Probiotics and Antimicrobial Proteins (2022) 14:603–612
https://doi.org/10.1007/s12602-022-09948-y

Expanding the Database of Signal-Anchor-Release Domain Endolysins Through Metagenomics

Marco Túlio Pardini Gontijo[1] · **Mateus Pereira Teles**[1,2] · **Pedro Marcus Pereira Vidigal**[3] · **Marcelo Brocchi**[1]

Accepted: 3 May 2022 / Published online: 7 May 2022

Abstract

Endolysins are bacteriophage-derived lytic enzymes with antimicrobial activity. The action of endolysins against Gram-negative bacteria remains a challenge due to the physical protection of the outer membrane. However, recent research has demonstrated that signal-anchor-release (SAR) endolysins permeate the outer membrane of Gram-negative bacteria. This study investigates 2628 putative endolysin genes identified in 183,298 bacteriophage genomes. Previously, bioinformatic approaches resulted in a database of 66 SAR endolysins. This manuscript almost doubles the list with 53 additional SAR endolysin candidates. Forty-eight of the putative SAR endolysins described in this study contained one muramidase catalytic domain, and five included additional cell wall–binding domains at the C-terminus. For the moment, SAR domains are found in four protein families: glycoside hydrolase family 19 (GH19), glycoside hydrolase family 24 (GH24), glycoside hydrolase family 25 (GH25), and glycoside hydrolase family 108 (GH108). These SAR lysis are clustered in eight groups based on biochemical properties and domain presence/absence. Therefore, in this study, we expand the arsenal of endolysin candidates that might act against Gram-negative bacteria and develop a consult database for antimicrobial proteins derived from bacteriophages.

Keywords Endolysin · Antimicrobial · Bacteriophage-mediated lysis · Enzybiotic · Antibiotic substitute · Novel therapy

Introduction

Bacteriophages (phages) are viruses that infect and replicate within bacterial cells [1]. These viruses follow two distinct life cycles once in contact with the host. Bacteriophages replicate and release new viral particles through lysis in the lytic cycle. In contrast, phages integrate their genomes into the chromosome of their bacterial host in the lysogenic cycle [2]. Bacteriophages have long been studied to control undesirable bacteria [3, 4].

Bacteriophages are the most diverse organisms on the planet. These viruses are identified in the environments where prokaryotes are present [5]. At the end of the lytic infectious cycle, in particular, the viral progeny is released from the cytoplasm of their bacterial host after the cleavage of the peptidoglycan (PG) [6]. The combined action of bacteriophage-derived lytic enzymes causes PG cleavage [7].

The enzymes directly responsible for PG cleavage are subdivided into five superfamilies: (I) muramidases, which hydrolyze the β-1,4 bonds between N-acetylmuramic acid (M) and N-acetyl-D-glucosamine (G) residues of the PG; (II) amidases, which hydrolyze the bond between M

✉ Marco Túlio Pardini Gontijo
m264546@dac.unicamp.br; pardinibiotech@gmail.com

Mateus Pereira Teles
m241581@dac.unicamp.br

Pedro Marcus Pereira Vidigal
pedro.vidigal@ufv.br

Marcelo Brocchi
mbrocchi@unicamp.br

[1] Departamento de Genética, Evolução, Microbiologia e Imunologia, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Rua Monteiro Lobato 255, Campinas, São Paulo 13083-862, Brazil

[2] Faculdade de Farmácia, Universidade Estadual de Campinas (UNICAMP), Cidade Universitária Zeferino Vaz, Rua Cândido Portinari 200, Campinas, São Paulo 13083-862, Brazil

[3] Núcleo de Análise de Biomoléculas (NuBioMol), Universidade Federal de Viçosa (UFV), Vila Gianetti, Casa 21, Campus da UFV, Viçosa, Minas Gerais 36570-900, Brazil

residue and L-alanine of the PG; (III) transglycosylases, which act on the β-1,4 bonds between M and G by a non-hydrolytic cleavage; (IV) peptidases, which cleave the peptide bonds in the interpeptide PG bridges; and (V) glucosaminidases, which cleave the β-1,4 glycosidic bonds between G and M monomers [8].

