The rhomboids: a nearly ubiquitous family of intramembrane serine proteases that probably evolved by multiple ancient horizontal gene transfers

Eugene V Koonin*, Kira S Makarova*, Igor B Rogozin*, Laetitia Davidovic†, Marie-Claude Letellier† and Luca Pellegrini†

Addresses: *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. †Centre de Recherche Université Laval Robert Giffard, Université Laval, Chemin de la Canadiere, G1J 2G3 Quebec, Canada.

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

Abstract

Background: The rhomboid family of polytopic membrane proteins shows a level of evolutionary conservation unique among membrane proteins. They are present in nearly all the 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 carried out using several independent methods for tree constructions and the corresponding statistical tests suggests that, despite its broad distribution in all three superkingdoms, the rhomboid family was not present in the last universal common ancestor of extant life forms. Instead, we propose that rhomboids evolved in bacteria and have been acquired by archaea and eukaryotes through several independent horizontal gene transfers. In eukaryotes, two distinct, ancient acquisitions apparently gave rise to the two major subfamilies, typified by rhomboid and PARL (presenilins-associated rhomboid-like protein), respectively. Subsequent evolution of the rhomboid family in eukaryotes proceeded by 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 a likely bacterial origin with subsequent dissemination by horizontal gene transfer. This emphasizes the importance of explicit phylogenetic analysis for the reconstruction of ancestral life forms. A hypothetical scenario for the 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 nonenzymatic 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. As 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 TGFα-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 seven 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 transduction cascades. Until recently, the only characterized cases of RIP in eukaryotes involved presenilin-1, an aspartyl protease, which cleaves a transmembrane helix in type-1 membrane proteins such as amyloid β-precursor protein (AβPP), Notch and Ire1 [13], and the metalloprotease S2P, 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 presenilins also might be homologous to bacterial and archaeal type IV prepilin peptidases, although, in this case, the sequence similarity is 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 metazoan, 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, can cleave the EGF 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 [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. Eukaryotic rhomboid family proteins seem to show considerable functional variability; in particular, cross-talk might exist between different RIP pathways. A distinct representative of the rhomboid family has been shown to physically interact with presenilins 1 and 2, and was accordingly named presenilins-associated rhomboid-like protein (PARL) [6]. The yeast ortholog of PARL has been suggested to participate in the processing of cytochrome c peroxidase precursor during its import into the mitochondrion [21].

The near ubiquity of the rhomboid family among bacteria, archaea and eukaryotes, along with the remarkable functional conservation, suggests 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 of the rhomboid family

Although the sequence similarity between eukaryotic and prokaryotic rhomboid family proteins is relatively low (around 10-15% 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 TMHs (Figure 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 (Figure 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 asparagine is required for the cleavage of Spitz by RHO [7].

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 (Figure 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 six TMH, whereas the eukaryotic members typically have an additional seventh TMH, which may be attached to the core either from the amino terminus or from the carboxyl 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 microsporidian Encephalitozoon cuniculi, a eukaryotic intracellular parasite with a highly degraded genome, the archaea Methanothermobacter thermoautotrophicus and Thermoplusma volcanium, and several bacterial species, primarily parasites with small genomes but also species with moderately sized genomes, such as Xylella fastidiosa (see COG0705 at [22]). In two instances, 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 have a single gene coding for a rhomboid-family protein, although some have two or three paralogs (see COG0705 [22]); in contrast, eukaryotes show expansion of the rhomboid family, with seven 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 (Figure 1) was employed to construct a phylogenetic tree using the least-squares algorithm with subsequent optimization using the maximum likelihood (ML) 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 noise from potentially misaligned residues (except for the Bayesian analysis, which used the complete alignment; see Materials and methods). The alignment used for phylogenetic reconstructions included 87 sequences and 149 aligned sites. The phylogenetic tree of the rhomboid family presents a complex and unexpected picture (Figure 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 (Figure 2). The first subfamily, which includes six of the seven Drosophila rhomboids, clusters with a distinct prokaryotic assemblage, consisting primarily of Gram-positive bacteria as well as a subset of archaea; this clade is strongly supported by bootstrap analysis (Figure 2). The proteins in this group of eukaryotic rhomboids, which we designated the RHO subfamily, typically have an extra TMH added carboxy-terminally to the 6-TMH core; some of these proteins also contain EF-hand calcium-binding domains amino-terminally of the core (Figure 2).

