Deposited research article

The rhomboids: a near ubiquitous family of intramembrane serine proteases evolved via multiple horizontal gene transfers

Eugene V Koonin[1], Kira S Makarova[1], Laetitia Davidovic[2] and Luca Pellegrini[2]

Addresses: [1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. [2]Laboratory of molecular neurobiology, Centre de Recherche Université Laval Robert Giffard, Université Laval, Quebec, Canada.

Correspondence: Luca Pellegrini. E-mail: Luca.Pellegrini@crulrg.ulaval.ca

Posted: 3 October 2002

Genome ***Biology*** 2002, **3(11)**:preprint0010.1-0010.26

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/11/preprint/0010

Received: 30 September 2002

This is the first version of this article to be made available publicly. A peer-reviewed and modified version is now available in full at http://genomebiology.com/2003/4/3/R19

The rhomboids: a near ubiquitous family of intramembrane serine proteases evolved via multiple horizontal gene transfers

Eugene V. Koonin[1], Kira S. Makarova[1], Laetitia Davidovic[2], and Luca Pellegrini[2,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; [2,*] Laboratory of molecular neurobiology, Centre de Recherche Université Laval Robert Giffard, Université Laval, Quebec, Canada

[*]To whom correspondence should be addressed:
CRULRG, Rm F-5519, 2601 Chemin de la Canardiere, G1J 2G3 Quebec, QC, Canada
Tel. 1 418 663 5000 ext 6879
Fax 1 418 663 5775
Email: Luca.Pellegrini@crulrg.ulaval.ca

Abstract

Background. The rhomboid family consists of polytopic membrane proteins, which show a level of evolutionary conservation that is unique among membrane proteins. The rhomboids are present in nearly all sequenced genomes of archaea, bacteria and eukaryotes, with the exception of several species with small genomes. On the basis of experimental studies with the developmental regulator Rhomboid from Drosophila and the AarA protein from the bacterium *Providencia stuartii*, the rhomboids are thought to be intramembrane serine proteases whose signaling function is conserved in eukaryotes and prokaryotes.

Results. Phylogenetic tree analysis suggests that, despite the broad distribution in all three kingdoms of life, the rhomboid family was not present in the last universal common ancestor of the extant life forms, but instead evolved in bacteria and has been acquired by archaea and eukaryotes via several independent horizontal gene transfer events. In eukaryotes, two distinct, ancient horizontal acquisitions apparently gave rise to the two major subfamilies typified by Rhomboid and PARL (presenilin-associated Rhomboid-

like protein), respectively. The subsequent evolution of the rhomboid family in eukaryotes proceeded via multiple duplications and functional diversification through the addition of extra transmembrane helices and other domains in different orientations relative to the conserved core that harbors the protease activity.

Conclusions. Although the near universal presence of the rhomboid family in bacteria, archaea and eukaryotes appears to suggest that this protein is part of the heritage of the last universal common ancestor, phylogenetic tree analysis indicates bacterial origin with subsequent dissemination via horizontal gene transfer. This emphasizes the importance of explicit phylogenetic analysis for the reconstruction of ancestral life forms. A hypothetical scenario of origin of intracellular membrane proteases from membrane transporters is proposed.

Background

Polytopic transmembrane proteins are, in general, not particularly strongly conserved during evolution. Inspection of the database of Clusters of Orthologous Groups of proteins (COGs) [1] revealed only one family of such proteins that is represented in most of the sequenced bacterial, archaeal and eukaryotic genomes. The prototype of this family is the Rhomboid (RHO) protein from *Drosophila melanogaster*, a developmental regulator involved in epidermal growth factor (EGF)-dependent signaling pathways [2-4]. Not only were homologs of Rhomboid detected in prokaryotes and eukaryotes, but the pattern of sequence conservation in this family appeared uncharacteristic of non-enzymatic membrane proteins, such as transporters [5, 6]. Specifically, several polar amino acid residues are conserved in nearly all members of the Rhomboid family, suggesting the possibility of an enzymatic activity. Since three of these conserved residues were histidines, it has been hypothesized that rhomboid family proteins could function as metal-dependent membrane proteases [5, 6]. Recently, however, it has been shown that RHO cleaves a transmembrane helix (TMH) in the membrane-bound precursor of the TGFa-like growth factor Spitz, enabling the released Spitz to activate the EGF receptor, and that a conserved serine and a conserved histidine in RHO are essential for this cleavage [7, 8]. Thus, it appears that Rhomboid family proteins are a distinct group of intramembrane serine proteases. Altogether, the genome of Drosophila encodes 7 RHO paralogs (now designated RHO1-7, with the original Rhomboid becoming RHO-1), at least three of which are involved in distinct EGF-dependent pathways, apparently through proteolytic activation of diverse ligands of the EGF receptor [9, 10] .

The newly discovered intramembrane proteolytic activity of RHO places the rhomboid family within the framework of regulated intramembrane proteolysis (RIP), a new paradigm of signal transduction, which appears to be prominent in all forms of life [11, 12]. Under RIP, signaling proteins undergo site-specific proteolysis within TMH, resulting in the release of active fragments, which are the actual effectors in signal tranduction cascades. Until recently, the only characterized cases of RIP in eukaryotes involved presenilin, an aspartyl protease, which cleaves a transmembrane helix in type I membrane proteins, such as amyloid precursor protein (APP), Notch, and Ire1 [13], and S2P, a metalloprotease, which cleaves a TMH in a type 2 transmembrane protein, the sterol-dependent transcription factor SREBP [11]. Notably, S2P has highly conserved bacterial homologs, and the protease domain of presenilin also might be homologous to bacterial and archaeal type IV prepilin peptidases, although, in this case, the sequence similarity is very low [14, 15].

