaperone [8] that modulates protein folding [9–12], at least in some TLPs. In addition, the proregion of M4 peptidases can inhibit the corresponding mature proteins [12–15] and mediate their secretion [9, 16]. The C-terminal regions of TLP precursors seem to mediate their binding to insoluble substrates, as supported by the limited data available [17–21].

Our previous studies [22, 23] and ample data obtained by other groups indicate several structural types of TLP precursors. In this work, the primary structures of the full-length precursors of the M4 peptidase family were systemically analyzed.

2 Materials and Methods

Multiple sequence alignments were constructed by ClustalX 1.8 using the Gonnet series of protein weight matrices [24].

The set of sequences of full-length TLP precursors used in this work (Table 1) was compiled as follows. The sequences of M4 peptidases available in the MEROPS database (release 7.20) were complemented by four TLP sequences translated from GenBank: Ser_pro, Brb_lie, Gib_zea, and V_har-1. All of these 146 sequences were aligned, and the following sequences representing fragments of the full-length precursors were excluded: MER01042, MER01048, MER03790, MER03915, MER05572, MER12295, MER16511, MER 19065, MER20843, MER27031, MER27032, MER27033, MER27034, MER40474, MER41301, MER48741 and MER50323. Among the remaining sequences, some that were closely related (identity exceeding 85%) were recognized: MER01025, MER01034, MER01353, and MER019 27; MER01026 and MER01027; MER01035 and MER01 038; MER01030, MER01354, MER03181, and MER21824; MER04711 and MER14408; MER05727, MER11853, MER29943, and MER30073; MER19561 and MER27496; MER20835, MER28888, and MER39814; MER20840, MER28890, and MER39811; MER21804, MER28887, and MER39810; MER25370, MER28622, MER29961, MER40 142, and MER50804; MER26466 and MER45739; MER2 7846 and MER52672; MER28889 and MER39813; MER30 706 and MER30781; MER39812 and MER43807; and MER48895 and MER50231. A single representative of each group (underlined) was included in the final set of 100 sequences so as to avoid overweighting closely related family members. The resulting set of sequences was realigned.[1]

For the alignment of the amino-terminal regions (ATRs) of the precursors, all mature parts were discarded by two steps. At the first step, the cutoff point was set after the first unambiguously alignable cluster of the mature protein (after position 94 in thermolysin precursor; the numbering starts from the first amino acid in the mature protein). At the second step, the realigned sequences were cut off before the experimentally identified processing sites (after position −9 in thermolysin precursor). Note that proteal-ysin-like short ATRs in the resulting sequence set still included 15–20 amino acids of the mature part. The sequences of Par_sp and Par_sp-1 deduced from the *Parachlamidia* sp. UWE25 genomic sequence demonstrate good similarity with other TLPs in the region of the Zn^{2+}-binding motif (HEXXH_E) only. They could not be aligned with the whole set of sequences analyzed. Hence, Par_sp and Par_sp-1 sequences were aligned with Ser_pro and thermolysin sequences alone, and their C-terminal sequences after position −9 in thermolysin were excluded. The resulting N-terminal regions were added to the resulting set. This set of aligned sequences was used to construct the dendrogram shown in Fig. 1.

In the case of the carboxy-terminal regions (CTRs), long sequences corresponding to the pre- and propeptides, as well as mature parts, were excluded from the alignment of the whole set of the full-length precursors. The resulting sequence set included the regions following the last unambiguously alignable cluster of the mature part (following amino acid 316 in thermolysin). Then, the sequences shorter than 30 amino acids were excluded and the remaining sequences were realigned.

Signal peptides were identified in the amino acid sequences using the SignalP 3.0 server (http://www.cbs.dtu.dk/services/SignalP). The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models [25].

The sequence logos [26] were generated using the WebLogo tool (http://weblogo.berkeley.edu) [27]. The regions −144…−106, −90…−78, and −47…−16 of long ATRs and PPL motifs were extracted, and columns containing gaps in at least 95% of the sequences were eliminated.

Phylogenetic analysis was performed using PHYLIP (version 3.65) [28]. Protein distances were calculated using the Jones–Taylor–Thoruton matrix. The unweighted pair group method with arithmetic mean (UPGMA) was applied to clustering distance data.

3 Results and Discussion

3.1 Structure and Function of Precursor Amino-Terminal Regions

Comparison of the amino acid sequences suggested the division of M4 family ATRs into two main groups (Fig. 1).

The first group includes 74 enzymes (from the resulting set of 100 sequences described in Sect. 2) with 183–287 amino acid long amino-terminal regions (long ATRs hereafter) removed during maturation in all studied cases. (ATR size is given for the enzymes with known processing sites.) All experimentally described proteins of this group are secreted extracellular enzymes with a typical leader peptide (presequence) at the N-terminus that can be reliably identified by prediction algorithms (e.g., SignalP 3.0 [25]). In the proteins with long ATRs, the signal peptide is followed by a region called prosequence, propeptide, or proregion. The propeptide length ranges from 162 to 264 amino acids.

[1] The current MEROPS database (release 7.70) was supplemented by new M4 peptidase sequences. Each of these sequences was analyzed individually. The results of the analysis are summarized in the footnotes to Table 1.

Table 1 List of peptidases of the M4 family analyzed in this work

Designation	Source	MEROPS number or GenBank protein ID	Notes[a]
Aer_cav	*Aeromonas caviae* T-64 *(Aeromonas punctata)*	MER29943	L, C
Alt_sp	*Alteromonas* sp. O-7	MER19099	L, C
Alt_sp-1	*Alteromonas* sp. O-7	MER24591	L, C
Ant_bac	*Antarctic bacterium* str. 643	MER12255	L, C
Art_sp[b]	*Arthrobacter* sp. FB24	MER50759	S
B_amy	*Bacillus amyloliquefaciens*	MER01035	L
B_cal	*Bacillus caldolyticus*	MER01034	L
B_cer	*Bacillus cereus* ATCC14579	MER01030	L
B_cer-1	*Bacillus cereus* ATCC14579	MER28887	L
B_cer-2[b]	*Bacillus cereus* ATCC14579	MER28889	L, C
B_cer-3	*Bacillus cereus* ATCC14579	MER28890	L
B_cer-4	*Bacillus cereus* ATCC14579	MER28888	L
B_cer-5	*Bacillus cereus* ATCC14579	MER11123	L
B_meg	*Bacillus megaterium* ATCC14581	MER01031	L
B_sub	*Bacillus subtilis* 168	MER01029	L
B_sub-1	*Bacillus subtilis* 168	MER01032	L
B_thu[b]	*Bacillus thuringiensis* sv. *konkukian* str. 97-27	MER39812	L
B_vie[b]	*Bacillus vietnamensis*	MER38281	L
TLN	*Bacillus thermoproteolyticus*	MER01026	L
Bor_bro	*Bordetella bronchiseptica* RB50	MER30781	S
Bra_jap	*Bradyrhizobium japonicum* USDA110	MER26988	
Bre_bre	*Brevibacillus brevis*	MER01028	L
Brb_lie	*Brevibacterium liens* BL2	ZP_00380881	S, C
Bur_cen[b]	*Burkholderia cenocepacia* AU1054	MER50859	L
Bur_cep[b]	*Burkholderia cepacia*	MER28622	L
Chr_vio	*Chromobacterium violaceum* ATCC12472	MER27350	L
Chh_sal	*Chromohalobacter salexigens* DSM3043	MER50897	S
Clo_ace	*Clostridium acetobutylicum* ATCC824	MER14937	L, C
Clo_ace-1	*Clostridium acetobutylicum* ATCC824	MER14941	L
Clo_per	*Clostridium perfringens*	MER02103	L
Col_psy[b]	*Colwellia psychrerythraea* 34H	MER52723	C
Cyt_hut	*Cytophaga hutchinsonii*	MER23927	L, C
Ent_fae	*Enterococcus faecalis* V583	MER02810	L
Fra_sp	*Frankia* sp. CcI3	MER51510	S
Fra_sp-1[b]	*Frankia* sp. EAN1pec	MER51747	S
Gib_zea	*Gibberella zeae* PH-1	EAA76052	S, C
Leg_lon	*Legionella longbeachae*	MER02394	L
Leg_pne	*Legionella pneumophila* str. Lens	MER01039	L
Leg_pne-1	*Legionella pneumophila* str. Lens	MER40780	L
Leg_pne-2	*Legionella pneumophila* str. Lens	MER40781	L
Leg_pne-3	*Legionella pneumophila* str. Lens	MER40782	L
Lep_int	*Leptospira interrogans* serovar *lai* str. 56601	MER22286	L, C
Lep_int-1	*Leptospira interrogans* serovar *lai* str. 56601	MER24877	L, C
Lep_int-2[b]	*Leptospira interrogans* serovar *lai* str. 56601	MER24878	L, C
Lis_iva	*Listeria ivanovii* subsp. *Ivanovii*	MER26466	L
Lis_mon	*Listeria monocytogenes* EGD	MER01047	L
Lis_see	*Listeria seeligeri*	MER45740	L
Met_ace	*Methanosarcina acetivorans* C2A	MER17697	L, C