Typically, phage holins disrupt the cytoplasmatic membrane to give endolysins access to the PG [9]. Holins are accessory phage-encoded proteins that act on the cytoplasmatic membrane of bacteria to provide access to the PG for endolysins. In contrast, endolysins that contain a canonical signal peptide (SP) cross the cytoplasmic membrane by the bacterial Sec secretion system [9]. Other phage-encoded endolysins contain signal-anchor-release (SAR) domains at the N-terminus. SAR domains are transmembrane regions comprising an elevated content of glycine and alanine and low content of basic amino acids [10]. SAR endolysins concentrate in the cytoplasmatic membrane until pinholins; also, phage-encoded lytic proteins disrupt the proton motive force. Xu et al. [10] also observed that SAR endolysins might be secreted to the periplasm space by the bacterial Sec secretion system. SAR endolysins represent a promising source of antimicrobial proteins that act exogenously on Gram-negative bacteria [11–13].

Endolysins act on the peptidoglycan of bacteria. Several studies have accounted for the antibiotic action of endolysins against Gram-positive bacteria [14, 15]. In addition, the peptidoglycan is naturally exposed in Gram-positive bacteria. On the other hand, the application of endolysin-based treatments against Gram-negative bacteria is still a challenge [7, 11]. SAR endolysins emerge as a potential class of endolysins that act exogenously against Gram-negative bacteria.

SAR endolysins were first described by Xu et al. [10]. At this time, the signal sequence was named signal-arrest-release. Since its discovery, only a handful of papers have studied SAR lysins [16, 17]. The signal peptide was then renamed signal-anchor-release [18]. The only study assessed the antimicrobial activity of SAR endolysins against bacteria. Lim et al. [19] showed that the native SAR endolysin SPN9CC is active against Gram-negative and Gram-positive bacteria. However, the antimicrobial action against Gram-negative bacteria was more significant than against Gram-positive bacteria.

Some researchers made efforts to describe SAR lysins in previously annotated genomes [8, 20, 21]. Fernández-Ruiz et al. [22] found 2628 putative endolysin genes in the genomes of more than 180 thousand unculturable bacteriophages. However, the authors did not evaluate the presence of SAR domains within the sequences. Thus, this study expands the database of SAR endolysins through a metagenomic screening.

Materials and Methods

Dataset

We evaluated 2628 putative endolysin genes found in 183,298 genomic sequences of uncultured viral genomes previously identified by Fernández-Ruiz et al. [22]. The putative endolysin genes were identified in genomic sequences gathered from the surface and deep ocean [23–26], several locations on the earth's surface [27], and prophage signatures found in bacterial and archaeal genomes [28]. To examine the complete dataset, please access Supplementary File S2 in the work of Fernández-Ruiz et al. [22].

SAR Endolysin Search

We screened SAR endolysins as Oliveira et al. [20] and Gontijo et al. [11] described. We identified transmembrane domains using SOSUI version 1.1 [29], TMHMM version 2.0 [30], Phobius version 1.01 [31], and Topcons version 1.0 [32]. Putative proteins containing transmembrane domains at the N-terminus with a glycine and alanine content of 40 to 60% and a limit of two basic residues (lysine, arginine, and histidine) were annotated as a putative SAR endolysin [10]. Length, size, mass, and isoelectric points (IP) of proteins were determined by Isoelectric Point Calculator (IPC) version 2.0 [33]. The HHpred web server [34] and Protein family (Pfam) database [35] were used for screening the SAR endolysins. The parameters used were 80% query coverage and an *E* value cutoff of 1e −5. MAFFT version 7.450 [36] was used for protein sequences aligned and WebLogo version 2.8.2 [37] to build the sequence logo.

Clustering of Putative SAR Endolysins

The putative SAR endolysins identified in this study, along with the ones identified by Oliveira et al. [20], Valero-Rello [21], and Gontijo et al. [11], were clustered with principal component analysis (PCA) using ClustVis version 1.0 [38]. For the PCA, we considered the quantitative values of the biochemical characteristics of the proteins, such as size (aa), mass (kDa), and isoelectric point (IP). We considered "1" when a certain domain was present and "0" in its absence for the qualitative data. We constructed the phylogenetic relationship of SAR endolysins using phyloXML version 1.0.0 interactive platform [39].

Structure Prediction

The structures of representative SAR endolysins identified in this study, along with the ones specified by Oliveira et al.