In the case of the rhomboid family, the existence of homologs of RHO in most prokaryotes is particularly remarkable because animal RHO proteins are involved in signaling pathways that are not found outside metazoa, which seems to make functional conservation in prokaryotes a remote possibility. The only prokaryotic protein of the rhomboid family that has been characterized experimentally in considerable detail is AarA from the bacterium *Providencia stuartii* [16, 17]. This protein is involved in the export of a quorum-sensing peptide, a function that, in physiological terms, resembles that of RHO, although the signaling molecules, other than RHO and AarA, are obviously unrelated [18]. In a striking recent development, two independent research groups have shown that several bacterial rhomboid family proteins, including AarA, were capable of cleaving the EGFr receptor ligands (Spitz, Keren and Gurken) that are normally cleaved

by RHO paralogs [19, 20]. The cleavage depended on the conserved serine and histidine residues paralogs [19] and, moreover, transgenic flies that expressed AarA developed a phenotype indistinguishable from that induced by overexpression of RHO, whereas RHO could substitute for AarA in *Providencia stuartii* [20]. These unexpected findings demonstrated the conservation of a RIP mechanism producing extracellular signals in eukaryotes and prokaryotes.

The near ubiquity of rhomboid family proteins among bacteria, archaea and eukaryotes, along with the remarkable functional conservation, suggest that a signaling mechanism mediated by rhomboids might have functioned already in the last common ancestor of all extant life forms, with subsequent loss in several lineages. To address this possibility, we performed a detailed phylogenetic analysis of the rhomboid family.

Results and Discussion

Sequence and structural features and phyletic distribution and of the rhomboid family

Although the sequence similarity between eukaryotic and prokaryotic rhomboid family proteins is relatively low (at the level of 15-10% identity in the conserved region), the entire superfamily could be retrieved from the protein sequence databases within three iterations of the PSI-BLAST program with a high statistical significance and without any false positives. The conserved core of the rhomboid family consists of six conserved TMH (Fig. 1). The predicted catalytic serine is located in TMH5, whereas the predicted catalytic histidine is in TMH7; TMH3 contains two additional histidines and an asparagine, which are conserved in the great majority of the rhomboid family proteins (Fig. 1). The roles of these conserved residues are not known, but, given the remarkable

evolutionary conservation, it seems likely that they also contribute to catalysis; indeed, it has been shown that the conserved asparagines is required for the cleavage of Spitz by RHO.

When examining the multiple alignment of the Rhomboid superfamily proteins, we noticed that several eukaryotic members appear to be inactivated proteases as indicated by the loss of the predicted catalytic serine or histidine (Fig. 1 and data not shown); these inactivated forms could be regulators of active rhomboid proteases. Several other proteins lack one or more of the conserved residues in TMH3; it remains unclear whether or not these are active proteases.

Bacterial and archaeal members of the Rhomboid superfamily contain 6 TMH, whereas the eukaryotic members typically have an additional, 7th TMH, which may be attached to the core either form the N-terminus or from the C-terminus as discussed below.

The phyletic distribution pattern of the rhomboid family shows that this intramembrane protease is extremely common in all three kingdoms of life, but is not necessarily essential for cell function. Rhomboids are missing in the microsporidium *Encephalitozoon cuniculi*, a eukaryotic intracellular parasite with a highly degraded genome, the archaea *Methanothermobacter thermoautotrophicus* and *Thermoplasma volcanium*, and several bacterial species, primarily parasites with small genomes but also species with moderate-size genomes, such as *Xylella fastidiosum* (see COG0705 at http://www.ncbi.nlm.nih.gov/COG/). On two occasions, a representative of the rhomboid family is present in only one of a pair of relatively close genomes (present in *T. acidophilum* but missing in T. volcanium; present in the spirochete *Treponema pallidum*

but missing in the related bacterium *Borrelia burgdorferi*), which suggests relatively recent, repeated losses of this gene. Most of the prokaryotic species encode a single gene coding for a rhomboid family protein, although some have two-three paralogs (see COG0705 at http://www.ncbi.nlm.nih.gov/COG/); in contrast, eukaryotes show expansion of the rhomboid family, with 7 members in Drosophila, and as many as 13 in Arabidopsis.

Phylogeny and evolutionary history of the rhomboid family

The multiple alignment of the 6-TMH core of the rhomboid family (Fig. 1) was employed to construct a phylogenetic tree using the least-square algorithm with subsequent optimization using the maximum likelihood method (see Materials and Methods). Only the conserved regions including the TMH and short adjacent stretches shown in Figure 1 were used as the input for tree building, whereas the poorly conserved intervening regions were omitted to avoid the noise from potentially misaligned residues. The resulting phylogenetic tree of the rhomboid family presents a complex and unexpected picture (Fig. 2). Neither the eukaryotic nor the archaeal subsets of the family appear to form monophyletic clades. Instead, the eukaryotic rhomboids are split between two major subfamilies, which are positioned in the midst of different prokaryotic branches (Fig. 2). The first subfamily, which includes 6 of the 7 Drosophila rhomboids, clusters with a distinct prokaryotic assemblage, which consists primarily of Gram-positive bacteria as well as a subset of archaeal rhomboids; this clade is strongly supported by bootstrap analysis (Fig. 2). The proteins in this group of eukaryotic rhomboids, which we designated the RHO-subfamily, typically have an extra TMH

added C-terminally of the 6-TMH core; some of these proteins also contain EF-hand Ca-binding domains N-terminally of the core (Fig. 2).