Table 1 continued

Designation	Source	MEROPS number or GenBank protein ID	Notes[a]
Met_ace-1	*Methanosarcina acetivorans* C2A	MER17698	S
Myx_xan	*Myxococcus xanthus*	MER17624	L, C
Neu_cra[b]	*Neurospora crassa*	MER29119	
Noc_sp	*Nocardioides* sp. JS614	MER49523	S, C
Nos_pun	*Nostoc punctiforme* PCC73102	MER24259	S
Nos_sp[b]	*Nostoc* sp. PCC7120	MER16719	S
Oce_ihe	*Oceanobacillus iheyensis* HTE83	MER22038	L
Pae_pol	*Paenibacillus polymyxa*	MER01033	L
Par_sp[b]	*Parachlamydia* sp. UWE25	MER38366	
Par_sp-1[b]	*Parachlamydia* sp. UWE25	MER41302	
Pec_car	*Pectobacterium carotovorum*	MER01045	S
Pho_lum	*Photorhabdus luminescens* subsp. *laumondii* TTO1	MER33481	S
Pse_sp	*Pseudoalteromonas* sp. A28	MER19098	L, C
P_aer	*Pseudomonas aeruginosa* PAO1	MER01024	L
P_syr	*Pseudomonas syringae* pv. *tomato* str. DC3000	MER27846	S
Ren_sal	*Renibacterium salmoninarum*	MER02083	
Ser_mar	*Serratia marcescens* ATCC21074	MER01046	S
Ser_pro	*Serratia proteamaculans* 94	MER59439 (AAV88082)	S
She_ama[b]	*Shewanella amazonensis* SB2B	MER48811	L, C
She_ama-1[b]	*Shewanella amazonensis* SB2B	MER49928	L
She_bal	*Shewanella baltica* OS155	MER48895	L, C
S_aur	*Staphylococcus aureus* subsp. *aureus*	MER04711	L
S_chr	*Staphylococcus chromogenes*	MER11075	L
S_epi	*Staphylococcus epidermidis* ATCC12228	MER01869	L
Str_ave[b]	*Streptomyces avermitilis* MA-4680	MER28561	L
Str_ave-1	*Streptomyces avermitilis* MA-4680	MER28563	L, C
Str_ave-2	*Streptomyces avermitilis* MA-4680	MER28564	L
Str_ave-3	*Streptomyces avermitilis* MA-4680	MER28519	L, C
Str_ave-4	*Streptomyces avermitilis* MA-4680	MER28565	L, C
Str_ave-5	*Streptomyces avermitilis* MA-4680	MER28566	L
Str_ave-6	*Streptomyces avermitilis* MA-4680	MER28562	S
Str_ave-7	*Streptomyces avermitilis* MA-4680	MER28567	L, C
Str_coe	*Streptomyces coelicolor* A3(2)	MER19351	L
Str_coe-1	*Streptomyces coelicolor* A3(2)	MER11085	L, C
Str_coe-2	*Streptomyces coelicolor* A3(2)	MER12275	L
Str_coe-3	*Streptomyces coelicolor* A3(2)	MER11082	S
Str_exf[b]	*Streptomyces exfoliatus* SMF13	MER52079	L, C
Str_gri	*Streptomyces griseus*	MER04744	L, C
Ther_sp	*Thermoactinomyces* sp. 27a	MER29719	L, C
V_ang	*Vibrio anguillarum*	MER01044	L, C
V_cho	*Vibrio cholerae* O1	MER01041	L, C
V_flu	*Vibrio fluvialis*	MER19097	L, C
V_har	*Vibrio harveyi*	MER20240	L, C
V_har-1	*Vibrio harveyi*	AAT68711	L, C
V_par	*Vibrio parahaemolyticus* RIMD2210633	MER27936	L, C
V_pro	*Vibrio proteolyticus*	MER01043	L, C
V_sp	*Vibrio* sp. T-1800	MER29796	L, C

Table 1 continued

Designation	Source	MEROPS number or GenBank protein ID	Notes[a]
V_vul	*Vibrio vulnificus* CMCP6/YJ016	MER03353	L, C
V_vul-1	*Vibrio vulnificus* CMCP6/YJ016	MER25442	L, C
Xan_axo[b]	*Xanthomonas axonopodis* pv. *citri* str. 306	MER19561	S
Xan_axo-1[b]	*Xanthomonas axonopodis* pv. *citri* str. 306	MER19560	L
Xan_cam	*Xanthomonas campestris* pv. *campestris* str. ATCC33913	MER19416	S

[a] L, precursor includes a long ATR; S, precursor includes a short ATR; C, precursor includes an additional CTR