[20], Valero-Rello [21], and Gontijo et al. [11], were predicted as described by Takahashi et al. [40] using RoseTTAFold modeling [41, 42]. We compared structures using FATCAT pairwise alignment version 2.0 [43]. The protein structures were visualized using Mol* Viewer version 3.0.2 [44].

Results

Metagenomic Data Revealed 53 Unannotated Putative SAR Endolysins

Among the 2628 putative endolysins identified in the 183,298 bacteriophage genomes, 53 (2.02%) contained putative SAR domains at the N-terminus (Tables S1 to S4). The putative SAR endolysins described in this study comprised three muramidase catalytic domain (MCD) types and two cell wall–binding domains (CBDs), comprehending five architectural organizations. The complete list of the SAR endolysins is found in Supplementary File 1.

The length of the SAR domain varied from 17 to 23 amino acids. All putative SAR endolysins described in this study contained a single MCD adjacent to the SAR domain (Table S5). The 53 putative SAR endolysins were classified into three types of MCDs: 46 had the glycoside hydrolase family 24 (GH24) catalytic domain (PF00959), 6 had the glycoside hydrolase family 25 (GH25) domain (PF01183), and 1 contained the glycoside hydrolase family 108 (GH108) domain (PF05838).

The size of GH24 proteins varied from 143 (15.226 kDa) to 194 (21.505 kDa) amino acids. Their catalytic domain corresponded to about 66% of the amino acids, while the SAR domain corresponded to about 12%. The majority ($n = 42$) of the GH24 proteins had isoelectric points (IP) superior to the cytoplasmic pH (7.40).

In addition, GH25 proteins size varied from 250 (27.499 kDa) to 448 (48.456 kDa) amino acids, and their catalytic domain corresponded to about 45% of the protein. The SAR domain corresponds to about 5%. The IP of all putative GH25 endolysins was inferior to 7.40. Four endolysins having the GH25 domain contained one or two copies of the LysM domain (PF01476) at the C-terminus. The LysM domain is a cell wall–binding domain (CBD). Each CBD corresponds to about 11% of the whole protein.

Furthermore, the size of the putative GH108 endolysin described in this study is 207 amino acids (22.319 kDa). Its catalytic domain corresponds to 40% of the protein and the SAR domain about 10%. The IP of the putative GH108 is 8.18. The putative SAR endolysin GH108 was associated with one peptidoglycan binding family 3 (PG3) domain (PF09374) at the C-terminus, corresponding to 40% of the protein.

LysM and PG3 are related to cell wall–binding activity. The PG3 domain is longer than the LysM domains.

The Position of the SAR Domains Within the First 40 Amino Acids Is Variable

To determine the sequence variation in the SAR domains, we aligned the N-terminus end of all putative SAR endolysins described in the literature [8, 20, 21]. Multiple sequence alignment of the first 40 amino acids revealed five highly conserved regions at the N-terminus, three of which correspond to the SAR domain (Fig. 1a). The color scheme in Fig. 1a suggests preserving the degree of hydrophobicity within the SAR domain despite the amino acid variations. In addition, despite the position of the transmembrane domain,