The second eukaryotic subfamily, which we designated the PARL-subfamily, after presenilin-associated rhomboid-like protein (PARL), the human ortholog of Drosophila Rhomboid 7 [6], resides within a large, heterogeneous prokaryotic cluster (Fig. 2). Within this subfamily, PARL and its orthologs from other animals and fungi, have a distinct domain architecture, with an extra TMH added to the N-terminus of the core, whereas the rest have only the core (a C-terminal TMH and a ubiquitin-associated domain are appended in one Arabidopsis protein; Fig. 2). Thus, the existence of two distinct subfamilies of eukaryotic rhomboids is supported by features of domain architectures that appear to comprise shared derived characters. Within these two major eukaryotic subfamilies, evolution apparently proceeded via both ancient and more recent duplications. Several lineage-specific expansions of paralogs [21] are noticeable, in insects, mammals and plants (Fig. 2).

Archaeal rhomboids are scattered over the phylogenetic tree, with two major clusters and three more isolated proteins joining different bacterial branches (Fig. 2). There is no indication of an affinity between any of the archaeal and eukaryotic rhomboids. Although many of the bacterial rhomboids form phylogenetically coherent clusters corresponding to the established bacterial lineages, there are also several clusters that have odd composition, such as grouping of proteobacterial and Gram-positive species; some of these clusters are well supported by bootstrap (see clusters 1-4 in Fig. 2).

The phylogenetic tree of the rhomboid family tree shown in Figure 2 clearly follows neither the "standard model" scenario [22, 23], with the major split between the archaeo-eukaryotic and bacterial lineages nor the "mitochondrial" scenario, which postulates acquisition of a gene by eukaryotes from the pro-mitochondrial endosymbiont. Neither can this tree be explained by postulating a small number of lineage-specific gene losses. The parsimonious interpretation of the rhomboid family seems to be that the evolutionary history of this family had been replete with horizontal gene transfer (HGT) and lineage-specific gene loss events. In particular, in spite of the presence of rhomboids in the majority of modern life forms from all three primary kingdoms, phylogenetic analysis suggests that this family had not been inherited from LUCA. Instead, the tree topology seems to indicate that this family emerged in some bacterial lineage and afterwards had been widely disseminated via HGT, and then lost in some lineages. Both archaea and eukaryotes seems to have acquired rhomboids on several independent occasions. In particular, at least two HGT events seem to have contributed to the origin of eukaryotic rhomboids, one of them yielding the RHO-subfamily and the other one the PARL-subfamily, with a possible additional HGT in plants (Fig. 2). Given the broad phyletic representation of both subfamilies of eukaryotic rhomboids, both the RHO-subfamily and the PARL-subfamily must have been acquired via HGT at an early stage of eukaryotic evolution, definitely before the divergence of the major crown-group lineages. This early epoch in eukaryotic evolution is thought to have been dominated by HGT from multiple bacterial symbionts [24, 25].

Two alternatives to this multiple-HGT scenario may be considered. One of them would postulate that LUCA already had multiple, paralogous rhomboids, which evolved

via a series of ancient gene duplications, and the odd topology of the phylogenetic tree is due primarily to differential loss of these ancient paralogs. Although this cannot be ruled formally, this hypothesis implies the existence of extremely elaborate signaling system in LUCA, which is hardly compatible with the existing general notions regarding this primitive life form. The second possibility is that the topology of the tree in Figure 2 is simply random. However, the strong bootstrap support for many nodes and the presence of several phylogenetically coherent clusters (above all, the RHO and PARL subfamilies in eukaryotes, but also some of the archaeal and bacterial clusters) seem to argue against this explanation.

The multiple-HGT interpretation of the evolutionary history of the rhomboid family is, at least at first glance, distinctly counter-intuitive, given that this family is nearly ubiquitous among the extant life forms. Indeed, when attempts are made to construct parsimonious evolutionary scenarios on the basis of phyletic patterns [25, 26], there is no chance that such a widespread family is not assigned to LUCA. It should be realized, however, that these approaches are inherently probabilistic and extensive HGT can fool them. For the rhomboid family, this mode of evolution seems to be particularly plausible (Fig. 3). It seems likely that the ultimate ancestor of the rhomboid family evolved from a non-enzymatic integral membrane protein, probably a transporter that might have been involved in an early, primitive form of export of signaling peptides in bacteria. The protease active center might have evolved in such a transporter by chance emergence of the suitable catalytic amino acid within two or three of the TMH (Fig. 3). This would enable the transition from simple transport to the RIP mode of controlled export of signaling molecules. Emergence of RIP could have conferred a major selective

advantage on the respective bacteria and might have resulted in an evolutionary sweep whereby the gene carrying this trait had been repeatedly fixed, rather than eliminated, after HGT. In terms of the evolution of sequence itself, the requirements for the conservation of the protease activity apparently "locked" the rhomboid family in the regime of relatively slow evolution, which ensures the significant sequence similarity among all family members (Fig. 1). The scenario of origin from non-catalytic transporters might potentially apply to other integral membrane enzymes, including intramembrane proteases involved in RIP, such as presenilins and their homologs [14, 15] and the archaeo-eukaryotic signal peptide peptidase [27].

Conclusions

The rhomboid family may be the most widespread and conserved group of integral membrane proteins. In and by itself, this would suggest that this family is part of the gene repertoire of LUCA. However, phylogenetic analysis strongly suggests a different scenario, one of emergence in a bacterial lineage with subsequent multiple independent HGT events and gene losses. In particular, eukaryotes probably acquired their two major rhomboid subfamilies, RHO and PARL, as the result of two independent, early HGT events. These events introducing RIP as a means of intercellular communication might have been pivotal in the evolution of eukaryotic multicellularity along the lines discussed previously with regard to the apparent bacterial origin of key components of eukaryotic programmed cell death machinery [28]. Subsequent evolution of rhomboids in eukaryotes proceeded via lineage-specific expansion of paralogs [21], followed by diversification through the addition of an extra TMH in different positions

relative to the catalytic core, some limited domain accretion (Fig. 2), and sequence divergence.