The second eukaryotic subfamily, which we designated the PARL subfamily, after PARL, the human ortholog of...
Figure 1 (see legend on the previous page)
| Gene  | Venn Diagrams |
|-------|---------------|
| Synsp | [Venn Diagram] |
| Bacsu | [Venn Diagram] |
| Mycle | [Venn Diagram] |
| Aquae | [Venn Diagram] |
| Prost | [Venn Diagram] |
| Xanca | [Venn Diagram] |
| Sulso | [Venn Diagram] |

16332120 Synsp    1
16077528 Bacsu    5
15827590 Mycle    6
15606530 Aquae    9
1168254  Prost   19
21230863 Xanca   12
15897391 Sulso    5
8923409  Homsa    4
15222545 Arath   10

**VVGA**
**VLGA**
**HIGA**
**LIGA**
**LVGA**
**SSGA**
**AVGF**
**SVGA**
**VVGA**
**AVGY**
**LVGA**
**LAGA**
**LVGA**
**EVGP**
**AVGP**
**LLGA**
**GLGA**
**SLGA**

**IAGL**

**S**

**TSLYGLFAAIIVLRYAT  4 IQ**
**GAVSAVLFAAVLLQPWA  7 PA**
**GTLHGLFAYYALNEALN  5 WL**
**GAIAGVLGAYLVLYPRA 14 RL**
**GAIFGLFGATAALVR--  1 LN**
**GAISAVLGAFLFLFPRA 14 RF**
**GAIAGIMGAYFVLFPSA 16 PI**
**GGVFGLLLAYAVLFPRR  9 PM**
**GAIFGIMGALAILAPHL  8 IP**
**GAIYAIAAATSYFFPNA  6 LP**
**GGVYALLAAQLASLLLN  8 IQ**

**GWCFTLFAYYSFKESQI  9 DY**

2003,

**VSWQG**
**IDNSG**
**IAWDA**
**ISWEA**
**ISWAC**
**VAYVA**
**IAWWA**
**IAHFA**
**VDIWA**
**IDNWG**
**VSHLA**
**VSYVA**
**IGYVA**
**PSFVA**
**VSFAA**
**VAFDA**
**FDYAA**
**ASFIG**

**H**

**IGGFVFGIILAPIFGLF  220**
**LFGFIGGAIAAWLIARE  191**
**VGGILAGAIMVIFMRRP  239**
**LSGAAFGVVFMLCMEPQ  191**
**IGGFVYGMIWGYILRMR  215**
**LAGAAFGVMFMLIMEPR  187**
**IGGFIAGFFGIPLMDRP  226**
**VGGLIGGCLLACILPAR  393**
**IGGILGGIMYAVVYYLI  207**
**VGGFLTGVALAPLLVDK  240**
**FGGLVTGLVLGYFYGIW  197**
**IGGLLGGTAMTWLLGPQ  336**
**IFGFLSGLLLAFAFLPY  553**
**VSGTFFGVVSSLFLLPA  368**
**LGGAAVAA-IAWARIRK  331**
**LGGSMMGVLYGWYISKA  330**

**IAGL**

**S**

**TSLYGLFAAIIVLRYAT  4 IQ**
**GAVSAVLFAAVLLQPWA  7 PA**
**GTLHGLFAYYALNEALN  5 WL**
**GAIAGVLGAYLVLYPRA 14 RL**
**GAIFGLFGATAALVR--  1 LN**
**GAISAVLGAFLFLFPRA 14 RF**
**GAIAGIMGAYFVLFPSA 16 PI**
**GGVFGLLLAYAVLFPRR  9 PM**
**GAIFGIMGALAILAPHL  8 IP**
**GAIYAIAAATSYFFPNA  6 LP**
**GGVYALLAAQLASLLLN  8 IQ**

**GWCFTLFAYYSFKESQI  9 DY**

2003,
Drosophila RHO [6], resides within a large, heterogeneous prokaryotic cluster (Figure 2). Within this subfamily, PARL and its orthologs from other animals and from fungi have distinct domain architecture, with an extra TMH added to the amino terminus of the core, whereas the rest have only the core (a carboxy-terminal TMH and a ubiquitin-associated domain are appended in one Arabidopsis protein; Figure 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 by both ancient and more recent duplications. Several lineage-specific expansions of paralogs [23] are noticeable, in insects, mammals and plants (Figure 2).