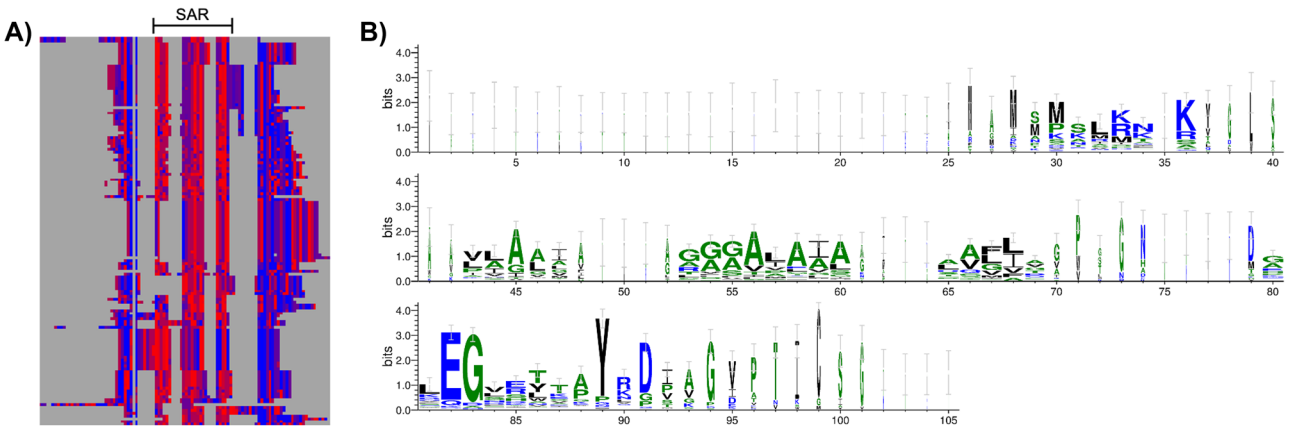


Fig. 1 **a** Multiple sequence alignment of the first 40 amino acids of all SAR endolysins identified in the literature. The color scale ranges from the most (red) to the least hydrophobic (blue) amino acid. **b** Sequence logo of the first 40 amino acids of all putative SAR endolysins identified in the literature. The height of the letters (*y*-axis) indicates the content of the alignment position (*x*-axis), in bits

we observed a conserved eight amino acid motif (G/A)(G/A)(G/A)A(L/I)A(A/I)A (Fig. 1b).

SAR Endolysins Described in the Literature Are Clustered in Eight Groups

We first determine the phylogenetic relationship of all putative SAR endolysins described in the literature [8, 20, 21]. A simple phylogenetic tree revealed two major clades, both containing proteins of the muramidase superfamily. The SAR endolysin of the glucosaminidase superfamily identified by Valero-Rello [21] is located in the interface of these two clades (Fig. 2a). Clade a contains all putative GH25 (PF01183) endolysins and the glycoside hydrolase family 19 (GH10; PF00182) protein in the interface. Clade b is composed only of proteins of the muramidase superfamily: GH24 (PF00959) and GH108 (PF05838) families.

Given the variability of biochemical characteristics and the possibility of uncharacterized CBDs, we hypothesized that other factors besides the type of the catalytic domain change the possible endolysin activity. To test this, we clustered all putative SAR endolysins ever described according to the presence/absence and variety of catalytic and cell wall–binding domains. PCA huddled putative SAR endolysins in eight groups (Fig. 2b). The analysis considered the biochemical characteristics of the proteins, such as size (aa), mass (kDa), and isoelectric point (IP). We considered possible unannotated CBDs as absent.

GH24 proteins (PF00959) were clustered in one group containing 123 proteins. In addition, GH25 putative proteins (PF01183; $n=10$) were clustered in 4 groups based on the presence and number of LysM domains (PF01476) and the size of the peptide. Furthermore, GH108 proteins (PF05838; $n=6$) were clustered in two groups, one of which was associated with the PG3 domain (PF09364). The last group contains the glucosaminidase GH19 (PF00182).

Structural Prediction Revealed Putative Unannotated Cell Wall–Binding Domains

To better characterize the SAR endolysins, we performed an in silico structural determination of representative protein sequences of each PCA group. First, to assess the accuracy of the prediction of RoseTTAFold modeling, we compared the predicted structure of the putative SAR endolysin from *Klebsiella* phage KpV475 (UniProtKB: A0A1B0Z137) with R^{21}, the SAR lysozyme of coliphage 21 (UniProtKB: P27359), whose structure was determined by crystallography [45]. The superimposed structures (Fig. S1) revealed the significance of the detected structural similarities (p value 1.67e − 11). After verifying the accuracy of the prediction, we proceed with the in silico structural analysis. The structural study suggests that MCDs associated with CBDs contain a flexible hinge between the domains (Fig. 3a, b, d, j, k). When the endolysin contains more than one CBD, which is the case for some GH25 proteins, the enzyme comprises another flexible hinge (Fig. 3a, b, d).

We also find CBD structural homologs. The superimposed structure of the G25 representant that contains two annotated CBDs (Fig. 3a) combined with other GH25 not fully annotated revealed possible LysM structural homologs. These results are represented in Fig. 3b (p value 4.11e −15) and Fig. 3d (p value 1.54e −10). The same structure (Fig. 3a) combined with the shortest GH25 protein (Fig. 3f) does not suggest changes in the structure of the MCD due to the presence of CBDs (p value 5.96e −12). We reached similar conclusions with the SAR endolysins of the GH108 family. This putative protein (Fig. 3k) does not contain an annotated CBD. However, the superimposed structures (Fig. 3l) of this protein and the GH108 combined with PG3 (Fig. 3j) revealed the significance of the detected structural similarities (p value 0.00e + 00). To examine the significance of structural prediction and the pairwise alignment, please access Fig. S2.