Phylogenetic analysis of the rhomboid family described here carries a general message for studies aimed at the reconstruction of ancestral life forms, particularly LUCA. Although most of the (nearly) ubiquitous protein families probably do derive from LUCA, explicit phylogenetic analysis is required to ascertain this in each individual case.

Material and Methods

The non-redundant (NR) protein sequence database at the National Center for Biotechnology Information (NIH, Bethesda) was searched iteratively using the PSI-BLAST program with multiple starting queries [29]. PSI-BLAST was normally run with expectation (E) value of 0.01 as the cut-off for inclusion of sequences into the position-specific scoring matrix. Multiple alignments of protein sequences were constructed using the ClustalW program [30] and manually adjusted on the basis of the examination of PSI-BLAST search outputs and the superposition of the predicted transmembrane helices. Transmembrane helices were predicted using the programs TMpred[31] and TMAP[32].

Phylogenetic trees were built using the least-square method [33] implemented in the FITCH program of the PHYLIP package [34], with subsequent local rearrangement using the PROTML program of the MOLPHY package to obtain the maximum likelihood tree [35]. The reliability of the tree topology was assessed using the RELL bootstrap method of MOLPHY, with 10000 replications [36].

Acknowledgments

L.P. is supported by a grant from NSERC Canada.

References

1. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes**. *Nucleic Acids Res* 2001, **29**:22-28.

2. Sturtevant MA, Roark M, Bier E: **The Drosophila rhomboid gene mediates the localized formation of wing veins and interacts genetically with components of the EGF-R signaling pathway**. *Genes Dev* 1993, **7**:961-973.

3. Sturtevant MA, Roark M, O'Neill JW, Biehs B, Colley N, Bier E: **The Drosophila rhomboid protein is concentrated in patches at the apical cell surface**. *Dev Biol* 1996, **174**:298-309.

4. Guichard A, Biehs B, Sturtevant MA, Wickline L, Chacko J, Howard K, Bier E: **rhomboid and Star interact synergistically to promote EGFR/MAPK signaling during Drosophila wing vein development**. *Development* 1999, **126**:2663-2676.

5. Mushegian AR, Koonin EV: **Sequence analysis of eukaryotic developmental proteins: ancient and novel domains**. *Genetics* 1996, **144**:817-828.

6. Pellegrini L, Passer BJ, Canelles M, Lefterov I, Ganjei JK, Fowlkes BJ, Koonin EV, D'Adamio L: **PAMP and PARL, two novel putative metalloproteases interacting with the COOH-terminus of Presenilin-1 and -2**. *J Alzheimers Dis* 2001, **3**:181-190.

7. Urban S, Lee JR, Freeman M: **Drosophila rhomboid-1 defines a family of putative intramembrane serine proteases**. *Cell* 2001, **107**:173-182.

8. Klambt C: **EGF receptor signalling: roles of star and rhomboid revealed**. *Curr Biol* 2002, **12**:R21-23.

9. Guichard A, Roark M, Ronshaugen M, Bier E: **brother of rhomboid, a rhomboid-related gene expressed during early Drosophila oogenesis, promotes EGF-R/MAPK signaling**. *Dev Biol* 2000, **226**:255-266.

10. Wasserman JD, Urban S, Freeman M: **A family of rhomboid-like genes: Drosophila rhomboid-1 and roughoid/rhomboid-3 cooperate to activate EGF receptor signaling**. *Genes Dev* 2000, **14**:1651-1663.

11. Brown MS, Ye J, Rawson RB, Goldstein JL: **Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans**. *Cell* 2000, **100**:391-398.

12. Urban S, Freeman M: **Intramembrane proteolysis controls diverse signalling pathways throughout evolution**. *Curr Opin Genet Dev* 2002, **12**:512.

13. Wolfe MS, Xia W, Ostaszewski BL, Diehl TS, Kimberly WT, Selkoe DJ: **Two transmembrane aspartates in presenilin-1 required for presenilin endoproteolysis and gamma-secretase activity**. *Nature* 1999, **398**:513-517.

14. Steiner H, Kostka M, Romig H, Basset G, Pesold B, Hardy J, Capell A, Meyn L, Grim ML, Baumeister R, Fechteler K, Haass C: **Glycine 384 is required for presenilin-1 function and is conserved in bacterial polytopic aspartyl proteases**. *Nat Cell Biol* 2000, **2**:848-851.

15. Sreekumar KR, Aravind L, Koonin EV: **Computational analysis of human disease-associated genes and their protein products**. *Curr Opin Genet Dev* 2001, **11**:247-257.

16. Rather PN, Orosz E: **Characterization of aarA, a pleiotrophic negative regulator of the 2'-N-acetyltransferase in Providencia stuartii**. *J Bacteriol* 1994, **176**:5140-5144.

17. Rather PN, Ding X, Baca-DeLancey RR, Siddiqui S: **Providencia stuartii genes activated by cell-to-cell signaling and identification of a gene required for production or activity of an extracellular factor**. *J Bacteriol* 1999, **181**:7185-7191.

18. Gallio M, Kylsten P: **Providencia may help find a function for a novel, widespread protein family**. *Curr Biol* 2000, **10**:R693-694.