[b] The current MEROPS database (release 7.70) was supplemented by new M4 peptidase sequences. In addition, some incomplete sequences excluded from our analysis were extended. Most of the added full-length precursor sequences have high sequence identity to those listed above: TLP from *Arthrobacter aurescens* TC1 (MER75195) is 73% identical to Art_sp; TLP from *Bacillus cereus* strain ATCC14579 (MER54507) is 81% identical to B_cer-2; TLPs from *Bacillus weihenstephanensis* KBAB4 (MER62602) and *Bacillus cereus* strain ATCC10987 (MER66067) are 92 and 94% identical to B_thu, respectively; TLPs from *Bacillus* sp. NRRL B-14911 (MER62589) and *Bacillus cereus* subsp. *cytotoxis* NVH 391-98 (MER62591 and MER62592) are 75 and 58% identical to B_vie, respectively; TLPs from *Burkholderia* sp. 383 (MER56883), *Burkholderia thailandensis* E264 (MER58397), and *Burkholderia cepacia* AMMD (MER73158) are 94, 80, and 95% identical to Bur_cen, respectively; TLP from *Burkholderia ambifaria* AMMD (MER55697) is 93% identical to Bur_cep; TLP from *Myxococcus xanthus* DK1622 (MER68095) is 55% identical to Fra_sp-1; TLP from *Leptospira interrogans* serovar *Copenhageni* strain Fiocruz L1-130 (MER69997) is 83% identical to Lep_int-2; TLP from *Anabaena variabilis* ATCC29413 (MER54976) is 96% identical to Nos_sp; TLPs from *Protochlamydia amoebophila* UWE25 (MER77521 and MER77544) are 100% identical to Par_sp-1 and Par_sp, respectively; TLP from *Shewanella amazonensis* SB2B (MER75749) is 100% identical to She_ama; TLPs from *Shewanella* sp. MR-7 (MER72768), *Shewanella* sp. MR-4 (MER73030), *Shewanella* sp. ANA-3 (MER73381), and *Shewanella amazonensis* SB2B (MER75808) are 68, 72, 70, and 100% identical to She_ama-1, respectively; TLP from *Nocardioides* sp. JS614 (MER75575) is 61% identical to Str_ave; TLP from *Streptomyces coelicolor* strain A3(2) (MER12295) is 62% identical to Str_exf; TLP from *Xanthomonas campestris* pv. *vesicatoria* strain 85-10 (MER70175) is 94% identical to Xan_axo; and TLP from *Xanthomonas campestris* pv. *vesicatoria* strain 85–10 (MER70193) is 90% identical to Xan_axo-1. At the same time, some new sequences have no close homologs among the analyzed ones, and are briefly described below. TLPs from *Actinoplanes teichomyceticus* (MER55324), *Hahella chejuensis* KCTC 2396 (MER58667), *Mycobacterium vanbaalenii* PYR-1 (MER62807), *Flavobacterium johnsoniae* UW101 (MER62832), and *Myxococcus xanthus* DK 1622 (MER68045 and MER68475) have typical long ATRs. TLPs from *Hahella chejuensis* KCTC 2396 (MER58667) and *Flavobacterium johnsoniae* UW101 (MER62832) have typical long ATRs and additional CTRs. The CTR in TLP from *Hahella chejuensis* KCTC 2396 (160 amino acids) contains one conserved F5/8 type C domain (PF00754) and the CTR in TLP from *Flavobacterium johnsoniae* UW101 (335 amino acids) contains one conserved fibronectin type III domain (PF00041). In addition, the MEROPS database release 7.70 assigned three putative proteins, which are 62–73% identical, deduced from candidatus *Protochlamydia amoebophila* UWE25 genomic sequence (MER77522, MER77524 and MER77537) to the M4 family. Overall, these sequences have no significant similarity to other TLPs (the identity of the full-length precursor and the mature part with thermolysin are about 10 and 16%, respectively) and demonstrate significant homology with other TLPs only in the region of the Zn^{2+}-binding motif (HEXXH_E) typical of many families of Zn^{2+}-dependent metallopeptidases. Finally, the MEROPS database, starting from release 7.50, recognized Col_psy and Neu_cra as non-peptidase homologs of the M4 family

As demonstrated for many TLPs with long ATRs, the proregion acts as an intramolecular chaperone essential for the production of the mature active enzyme [9–12, 29–31]. At the same time, the propeptide is not required to produce at least one catalytically active TLP [32]. The propeptides in TLPs with long ATRs function as inhibitors of the mature protein [12, 13, 15, 29–31]. Apparently, this function is significant in vivo to prevent premature release of the active enzyme, which can be harmful for the cell [33]. In addition, the propeptides can be essential for the secretion of the M4 proteinases with long ATRs [11, 16]. Note also that the propeptides of TLPs with long ATRs are cleaved autocatalytically [30, 31, 34, 35] and intramolecularly [34].

In complete agreement with the data obtained by other investigators, two primary conserved regions were identified in the propeptide of TLPs with long ATRs. These regions (Fig. 2a, b) correspond to amino acids −144…−106 and −47…−16 in the thermolysin sequence. The first one is similar to the fungalysin/thermolysin propeptide (FTP) motif (Protein families database of alignments and HMMs (PFAM) accession number PF07504) [36] and to the hydrophilic region (ProM) found in the middle of the *P. aeruginosa* elastase propeptide [37]. The second one is similar to the more hydrophobic (compared to ProM) region ProC in *P. aeruginosa* elastase propeptide proximal to the propeptide-processing site [37], and is a fragment of the peptidase propeptide and YPEB (PepSY) domain (PFAM accession number PF03413) [38]. Note that, although the FTP motif and PepSY domain are largely found in propeptides of the M4 peptidase family, they were also identified in non-M4 peptidases and many non-peptidase proteins [36]. A conserved Ala was found between these clusters (position −83) in most long ATRs (Fig. 2c). Note that the propeptides of M4 peptidases with long ATRs contain no amino acids conserved in all sequences analyzed. The data on the highly conserved amino acids identified in the prosequences (most of which have been