The SAR Domain Sometimes Is Also Part of the Endolysin Catalytic Core

To assess the differences in the amino acid sequence and, consequently, in the biochemical properties of the N-terminus of these endolysins, we again aligned the first 40 amino acids of one representative endolysins (Fig. 4a, b). The alignment shows that the region of SAR endolysins is more hydrophobic (Fig. 4b) when compared to the N-terminus of endolysins that do not contain an SAR domain (Fig. 4a).

To assess the structural similarities between SAR endolysins and endolysins that do not contain such domain, we superimposed the structure of representative SAR endolysins with similar endolysins that do not have an SAR domain (Fig. 4c–f). The pairwise alignment revealed a significant structural similarity for all types of catalytic domains. For the proteins of PF01183 (p value 1.11e −16) and PF00182 (p value 4.44e −16) families, the N-terminus portion of the peptides is almost entirely superimposed (Fig. 4c, d). In contrast, in the case of PF00959 MCD (p value 1.59e −14), the N-terminus of the peptides did not align but had a similar folding pattern (Fig. 4e). Furthermore, the N-terminus of the endolysins of the PF05838 family (p value 1.36e −12) is superimposed. Still, the α-helix of the protein that does not contain the SAR domain is slightly shorter (Fig. 4f). To examine the significance of the pairwise alignment, please access Fig. S3.