19. Urban S, Schlieper D, Freeman M: **Conservation of intramembrane proteolytic activity and substrate specificity in prokaryotic and eukaryotic rhomboids**. *Curr Biol* 2002, **12**:1507.

20. Gallio M, Sturgill G, Rather P, Kylsten P: **A conserved mechanism for extracellular signaling in eukaryotes and prokaryotes**. *Proc Natl Acad Sci U S A* 2002, **99**:12208-12213.

21. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes**. *Genome Res* 2002, **12**:1048-1059.

22. Brown JR, Doolittle WF: **Archaea and the prokaryote-to-eukaryote transition**. *Microbiol Mol Biol Rev* 1997, **61**:456-502.

23. Woese CR, Kandler O, Wheelis ML: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya**. *Proc Natl Acad Sci U S A* 1990, **87**:4576-4579.

24. Doolittle WF: **You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes [In Process Citation]**. *Trends Genet* 1998, **14**:307-311.

25. Koonin EV, Galperin MY: *Sequence - Evolution- Function. Computational Approaches in Comparative Genomics*. New York: Kluwer Acad Publ; 2002.

26. Snel B, Bork P, Huynen MA: **Genomes in flux: the evolution of archaeal and proteobacterial gene content**. *Genome Res* 2002, **12**:17-25.

27. Weihofen A, Binns K, Lemberg MK, Ashman K, Martoglio B: **Identification of signal peptide peptidase, a presenilin-type aspartic protease**. *Science* 2002, **296**:2215-2218.

28. Koonin EV, Aravind L: **Origin and evolution of eukaryotic apoptosis: the bacterial connection**. *Cell Death Differ* 2002, **9**:394-404.

29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

30. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.

31. Hofmann K, Stoffel W: **TMbase - A database of membrane spanning proteins segments**. *Biol. Chem. Hoppe-Seyler* 1993, **374**:166.

32. Persson B, Argos P: **Prediction of membrane protein topology utilizing multiple sequence alignments**. *J Protein Chem* 1997, **16**:453-457.

33. Fitch WM, Margoliash E: **Construction of phylogenetic trees**. *Science* 1967, **155**:279-284.

34. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods**. *Methods Enzymol* 1996, **266**:418-427.

35. Adachi J, Hasegawa M: *MOLPHY: Programs for Molecular Phylogenetics*. Tokyo: Institute of Statistical Mathematics; 1992.

36. Kishino H, Miyata T, Hasegawa M: **Maximum likelihood inference of protein phylogeny and the origin of chloroplasts**. *J. Mol. Evol.* 1990, **31**:151-160.

Figure legends

Figure 1. **Multiple alignment of the conserved core of the rhomboid family proteins**.

The alignment includes the majority of the detected rhomboid family proteins; some closely related sequences were omitted. Only the six conserved (predicted) transmembrane helices and short surrounding regions are shown. The boundaries of the predicted TMH are indicated by shading and overline and the TMH are numbered 1-6. The number of amino acid residues in the omitted terminal and internal regions are indicated. The consensus shows amino acid residues present in at least 90% of the aligned sequences; h stands for hydrophobic residues (A,C,I,L,V,M,F,Y,W) and s for small residues (G,A,S,D,N,V). The proposed catalytic serine (TMH4) and histidine (TMH6) as well as conserved residues in TMH3 with possible ancillary roles in catalysis are highlighted with color. The proteins are identified with the gene identification (GI) number from the non-redundant database and an abbreviated species name. Bacterial species are color-coded green, eukaryotic species blue and archaeal species orange. Species name abbreviations: Aerpe, *Aeropyrum pernix*, Agrtu, *Agrobacterium tumefaciens*, Anoga, *Anopheles gambiae*, Arath, *Arabidopsis thaliana*; Arcfu, *Archaeoglobus fulgidus*, Bacsu, *Bacillus subtilis*, Brume, *Brucella melitensis*, Caeel, *Caenorhabdtitis elegans*, Caucr, *Caulobacter crescentus*, Chlte, *Chlorobium tepidum*, Cloac, *Clostridium acetobutilicum*, Corgl, *Corynebacterium glutamicum*, Deira, *Deinococcus radiodurans*, Dicdi, *Dictyostelium discoideum*, Drome, *Drosophila melanogaster*, Escco, *Escherichia coli*, Haein, *Haemophilus influenzae*, Halsp, *Halobacterium sp*., Homsa, *Homo sapiens*, Lacla, *Lactococcus lactis*, Lisin, *Listeria*

innocua, Metja, *Methanoccocus jannaschii*, Metka, *Methanopyrus kandleri*, Metma, *Methanosarcina mazei*, Meslo, *Mesorhisobium loti*, Mycle, *Mycobacterium leprae*, Myctu, *Mycobacterium tuberculosis*, Neucr, *Neurospora crassa*, Nossp, *Nostoc sp*., Prost, *Providencia stuartii*, Pyrab, *Pyrococcus abyssi*, Pyrae, *Pyrobaculum aerophilum*, Ralso, *Ralstonia solanaraceum*, Sacce, *Saccharomyces cerevisiae*, Schpo, *Schizosaccharomyces pombe*, Sinme, *Sinorhisobium meliloti*, Strco, *Streptomyces coelicolor*, Strpn, *Streptococcus pneumoniae*, Sulso, *Sulfolobus solfataricus*, Sulto, *Sulfolobus tokodaii*, Synsp, *Synechocystis sp*., Theac, *Thermoplasma acidophilum*, Thema, *Thermotoga maritima*, Thete, *Thermus thermophilus*, Vibch, *Vibrio cholerae*, Xanca, *Xanthomonas campestris*, Xylfa, *Xylella fastidiosa*.