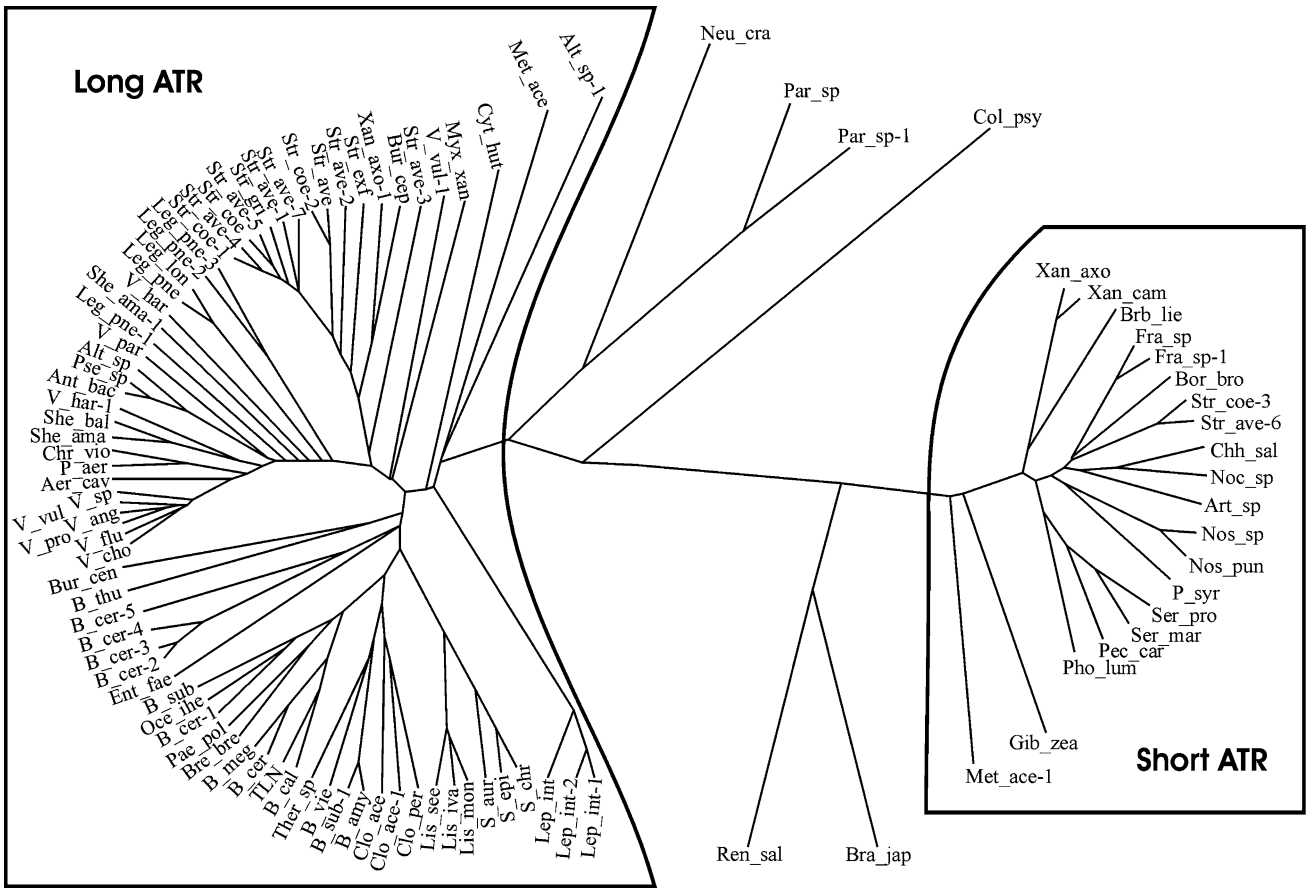

Fig. 1 Dendrogram of the N-terminal amino acid sequences in peptidase precursors of the M4 family

identified by other investigators) are summarized in Table 2. These amino acids are most likely crucial for the functioning of propeptides of TLPs with long ATRs.

The second group of M4 peptidases recognized here from the structure of precursor ATRs includes 20 enzymes with N-terminal regions that are typically 50–60 amino acid long (short ATRs hereafter; Fig. 1). Most proteins of the group are putative, and only three enzymes with short ATRs have been isolated and described. These include the protease of *Pectobacterium carotovorum* (*Erwinia carotovora* subsp. *carotovora*) (Pec_car) [39], minor protease of *Serratia marcescens* (Ser_mar) [40] and protealysin of *Serratia proteamaculans* (Ser_pro) [22]. No data on the functional role of the precursor ATRs are currently available for this group.

The previously studied Pec_car and Ser_mar were considered secretory proteins. When expressed in *E. coli* cells, Ser_mar was shown to accumulate in the extracellular space [40]. Analysis of the precursor ATRs in Pec_car and Ser_mar [39, 40] suggested that these regions contain a typical signal peptide. Indeed, some short ATRs in M4 peptidases contained a structure that, at first glance, resembled signal peptides. However, SignalP 3.0 algorithms failed to identify classical signal peptides in short ATRs. At the same time, a hydrophobic cluster in the region near the initiation methionine was found in the analyzed short ATRs (Fig. 3). This cluster, the PPL-motif, includes seven amino acids, three of which are invariant: two neighboring Pro residues and a Leu two residues downstream of them. An aromatic Tyr or His amino acid immediately follows the Pro–Pro site in most cases. The function of the PPL-motif remains unclear. The PPL-motif can represent a previously unknown cell sorting signal. Noteworthily, this motif can be widespread in nature, since it is found in the N-terminal regions of some putative nonproteolytic proteins, including eukaryotic ones [22].

Of note is a unique feature of the ATR in Gib_zea, among the sequences analyzed. This ATR is the longest (about 100 amino acids) of the short ATRs and contains two PPL-motifs; one is close to the initiation methionine, and the other starts at position 50 from the initiation Met. Comparison of the left and right halves of the Gib_zea ATR demonstrates their 31% identity and 44% similarity, which is largely due to the region (of about 20 amino acids) around the PPL-motif. This can exemplify a duplication in the short ATR of TLP precursors.

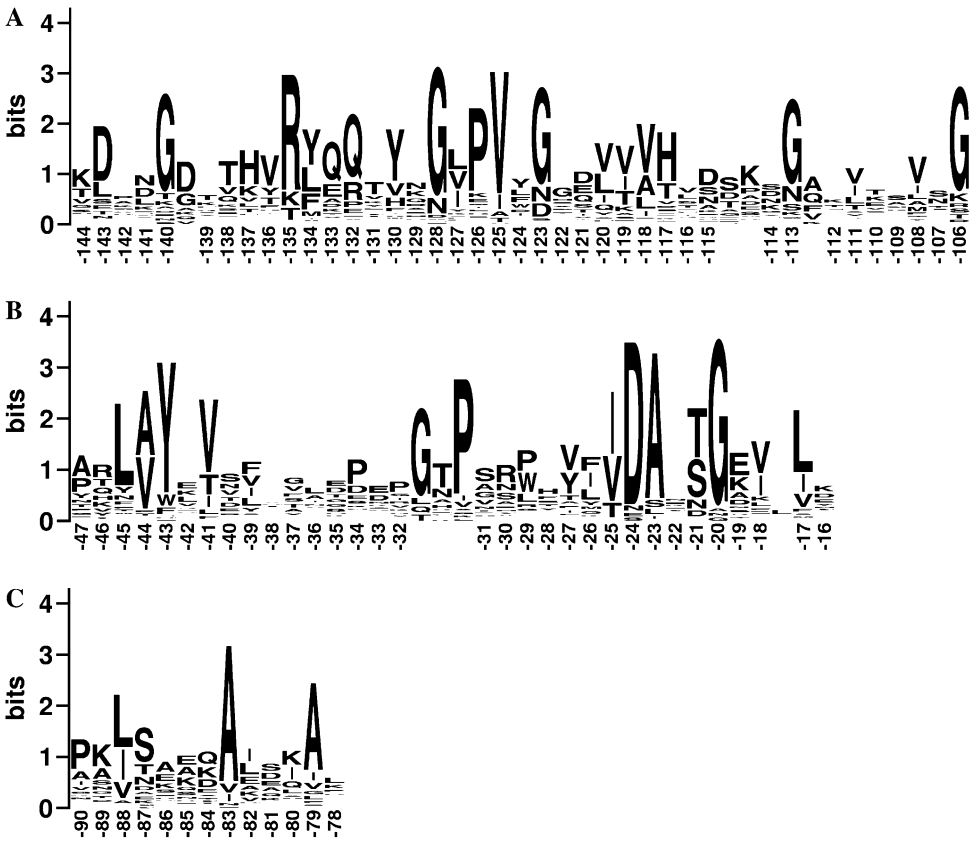

Fig. 2 LOGO presentation of the consensus sequences of the most conserved regions in long ATRs: −144…−106 (related to FTP motif) (**a**), −47…−16 (related to PepSY domain) (**b**), and region around Ala(−83) (**c**)

Previously, we proposed that the enzymes with short ATRs constitute a separate group in the M4 family [22]. Here, we analyzed many sequences of TLP precursors to demonstrate that the M4 family indeed includes two groups of enzymes with a different ATR size and structure. At the same time, no intermediate variants indicating an evolutionary transition from one ATR type to the other have been identified. These ATR structures are not species-specific, and the same species can produce enzymes with both short and long ATRs.