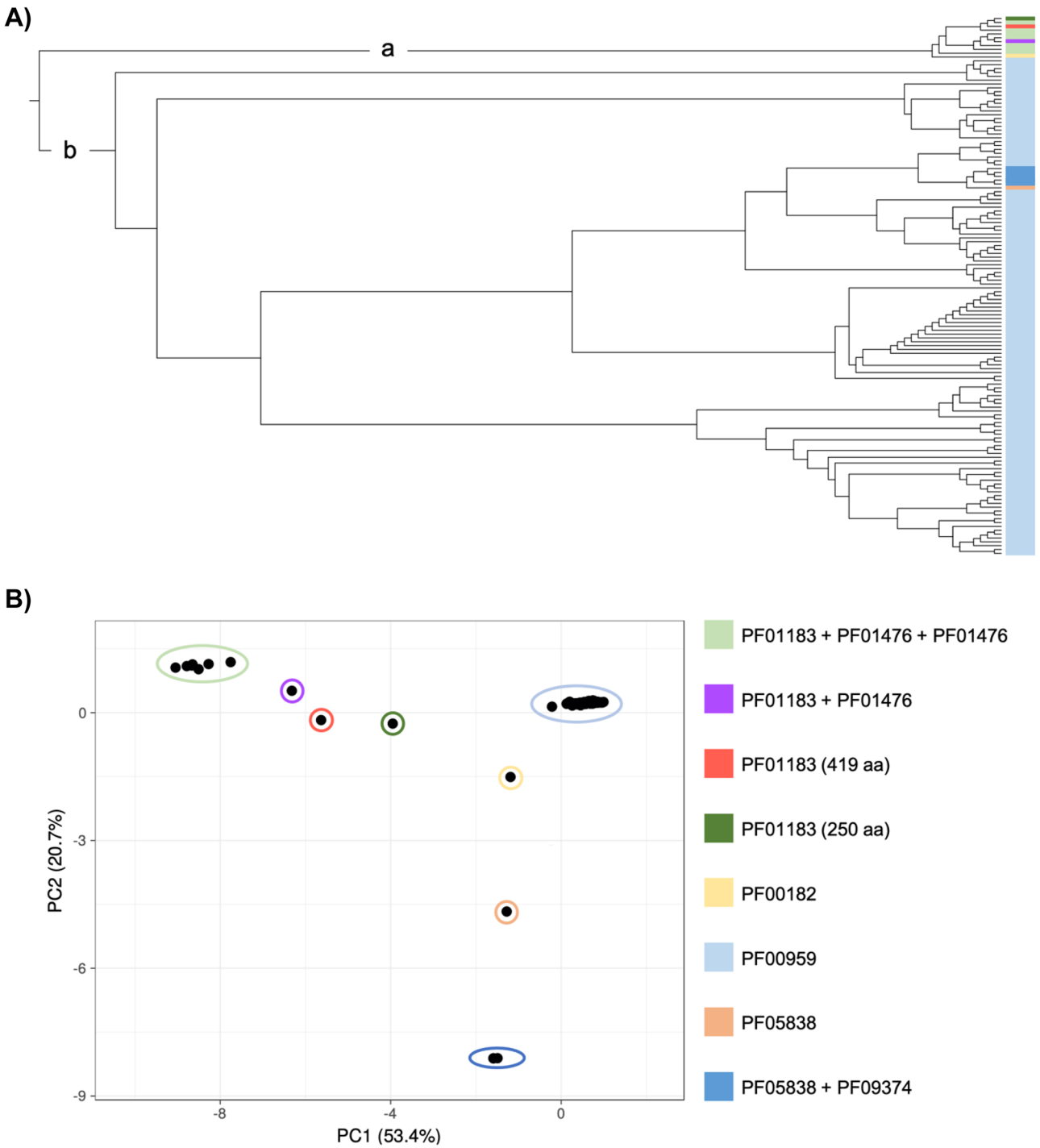
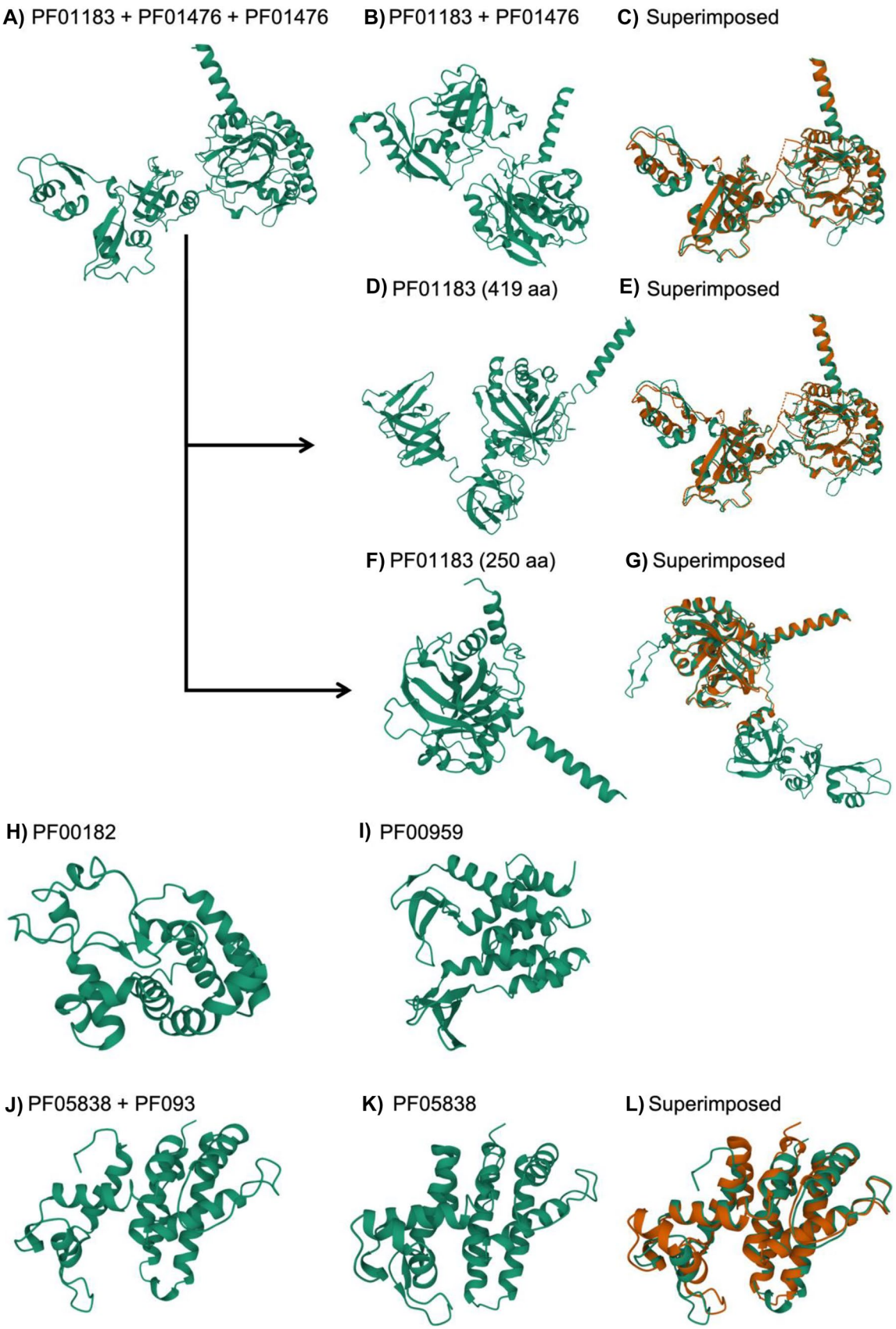