Figure 2. **Phylogenetic tree of the rhomboid family**.

The sequences and their regions used for the construction of the tree are exactly those shown in Fig. 1. The color coding and abbreviations are as in Fig. 1. The two major eukaryotic subfamilies are denoted as RHO and PARL (see text) and four clusters containing unexpected, from a phylogenetic viewpoint, sets of species are denoted 1-4. Although the tree is shown in a rooted form for convenience, this is an unrooted tree; in particular, the placement of the "root" in the midst of the PARL subfamily is arbitrary. Internal nodes with at least 70% RELL bootstrap supported are denoted by circles. Domain architectures are connected to the respective proteins by brackets or by lines. The domain key is shown in the bottom of the figure.

Figure 3

A hypothetical scenario for the origin and dissemination of the rhomboid family proteases.

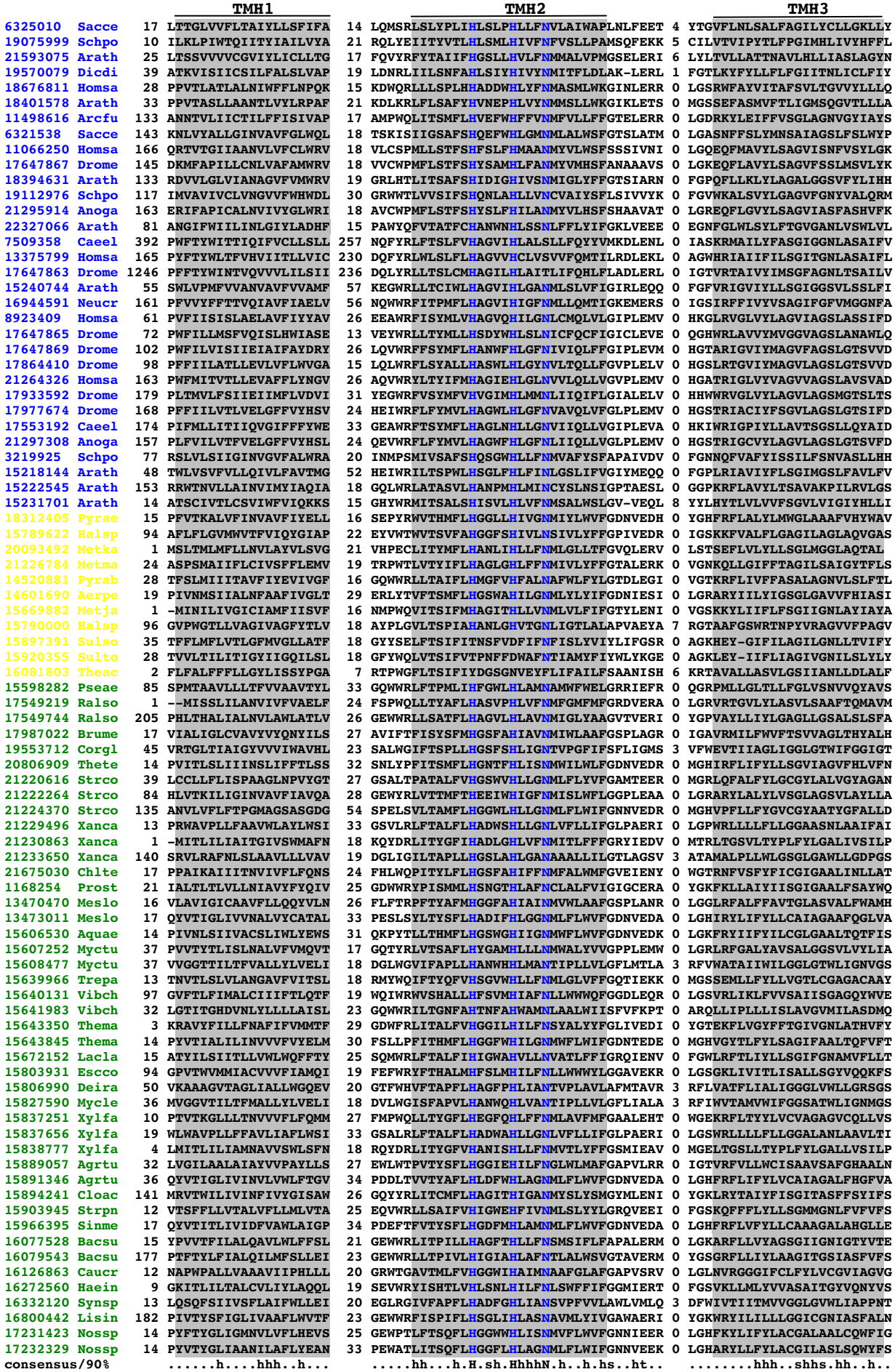
The figure schematically shows the proposed three stages of evolution of the rhomboid family:

I – the progenitor of the rhomboid family functions as a transporter for a regulatory peptide in some bacterial lineage