Our analysis also demonstrates the differences in the N-terminal regions of mature M4 proteinases with long and short ATRs of the precursors. Figure 4 demonstrates different distances from the first amino acid to the first unambiguously alignable cluster of the N-terminal regions of mature protealysin relative to thermolysin and *P. aeruginosa* elastase, classical representatives of the M4 family. Thus, the N-terminal part of mature M4 peptidases, precursors of which have short ATRs, is by about 30 amino acids shorter compared to those with long ATRs.

The analyzed set of proteins included six enzymes with ATRs which could not be unambiguously recognized as short or long ATRs (Fig. 1). Visual analysis suggests that the ATRs of Ren_sal and Col_psy precursors resemble long ATRs; Bra_jap and Neu_cra ATRs resemble short ATRs; while Par_sp and Par_sp-1 ATRs resemble neither long nor short ATRs.

Finally, we would like to emphasize an important structural feature of the ATR in M4 precursors. It is common knowledge that TLP prosequences are more tolerant to mutations compared to the mature parts. For instance, the proportion of identical amino acids in ATRs is lower compared to the mature regions (Fig. 5); the majority of modifications in the most conserved amino acids in TLP prosequences (such data are available for long ATRs only) do not completely abolish enzyme activity [37]; TLP propeptides can be replaced with heterologous ones [12]; and long propeptide regions can be replaced with heterologous ones [41] without disturbing protein folding and processing. At the same time, the analysis of mutations in the ATRs and mature regions of the proteins with both short and long ATRs (Fig. 5) demonstrates a concerted accumulation of mutations in the ATRs and mature regions, i.e., they evolved in parallel and no exchange of these domains between different enzymes took place (which is not the case for the precursor CTRs; see below).

One can speculate that this system of variable prosequences serves as a specific evolution module for changing the functional activity of mature proteins in vivo. On the one hand, mutations in prosequences can have an effect on the enzyme accumulation and localization in the cell [37]. On the other hand, there are experimental data demonstrating the direct effect of the changes in propeptides on the catalytic properties of mature proteinases. To date,

Table 2 Conserved amino acids in prosequences of TLPs with long ATRs

Amino acid[a]	Conservation rate[b]	Comments
Gly(−20)	0.92	
Ala(−23)	0.86	The Ala(−23) → Val substitution in *P. aeruginosa* elastase prosequence destabilized the propeptide–enzyme complex expressed in *Pseudomonas putida* cells, while the expression of the mutant in *P. aeruginosa* resulted in a severe growth retardation, leakage of the cells, and cell lysis [33]. The Ala(−23) → Cys substitution completely blocked active elastase accumulation in *P. aeruginosa* culture medium, while proelastase and elastase accumulation in the periplasm remained similar to wild type. Elastase degradation products were detected in the periplasmic fraction [37]
Asp(−24)	0.86	The Asp(−24) → Glu/Ser substitutions in *P. aeruginosa* elastase prosequence had no significant effect [37]
Ile/Val(−25)	0.49/0.36	The Ile(−25) → Val substitution in *P. aeruginosa* elastase prosequence insignificantly affected the accumulation and activity of extracellular elastase. The Ile(−25) → Glu substitution completely blocked active elastase production [37]
Tyr(−43)	0.82	
Ala(−83)	0.84	Identified in this work
Gly(−106)	0.76	
Val(−125)	0.80	The Val(−125) → Ile substitution in *P. aeruginosa* elastase prosequence caused a fivefold decrease in the extracellular activity and decreased the accumulation of extracellular elastase. The Val(−125) → Glu substitution insignificantly decreased the extracellular activity and increased proelastase quantities in the periplasm [37]
Gly(−128)	0.82	The Gly(−128) → Glu substitution in *P. aeruginosa* elastase prosequence caused a twofold decrease in the extracellular activity and decreased the accumulation of extracellular elastase [37]
Arg/Lys(−135)	0.77/0.11	The Arg(−135) → Lys substitution in *P. aeruginosa* elastase prosequence caused a threefold decrease in the extracellular activity and decreased the accumulation of extracellular elastase without affecting the accumulation of elastase and proelastase in the periplasm. The Arg(−135) → Val substitution completely inhibited the extracellular activity and the accumulation of proelastase, elastase, and elastase hydrolytic products in the periplasm [37]

[a] Thermolysin numbering system

[b] Proportion of sequences with a given amino acid among TLPs with long ATRs

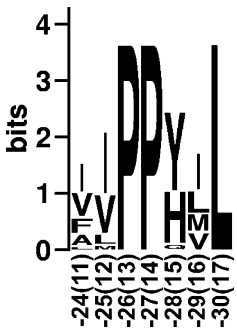

Fig. 3 LOGO presentation of the PPL-motif consensus. Protealysin numbering system, the position relative to the initiation methionine is given in parentheses

there are at least two such examples. First, a point mutation in subtilysin propeptide changes the secondary structure, thermostability, and substrate specificity of the mature protein [42, 43]. Second, cathepsin E with the propeptide from cathepsin D had a different catalytic efficiency and constant of inhibition by protein inhibitors [44].

3.2 Structure and Function of Precursor Carboxy-Terminal Regions

Our analysis of the deduced amino acid sequences of the full-length precursors of M4 peptidases demonstrates an additional C-terminal extension relative to the catalytic part in 37 enzymes from the resulting set of 100 sequences described in Sect. 2 (Table 1). Thirty-three of them are TLPs with long ATRs, three have short ATRs, and one (Col_psy) cannot be assigned to either group (see above).

The length of the precursor CTRs ranges from 110 to 670 amino acids, and their primary structure is even more heterogeneous than in ATRs. The absence of CTRs has been experimentally confirmed for certain mature active proteins [23, 45–50], which allows us to call them C-terminal prosequences. Conversely, other M4 proteinases were isolated from natural hosts exclusively in the form containing

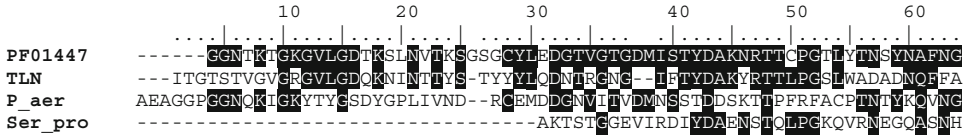

Fig. 4 Alignment of the N-terminal sequences in some TLPs, and the consensus sequence of the conserved peptidase_M4 domain (PF01447). Amino acids identical to PF01447 are shaded. Thermolysin numbering system