Archaeal rhomboids are scattered over the phylogenetic tree, with two major clusters and, in addition, three isolated proteins joining different bacterial branches (Figure 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 an odd composition, such as the grouping of proteobacterial and Gram-positive species; some of these clusters are well supported by bootstrap (see clusters 1-4 in Figure 2).

Unexpected tree topologies often emerge due to artifacts of phylogenetic analysis methods. This concern is particularly serious for highly divergent families of membrane proteins, such as the rhomboids, in which parallel amino-acid substitutions are likely. Therefore we investigated the phylogeny of the rhomboid family in greater detail using several independent phylogenetic methods and the corresponding statistical tests. First, we assessed the robustness of the topology of the tree shown in Figure 2 using the Kishino-Hasegawa (KH) test whereby the clade of interest is forced into various positions on the tree and the likelihoods of the resulting topologies are estimated. Specifically, the KH test was used to evaluate two alternative topologies, in which the RHO and PARL subfamilies formed a clade, and two topologies, in which the RHO subfamily formed a clade with archaeal rhomboids (Figure 2 and Table 1). Each of these alternative topologies had a significantly lower likelihood than the original topology shown in Figure 2 (see Table 1).

In addition, a tree of the rhomboid family was constructed using the Bayesian inference method, which has recently become a practical alternative to the more traditional methods of phylogenetic analysis [24,25]. The tree produced using the MRBAYES package [26] showed the same major clades as the tree in Figure 2 (data not shown); moreover, clustering of the RHO and PARL subfamilies of eukaryotic rhomboids with the respective prokaryotic clades was supported by high posterior probabilities (Figure 2).

We also attempted to construct a phylogenetic tree of the rhomboid family by using the maximum parsimony method [27]. The resulting tree contained the same major clades as the trees constructed using ML and MRBAYES; however, the number of parsimony-informative sites was insufficient to obtain high bootstrap support with this approach (data not shown).

We also tested alternative phylogenies using neighboring search with constraint trees [27]. The alternative phylogenies reflected two different hypotheses: first, clustering of the RHO and PARL subfamilies of eukaryotic rhomboids with the prokaryotic rhomboid families as suggested by the tree topology in Figure 2; and second, monophyly of the eukaryotic rhomboids (Figure 3). The phylogenies corresponding to these alternative hypotheses were compared to the best phylogeny using three statistical tests (Table 2). The

| Table 1 |
| Log-likelihood analysis of possible placements of selected branches of maximum likelihood trees for the proteins analyzed |
| Tree\(\text{a}\) | Diff lnL | SE\(\text{b}\) | RELL-BP\(\text{c}\) |
|----------|---------|---------|----------|
| Original tree | 0.0 | - | 0.9702 |
| A → B | -18.9 | 10.2 | 0.0264 |
| B → A | -46.6 | 14.6 | 0.0003 |
| A → C | -30.3 | 12.8 | 0.0031 |
| A → D | -47.9 | 15.6 | 0.0000 |

\(\text{a}\) A-D, clades that were subjected to local rearrangements in the tree as indicated in Figure 2 and discussed in the text. \(\text{b}\) Difference of the log-likelihoods relative to the best tree. \(\text{c}\) Standard error of Diff lnL.