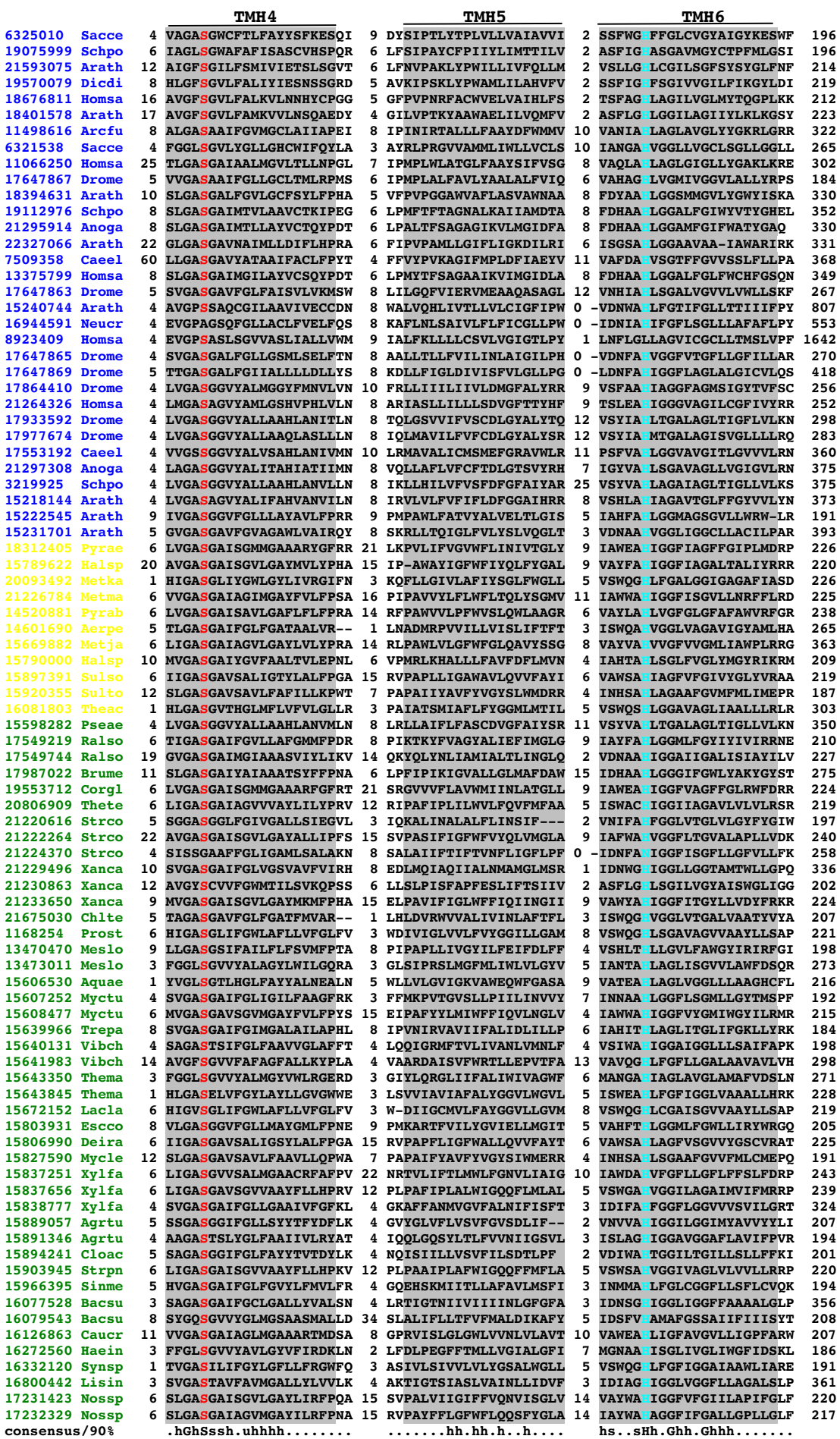
II – the catalytic site of the intramembrane protease evolves allowing the switch to RIP as the mechanism of the regulatory peptide release

III – the emergence of RIP is followed by a burst of HGT

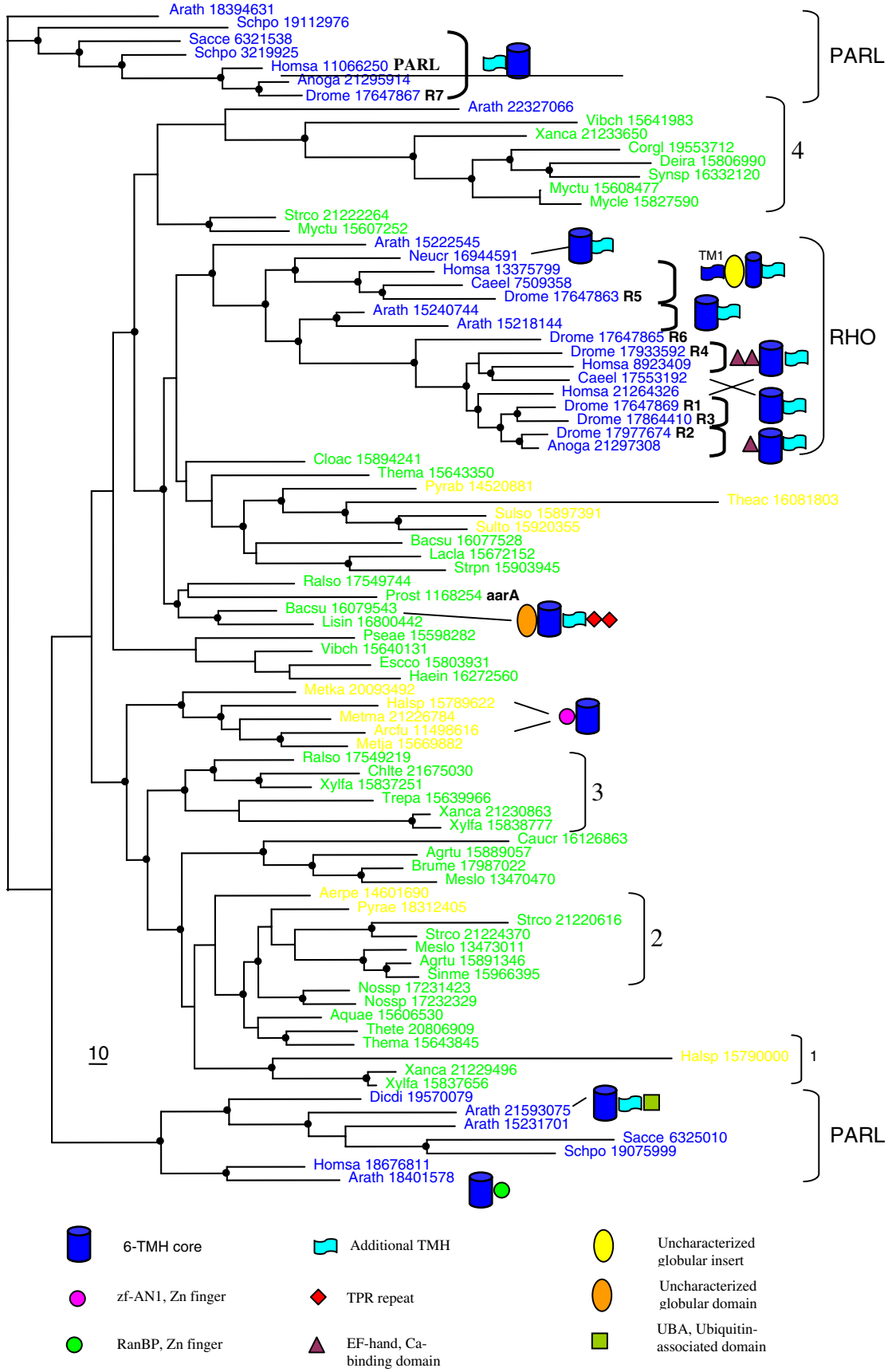
R, regulatory peptide; the transmembrane helices of rhomboid are designated as in Fig. 1, their topology in the membrane is based on that proposed in Ref. 7; the catalytic histidine and serine are shown and connected by a dotted line to indicate the proposed charge-relay system of the protease; possible ancillary catalytic residues are not shown.