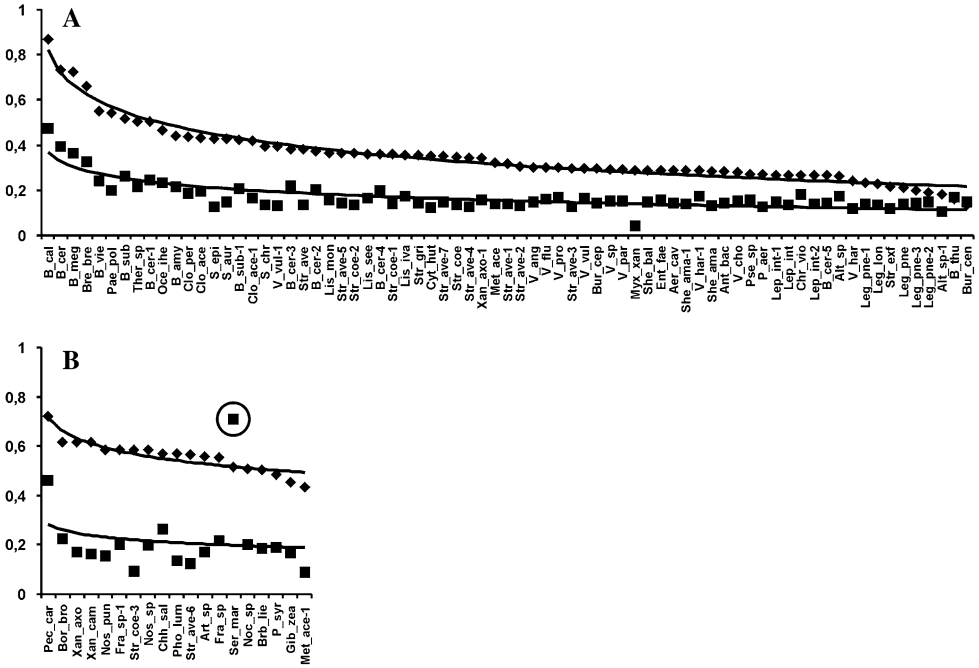

Fig. 5 The proportion of identical amino acids in mature domains (◆) and ATRs (■) of M4 peptidases. (**a**) Enzymes with long ATRs as compared to thermolysin. Pearson's product moment correlation coefficient between mature domains and ATRs was 0.85. (**b**) Enzymes with short ATRs as compared to protealysin. Point mismatch corresponding to Ser_mar ATR (encircled) is due to sequence errors in the mature domain of this enzyme [22]. Pearson's product moment correlation coefficient between mature domains and ATRs was 0.59 (regardless of Ser_mar)

Table 3 Conserved CTR domains in peptidase precursors of the M4 family

Domain name	Designation	PFAM access No.	Total number in analyzed TLPs
Pre-peptidase C-terminal domain	PPC	PF04151	17
Polycystic kidney disease protein PKD1 (polycystin-1) domain	PKD	PF00801	5
Proprotein convertase P-domain	P	PF01483	5
Putative Ig-like domain	He_PIG	PF05345	2
Bacterial Ig-like domain	Big_3	PF07523	1
Gram-positive anchor	GPA	PF00746	1
Clostridial hydrophobic with conserved W (tryptophan) domain	ChW	PF07538	1
Peptidase_M28 domain	peptidase_M28	PF04389	1

CTRs, and were also catalytically active [18, 19, 51–53]. However, thorough investigation of metalloproteinases with CTRs usually demonstrates the presence of both protein forms, with and without CTRs [17–19, 54]. In vitro, recombinant metalloproteinases are also transformed from the CTR-containing to the short form [30, 31, 53, 55], apparently by an autocatalytic mechanism [21, 30, 31]. To summarize the published data, it looks like most M4 peptidases lose their CTR, although the lifetime of the CTR-containing forms varies considerably between different proteins.

The most significant feature of CTRs in TLP precursors is the presence of previously identified conserved domains (Table 3 and Fig. 6) also typical for many other proteins. Such domains have been found in many bacterial and archaeal proteins, as well as mammalian and vertebrate

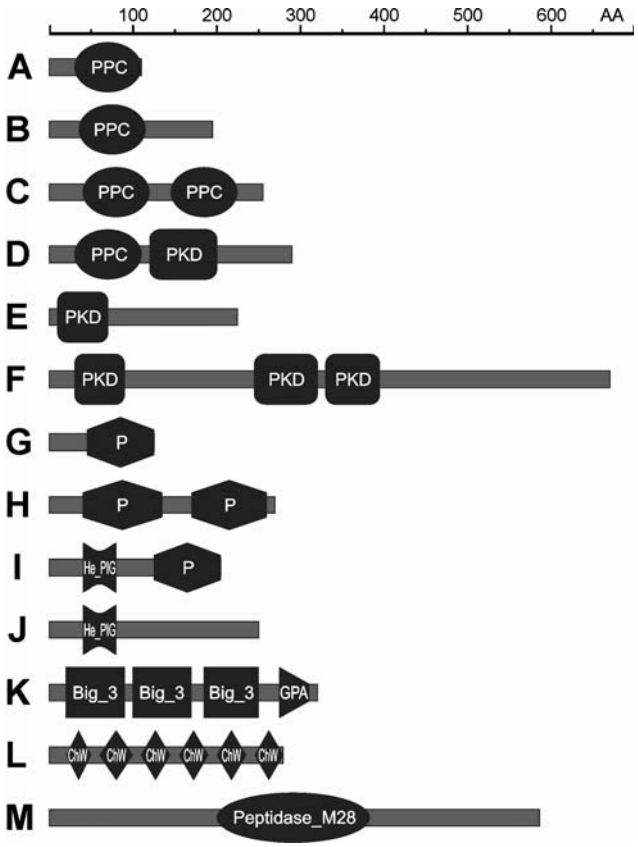

Fig. 6 Architecture of TLP C-terminal regions containing conserved domains. Pattern A is typical of V_cho, V_ang, V_flu, V_sp, V_pro, V_vul, Aer_cav, and Ther_sp; pattern B, V_par and V_har; pattern C, Pse_sp, Alt_sp, Ant_bac, and Myx_xan; pattern D, V_harv-1, She_ama, and She_bal; pattern E, Col_psy; pattern F, Alt_sp-1; pattern G, Str_ave-4, Str_coe-1, and Str_gri; pattern H, Met_ace; pattern I, Str_ave-1; pattern J, Str_ave-7; pattern K, B_cer-2; pattern L, Clo_ace; and pattern M, Str_exf. M4 peptidase designations are given in Table 1. Conserved domain designations are given in Table 3

proteins. The wide distribution of the conserved domains found in CTRs of M4 peptidases suggests domain shuffling between TLPs and proteins of different groups.

Experimental data on the functions of CTRs in M4 peptidases are not abundant. However, the available data suggest that these regions provide for the enzyme binding to insoluble protein and/or polysaccharide substrates. In the case of *Vibrio vulnificus* proteinase (V_vul), the absence of the PPC domain-containing CTR had no effect on the efficiency of hydrolysis of soluble proteins. At the same time, the form without CTR demonstrated an increased rate of hydrolysis of short peptide substrates and, conversely, a considerably decreased capacity to bind and hydrolyze insoluble proteins such as collagen and elastin. In addition, the CTR removal decreased the hemorrhagic activity in vivo [20, 21]. Similar data were obtained for *Vibrio fluvialis* proteinase (V_flu), which has a PPC domain in its CTR, too. In contrast to the CTR-containing form, and similar to V_vul, the CTR-less proteinase demonstrated insignificant activity towards insoluble elastin and was unable to agglutinate rabbit erythrocytes [53]. The CTR of *Myxococcus xanthus* metalloproteinase, with two PPC domains, proved to be bound to the extracellular matrix of this bacterium in vivo [17]. At the same time, the extracellular matrix of *M. xanthus* is arranged as fibrils composed of a carbohydrate backbone with associated proteins [56]. The isolated CTR of thermolysin-like metalloproteinase, a component of the chitinolytic complex in marine bacterium *Alteromonas* sp. (Alt_sp-1), contains three PKD domains and was shown to bind cellulose, chitosan, as well as α- and β-chitins [18, 19].