Fig. 2 **a** Phylogenetic tree of all putative SAR endolysins identified in the literature. The tree was constructed using the Neighbor-Joining method, model JTT. **b** Principal component analysis of all putative SAR endolysins identified in the literature. *X*- and *Y*-axis show principal components 1 and 2. The total variance observed for components 1 and 2 was 53.4% and 20.7%, respectively. Ellipses represent putative SAR endolysins with similar structures. $N = 140$ data points

Discussion

The lysis module of bacteriophages has constantly been evolving. Endolysins' signal-anchor-released mode of action is considered the most ancient lysis system [20]. Despite the primitive origin, the lysis module of secreted endolysins, sometimes named holin-independent endolysins, is carried in the genomes of some bacteriophages [9]. However, the


A) PF01183 + PF01476 + PF01476
B) PF01183 + PF01476
C) Superimposed
D) PF01183 (419 aa)
E) Superimposed
F) PF01183 (250 aa)
G) Superimposed
H) PF00182
I) PF00959
J) PF05838 + PF093
K) PF05838
L) Superimposed

◂**Fig. 3** Structural prediction of representative SAR endolysins identified in the literature. Putative SAR endolysins belonging to glycoside hydrolase family 25 (PF001183): **a** GH25 associated with two annotated CBDs LysM; **b** GH25 associated with one LysM domain; **c** superimposed structures of (**a**) and (**b**); **d** long-chain GH25 with no annotated CBDs; **e** superimposed structures of (**a**) and (**d**); **f** short-chain GH25; **g** superimposed structures of (**a**) and (**f**); **h** structure of the glycoside hydrolase family 19 putative SAR endolysins; **i** structure of the glycoside hydrolase family 24 (lysozyme) representative SAR endolysin; putative SAR endolysins belonging to glycoside hydrolase family 108 (PF05838): **j** GH108 associated with one CBD PG3; **k** long-chain GH108 with no annotated CBDs; **l** superimposed structures of (**j**) and (**k**)

evolutionary advantages of this lysis system remain unclear. In the study, we evaluate one type of secreted endolysin containing a particular type of signal peptide named signal-anchor-release (SAR). This type of signal peptide is believed to anchor the endolysin in the cytoplasmatic membrane of bacteria until the lysis-timing of pinholins [9, 10, 45]. We focus on this type of endolysins because the hydrophobic nature of the SAR domain probably confers outer membrane permeation to endolysins [11].

Based on our data, SAR domains are present in about 2.02% of the endolysins encoded in the genomes of uncultured viruses. SAR domains were mainly described in the muramidase superfamily [8, 20, 21]. Our study identified three muramidase domains (PF00959, PF01183, and PF05838). Among the 53 putative SAR endolysins described in this study, five were encoded by bacteriophages that infect Gram-positive bacteria and 30 by phages that act on Gram-negatives.

Oliveira et al. [20] was the first author that identified putative SAR endolysins in viral genomes (Supplementary File 1). The authors also observed a low frequency of SAR endolysins (5.26%; $n=38$) in the genomes of cultured bacteriophages that infect Gram-positive ($n=4$) and Gram-negative bacteria ($n=34$). This database of 38 putative SAR endolysins resulted in a list of 32 non-redundant candidates (Table S6). Among these candidates, Oliveira et al. [20] identified 26 GH24 proteins, four multimodular GH25 proteins, and two multimodular GH108 proteins.

Following the work of Oliveira et al. [20], Valero-Rello [21] identified 33 SAR domains in 8.89% of the putative endolysins annotated in the genomes of *Pseudomonas* bacteriophages (Supplementary File 1). Valero-Rello [21] revealed 21 non-redundant SAR endolysins (Table S7) and identified 19 GH24 proteins, two of which were multimodular and one GH108 putative protein. Interestingly, Valero-Rello [21] also characterized one putative SAR domain in one glucosaminidase GH19. So far, this is the only author that observes SAR domains outside of the muramidase superfamily.