Figure 2 (see figure on the next page)

Phylogenetic tree of the rhomboid family. The sequences and their regions used to construct the tree are exactly those shown in Figure 1. The color coding and abbreviations are as in Figure 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 I-4. The clades that were investigated in the KH test are denoted A through D. Although the tree is shown in a pseudorooted form for convenience, this is an unrooted tree. Internal nodes with at least 70% RELL bootstrap support are denoted by black circles and nodes with a 50-70% support by blue circles. The posterior probabilities reported by the MRBAYES program are indicated for some key internal branches. Domain architectures are connected to the respective proteins by brackets or lines. The domain key is shown at the bottom of the figure.
Figure 2 (see legend on the previous page)
The RHO and PARL subfamilies are denoted; the remaining clusters include prokaryotic rhomboids designated as in Figure 2 (with ‘a’ added to the GI number). Within each cluster, the branches were collapsed into a multifurcation. The RHO and PARL subfamilies are denoted; the remaining clusters include prokaryotic rhomboids designated as in Figure 2 (with ‘a’ added to the GI number). Within each cluster, the branches were collapsed into a multifurcation. The trees are unrooted, although shown in a pseudorooted form.

The phylogenetic tree of the rhomboid family shown in Figure 2 and supported by the additional tests described above follows neither the ‘standard model’ scenario [28,29], 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 tree seems to be that the evolutionary history of this family has 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 superkingdoms, phylogenetic analysis suggests that this family has not been inherited from the last universal common ancestor (LUCA). Instead, the tree topology seems to indicate that this family emerged in some bacterial lineage and afterwards had been widely disseminated by HGT, and then lost in some lineages. Both archaea and eukaryotes seem 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 (Figures 2,3).
Given the broad phylogenetic representation of both subfamilies of eukaryotic rhomboids, both the RHO subfamily and the PARL subfamily must have been acquired through 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 [30,31].

An alternative to this multiple-HGT scenario is that LUCA already had multiple, paralogous rhomboids, which evolved by 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 out formally, this hypothesis implies the existence of an elaborate signaling system in LUCA and, accordingly, suggests that LUCA was a complex organism, which might have had as many genes as modern bacteria. Theoretical analysis of evolutionary scenarios constructed on the basis of the phyletic patterns of COGs by applying the parsimony principle shows that the complexity of the inferred gene set of LUCA critically depends on the relative rates of gene loss and HGT at the early stages of evolution [32]. A complex LUCA with around 2,000 genes is predicted only when one assumes that the rate of gene loss is an order of magnitude greater than the rate of HGT. However, explicit reconstruction of the gene set of LUCA under the assumption of equal rates of gene loss and HGT leads to a hypothetical genome that consists of only around 600 genes but appears to be 'compatible with life', that is, it includes genes responsible for most, if not all, essential cellular functions [32]. We currently believe that this is the most realistic, albeit inevitably imprecise, reconstruction of LUCA's gene set. With respect to the rhomboid family and other families whose phylogenetic trees show similar patterns, this makes the multiple-HGT interpretation the scenario of choice. Further theoretical, comparative-genomic and experimental analyses aimed at determining relative rates of gene loss and HGT will help in a more objective assessment of the validity of this argument.

The multiple-HGT interpretation of the evolutionary history of the rhomboid family, while supported by the above argument, seems, at least at first glance, distinctly counter-intuitive, given that this family is nearly ubiquitous among extant life
forms. Indeed, when attempts are made to construct parsimonious evolutionary scenarios on the basis of phyletic patterns alone [31-33], 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 [34]. For the rhomboid family, the multiple-HGT mode of evolution seems to be particularly plausible. It seems likely that the ultimate ancestor of the rhomboid family evolved from a nonenzymatic 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 acids within two or three of the TMHs (Figure 4). 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 was 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 a regime of relatively slow evolution, which ensures significant sequence similarity between all family members (Figure 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 [35].

**Conclusions**

The rhomboid family might 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 suggests a different scenario, one of emergence in a bacterial lineage with subsequent multiple, independent HGT events and gene losses. Although caution is due in the evolutionary interpretation of phylogenetic trees for large families, particularly when membrane proteins with a relatively small number of conserved positions, such as the rhomboids, are involved, the multiple-HGT scenario seemed to be supported by several methods of tree analysis and statistical tests.