The conclusion that the CTR in M4 peptidases is a substrate-binding module is confirmed by the data on the properties of conserved domains in other enzyme groups. For instance, two C-terminal PPC domains in class I collagenase from *Clostridium histolyticum* (Col G) can bind different collagen types [57, 58]. According to the PFAM database, most PKD domains are found in the extracellular parts of proteins interacting with other proteins or polysaccharides. Substrate binding is one of the proposed functions of P-domains of proprotein convertases [59–61], although these elements also mediate proper cellular localization of the enzymes [60–64], proprotein convertase stability, and, possibly, folding of the catalytic domain and processing [64–67].

The CTR in M4 peptidases can also mediate cellular localization of the enzymes. For instance, the CTR in B_cer-2 includes conserved domains typical of a variety of bacterial surface proteins (Fig. 6), suggesting that the CTRs are responsible for the B_cer-2 localization on the bacterial cell surface.

Unusually, the catalytic domain peptidase_M28 is found in the CTR of M4 peptidase from *Streptomyces exfoliatus* (Str_exf). The M28 peptidase family includes amino- and carboxypeptidases from Bacteria, Archaea, Protozoa, Fungi, plants, and animals. The functional role of the combination of two peptidases is not known.

Thus, the data on the functions of TLP CTRs and their conserved domains suggest that most CTRs are modules binding high molecular weight insoluble substrates. At the same time, CTRs are absent from many mature M4 peptidases. What can be the function of such removed substrate-binding modules? The above-mentioned data for V_vul and V_flu demonstrate that CTR-less proteins better hydrolyze low molecular weight soluble substrates, while CTR-containing enzymes better hydrolyze high molecular weight insoluble ones [21, 53]. In this context, the following scenario can be proposed: a CTR-containing enzyme binds an insoluble substrate and starts its hydrolysis. Later, the enzyme is released from the CTR anchor and efficiently hydrolyzes the resulting low molecular weight products.

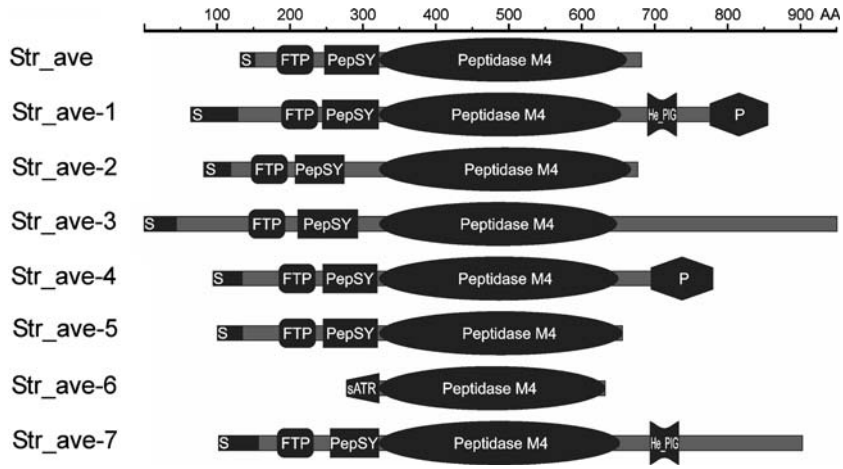

Fig. 7 Architecture of peptidase precursors of the M4 family in *Streptomyces avermitilis* MA-4680. Str_ave, Str_ave-1, Str_ave-2, Str_ave-4, Str_ave-5, and Str_ave-7 have 61-79% identity between the catalytic parts. Str_ave-1, Str_ave-4, and Str_ave-7 have additional CTRs with seemingly low similarity. However, thorough analysis of these CTR sequences demonstrates that Str_ave-4 and Str_ave-1 P-domains have 53% identity, while Str_ave-1 and Str_ave-7 He_PIG domains have 63% identity. Str_ave-3 significantly differs from other TLPs in *S. avermitilis* MA-4680. The similarity of the catalytic part with other *S. avermitilis* TLPs with long ATRs is low (about 20% identity) and the CTR of Str_ave-3 contains no known conserved domains. S, signal peptide; FTP, fungalysin/thermolysin propeptide motif (PF07504); PEPSY, peptidase propeptide and YPEB domain (PF03413); sATR, short ATR of TLPs; peptidase M4, TLP catalytic region including Peptidase_M4 (PF01447) and Peptidase_M4_C (PF02868) domains; P, proprotein convertase P-domain (PF01483); and He_PIG, putative Ig-like domain (PF05345)

3.3 TLPs Encoded within the Same Genome

Analysis of the available data on the genome structure indicates that some genomes include several peptidases of the M4 family (Table 1). In many cases, the genes in the same species code for TLPs with different precursor structures. Let us consider the most vivid example.

The highest number of genes of M4 proteinases is found in the genome of *S. avermitilis* MA-4680. One of eight TLPs in this species has a short ATR, while the other seven proteins have long ATRs and four of them have CTRs (Fig. 7). Analysis of their amino acid sequences demonstrated the general pattern observed for the whole M4 family. The precursor ATRs have a lower similarity compared to the mature parts, while the similarity level of the ATRs and mature parts between different TLPs from *S. avermitilis* correlated. The CTR structure of TLPs from *S. avermitilis* MA-4680 also supports the concept of domain shuffling.

4 Conclusions

In summary, this analysis has revealed the following significant facts concerning the structure of peptidase precursors of the M4 family:

- no precursor of M4 peptidases without amino-terminal regions (ATRs) in addition to the catalytic domain has been found;
- there are two ATR types: short and long ATRs of about 50 and 200 amino acids in length, respectively;
- long ATRs contain no amino acids conserved in all M4 peptidases;
- no classical signal peptides have been identified in short ATRs, but short ATRs proved to contain conserved a PPL-motif near the initiation methionine;
- the accumulation of mutations in ATRs of both types correlates with that in the catalytic domains;
- about one-third of TLP precursors have C-terminal extensions (CTRs); they are found in about half of precursors with long ATRs, but occur more rarely in precursors with short ATRs; CTRs contain previously identified conserved domains typical of many other proteins, too, and likely underlie the interaction with high molecular weight substrates.

Acknowledgments This work was supported by the Russian Foundation for Basic Research (project nos. 06-04-48678, 06-04-48690, and 06-04-08123).