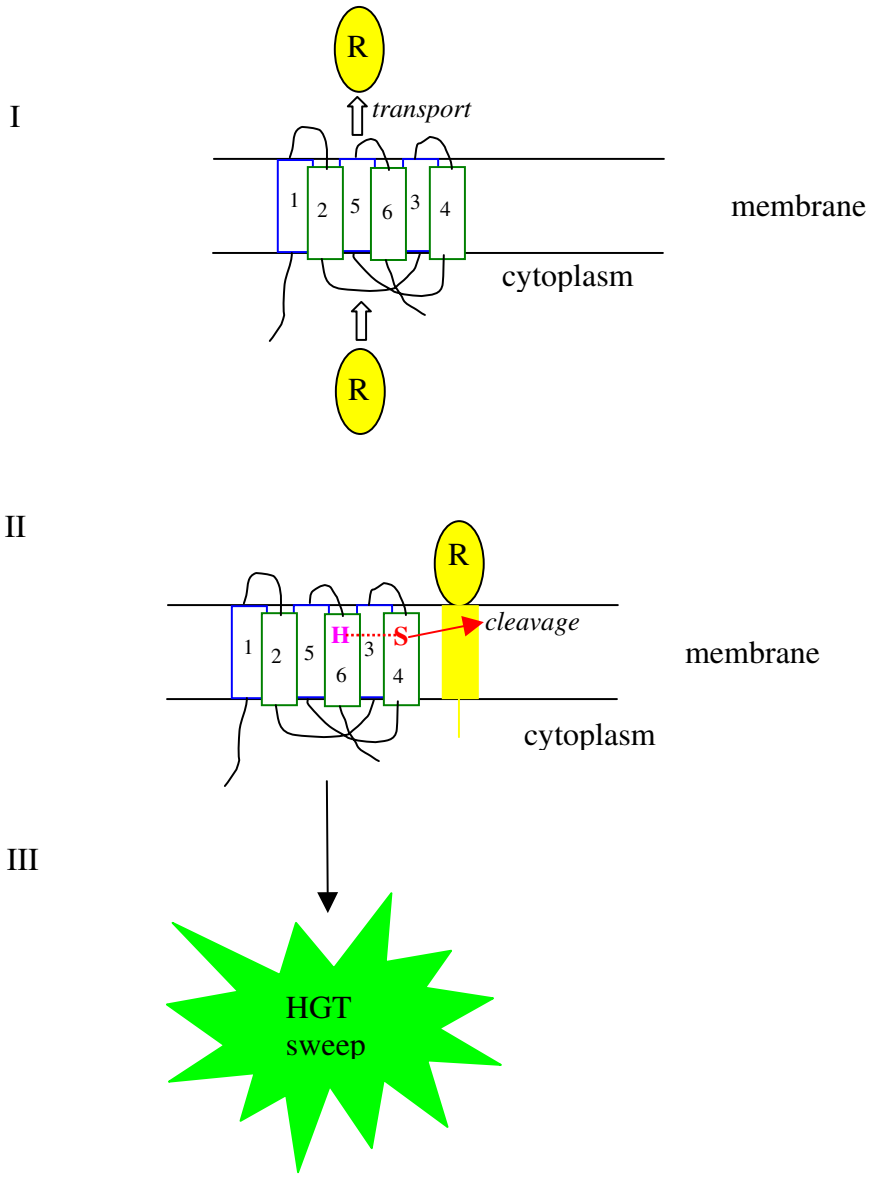
Koonin et al., Fig. 1



Koonin et al., Fig. 1, contd



Koonin et al., Fig. 2



Koonin et al., Fig. 3