Eukaryotes probably acquired their two major rhomboid subfamilies, RHO and PARL, as the result of two independent, early HGT events. These events, which might have introduced RIP as a means of intercellular communication, could 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 [36]. Subsequent evolution of rhomboids in eukaryotes proceeded by lineage-specific expansion of paralogs [23] followed by

---

**Figure 4**

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. In (a), the progenitor of the rhomboid family functions as a transporter for a regulatory peptide in some bacterial lineage. In (b), the catalytic site of the intramembrane protease evolves, allowing the switch to RIP as the mechanism of the regulatory peptide release. In (c), the emergence of RIP is followed by a HGT. R, regulatory peptide. The transmembrane helices of rhomboid are designated as in Figure 1; their topology in the membrane is based on that proposed in [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.
diversification through the addition of an extra TMH in different positions relative to the catalytic core, some limited domain accretion (see Figure 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 case.

Materials and methods
The nonredundant (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 [37]. 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 [38] and manually adjusted on the basis of the examination of PSI-BLAST search outputs and the superposition of the predicted TMHs, which were identified using the programs TMpred [39] and TMAP [40].

Phylogenetic trees were built using the least-squares method [41] implemented in the FITCH program of the PHYLIP package [42], with subsequent local rearrangement using the PROTML program of the MOLPHY package to obtain the maximum likelihood tree [43]. The reliability of the tree topology was assessed using the RELL (resampling of estimated log-likelihoods) bootstrap method of MOLPHY, with 10,000 replications [44]. Alternative placements of selected clades in maximum-likelihood trees were compared by using the rearrangement optimization method (Kishino-Hasegawa test) as implemented in the ProtML program of the PHYLIP package [42], with subsequent local rearrangement using the fitch program of the PHYLIP package [46].

Constraint trees for phylogenetic hypothesis testing were generated using the TreeView program [46]. Constraint trees were imported into PAUP* [27] and subjected to neighborhood search to generate the phylogenies corresponding to alternative hypotheses. These phylogenies were compared using the KH [45], Templeton (Wilcoxon signed-ranks) [47] and Winning-sites (sign) [48] tests implemented in PAUP*.

Acknowledgements
L.F. is supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References
1. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kyrizyn 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, 7:971-973.

3. Sturtevant MA, Roark M, O’Neill JW, Biels 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, Biels 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. Pellegreni L, Passer BJ, Canelles M, Lefterov I, Ganjie JK, Fowlwes BK, 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-R23.

9. Guichard A, Roark M, Ronshaugen M, Bier E: brother of rhomboid, a rhomboid-related gene expressed during early Drosophila oogenesis, promotes EGFR/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, 12:5-12.

13. Wolfe MS, Xia W, Ostaszewski BL, Diehl TS, Kimberly WT, Wolfe MS: COGS: phylogenetic classification of proteins encoded in complete genomes. [http://www.ncbi.nlm.nih.gov/COG]
23. 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.

24. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP: Bayesian inference of phylogeny and its impact on evolutionary biology. Science 2001, 294:2310-2314.

25. Huelsenbeck JP, Larget B, Miller RE, Ronquist F: Potential applications and pitfalls of bayesian inference of phylogeny. Syst Biol 2002, 51:673-688.

26. Huelsenbeck JP, Ronquist F: MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001, 17:754-755.

27. Swofford DL: PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Sunderland, MA: Sinauer; 1998.

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

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

30. Doolittle WF: You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet 1998, 14:307-311.

31. Koonin EV, Galperin MY: Sequence - Evolution - Function. Computational Approaches in Comparative Genomics. Boston: Kluwer; 2002.

32. Mirkin BG, Fenner TI, Galperin MY, Koonin EV: Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 2003, 3:2.

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

34. Gogarten JP, Doolittle WF, Lawrence JG: Prokaryotic evolution in light of gene transfer. Mol Biol Evol 2002, 19:2226-2238.

35. 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.

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

37. 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.

38. 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.

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

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

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

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

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

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

45. Kishino H, Hasegawa M: Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 1989, 29:170-179.

46. Page RD: TREEVIEW: an application to display phylogenetic trees on personal computers. Comput Appl Biosci 1996, 12:357-358.

47. Templeton AR: Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the humans and apes. Evolution 1983, 37:221-244.

48. Prager EM, Wilson AC: Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J Mol Evol 1988, 27:326-335.