References

1. Rawlings ND, Barrett AJ (1995) Methods Enzymol 248:183–228
2. Rawlings ND, Morton FR, Barrett AJ (2006) Nucleic Acids Res 34:D270–272
3. Matthews BW (1988) Acc Chem Res 21:333–340
4. de Kreij A, Venema G, van den Burg B (2000) J Biol Chem 275:31115–31120
5. Feder J (1967) Biochemistry 6:2088–2093

6. Feder J, Schuck JM (1970) Biochemistry 9:2784–2791
7. Schechter I, Berger A (1967) Biochem Biophys Res Commun 27:157–162
8. Shinde U, Inouye M (2000) Semin Cell Dev Biol 11:35–44
9. Braun P, Tommassen J, Filloux A (1996) Mol Microbiol 19:297–306
10. Marie-Claire C, Ruffet E, Beaumont A, Roques BP (1999) J Mol Biol 285:1911–1915
11. McIver KS, Kessler E, Olson JC, Ohman DE (1995) Mol Microbiol 18:877–889
12. Tang B, Nirasawa S, Kitaoka M, Marie-Claire C, Hayashi K (2003) Biochem Biophys Res Commun 301:1093–1098
13. Kessler E, Safrin M (1994) J Biol Chem 269:22726–22731
14. O'Donohue MJ, Roques BP, Beaumont A (1994) Biochem J 300(Pt 2):599–603
15. Serkina AV, Gorozhankina TF, Shevelev AB, Chestukhina GG (1999) FEBS Lett 456:215–219
16. Wetmore DR, Wong SL, Roche RS (1992) Mol Microbiol 6:1593–1604
17. Kearns DB, Bonner PJ, Smith DR, Shimkets LJ (2002) J Bacteriol 184:1678–1684
18. Miyamoto K, Nukui E, Hirose M, Nagai F, Sato T, Inamori Y, Tsujibo H (2002) Appl Environ Microbiol 68:5563–5570
19. Miyamoto K, Tsujibo H, Nukui E, Itoh H, Kaidzu Y, Inamori Y (2002) Biosci Biotechnol Biochem 66:416–421
20. Miyoshi S, Kawata K, Tomochika K, Shinoda S, Yamamoto S (2001) Toxicon 39:1883–1886
21. Miyoshi S, Wakae H, Tomochika K, Shinoda S (1997) J Bacteriol 179:7606–7609
22. Demidyuk IV, Kalashnikov AE, Gromova TY, Gasanov EV, Safina DR, Zabolotskaya MV, Rudenskaya GN, Kostrov SV (2006) Protein Expr Purif 47:551–561
23. Zabolotskaya MV, Demidyuk IV, Akimkina TV, Kostrov SV (2004) Protein J 23:483–492
24. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) Nucleic Acids Res 25:4876–4882
25. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) J Mol Biol 340:783–795
26. Schneider TD, Stephens RM (1990) Nucleic Acids Res 18:6097–6100
27. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) Genome Res 14:1188–1190
28. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle
29. O'Donohue MJ, Beaumont A (1996) J Biol Chem 271:26477–26481
30. Tang B, Nirasawa S, Kitaoka M, Hayashi K (2002) Biochim Biophys Acta 1596:16–27
31. Tang B, Nirasawa S, Kitaoka M, Hayashi K (2002) Biochem Biophys Res Commun 296:78–84
32. Mansfeld J, Petermann E, Durrschmidt P, Ulbrich-Hofmann R (2005) Protein Expr Purif 39:219–228
33. Braun P, Bitter W, Tommassen J (2000) Microbiology 146(Pt 10):2565–2572
34. Marie-Claire C, Roques BP, Beaumont A (1998) J Biol Chem 273:5697–5701
35. McIver K, Kessler E, Ohman DE (1991) J Bacteriol 173:7781–7789
36. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) Nucleic Acids Res 32:D138–141
37. McIver KS, Kessler E, Ohman DE (2004) Microbiology 150:3969–3977
38. Yeats C, Rawlings ND, Bateman A (2004) Trends Biochem Sci 29:169–172
39. Kyostio SR, Cramer CL, Lacy GH (1991) J Bacteriol 173:6537–6546
40. Kwon YT, Lee HH, Rho HM (1993) Gene 125:75–80
41. Novikova SI, Serkina AV, Konstantinova GE, Khlebalina OI, Chestukhina GG, Shevelev AB (2001) Vopr Med Khim 47:123–131
42. Shinde U, Fu X, Inouye M (1999) J Biol Chem 274:15615–15621
43. Shinde UP, Liu JJ, Inouye M (1997) Nature 389:520–522
44. Yasuda Y, Tsukuba T, Okamoto K, Kadowaki T, Yamamoto K (2005) J Biochem (Tokyo) 138:621–630
45. Hase CC, Finkelstein RA (1990) Infect Immun 58:4011–4015
46. Kato JY, Suzuki A, Yamazaki H, Ohnishi Y, Horinouchi S (2002) J Bacteriol 184:6016–6025
47. Lee SO, Kato J, Nakashima K, Kuroda A, Ikeda T, Takiguchi N, Ohtake H (2002) Biosci Biotechnol Biochem 66:1366–1369
48. Nirasawa S, Nakajima Y, Zhang ZZ, Yoshida M, Hayashi K (1999) Biochem J 341(Pt 1):25–31
49. Norqvist A, Norrman B, Wolf-Watz H (1990) Infect Immun 58:3731–3736
50. Oda K, Okayama K, Okutomi K, Shimada M, Sato R, Takahashi S (1996) Biosci Biotechnol Biochem 60:463–467
51. Chuang YC, Chang TM, Chang MC (1997) Gene 189:163–168
52. Miyoshi N, Shimizu C, Miyoshi S, Shinoda S (1987) Microbiol Immunol 31:13–25
53. Miyoshi S, Sonoda Y, Wakiyama H, Rahman MM, Tomochika K, Shinoda S, Yamamoto S, Tobe K (2002) Microb Pathog 33:127–134
54. David VA, Deutch AH, Sloma A, Pawlyk D, Ally A, Durham DR (1992) Gene 112:107–112
55. Teo JW, Zhang LH, Poh CL (2003) Gene 303:147–156
56. Behmlander RM, Dworkin M (1994) J Bacteriol 176:6295–6303
57. Matsushita O, Koide T, Kobayashi R, Nagata K, Okabe A (2001) J Biol Chem 276:8761–8770
58. Toyoshima T, Matsushita O, Minami J, Nishi N, Okabe A, Itano T (2001) Connect Tissue Res 42:281–290
59. Creemers JW, Siezen RJ, Roebroek AJ, Ayoubi TA, Huylebroeck D, Van de Ven WJ (1993) J Biol Chem 268:21826–21834
60. Gluschankof P, Fuller RS (1994) EMBO J 13:2280–2288
61. Rovere C, Luis J, Lissitzky JC, Basak A, Marvaldi J, Chretien M, Seidah NG (1999) J Biol Chem 274:12461–12467
62. Creemers JW, Usac EF, Bright NA, Van de Loo JW, Jansen E, Van de Ven WJ, Hutton JC (1996) J Biol Chem 271:25284–25291
63. Lusson J, Benjannet S, Hamelin J, Savaria D, Chretien M, Seidah NG (1997) Biochem J 326(Pt 3):737–744
64. Taylor NA, Shennan KI, Cutler DF, Docherty K (1997) Biochem J 321(Pt 2):367–373
65. Ueda K, Lipkind GM, Zhou A, Zhu X, Kuznetsov A, Philipson L, Gardner P, Zhang C, Steiner DF (2003) Proc Natl Acad Sci USA 100:5622–5627
66. Zhou A, Martin S, Lipkind G, LaMendola J, Steiner DF (1998) J Biol Chem 273:11107–11114
67. Zhu X, Muller L, Mains RE, Lindberg I (1998) J Biol Chem 273:1158–1164