Furthermore, Gontijo et al. [11] also observed a low frequency (5.50%; $n=16$) of SAR endolysins in the genomes of bacteriophages that infect Gram-negative bacteria (Supplementary File 1). The authors identified 13 non-redundant SAR muramidases (Table S8). The screening of SAR endolysin in annotated endolysins resulted in a non-redundant list with 66 candidates. In this work, we describe 53 additional non-redundant SAR endolysins.

The genomes of phages that infect either Gram-positive or Gram-negative bacteria contain secreted endolysins [20, 22]. SAR endolysins, in particular, are majorly found in bacteriophages that infect Gram-negative bacteria [8, 20]. However, despite the low frequency, holin-independent endolysins exist alongside holin-dependent endolysins, which suggests that this primitive lysis system is still effective.

Interestingly, SAR endolysins encoded in the genomes of bacteriophages that infect Gram-positive bacteria have the GH25 catalytic domain associated with the LysM CBD. However, the number of multimodular SAR endolysins is still too scarce to jump to any conclusions. Once the cell membranes are negatively charged, SAR endolysins with higher IP are more recommended to act against Gram-negative bacteria [23]. Interestingly, most putative proteins with IP values inferior to 7.40 are associated with CBDs. Among these 53 candidates, we describe six multimodular SAR endolysins. Previously, two multimodular SAR endolysins were described by Valero-Rello (2019), and 6 by Oliveira et al. [20]. Most of them had IP values inferior to 7.40. Therefore, we speculate that CBDs might help fixate endolysins in the membrane/PG despite the unfavorable net charge.

Endolysins containing catalytic domains combined with CBDs are less frequently found within bacteriophage genomes (< 1%). CBDs are associated with targeting different chemical receptors [20] or with broad lytic activity [12, 46]. The LysM domain, particularly in heterologous endolysins, is related to broad-spectrum activity [12, 47]. This work revealed three uncharacterized CBDs with similar structures to LysM ($n=3$) and PG3 domains ($n=1$). These multimodular SAR endolysins are promising candidates as potential antibacterial drugs against Gram-negative bacteria.

The frequency of SAR endolysins is inferior to 5%. In general, the presence of an SAR domain does not alter the size, the mass, or the IP of endolysins compared to non-SAR-containing proteins. Holin-dependent endolysins are believed to have evolved from SAR lysins [20]. In contrast with Xu et al. [10], we believe that the SAR domain might also be essential in the catalysis of SAR endolysins. Despite the differences in the sequence and the biochemical properties, we speculate that the changes at the N-terminus end of SAR endolysins must have been substantial enough to confer less hydrophobic properties to the endolysin but not enough to change protein folding. The SAR domain is probably also essential in terms of catalytic activity beyond endolysin transport. Lim et al. [19] showed that the native

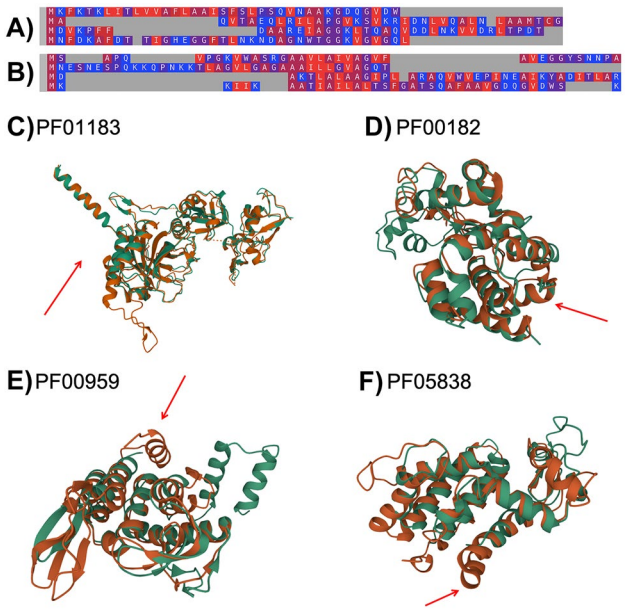


Fig. 4 Multiple sequence alignment of the first 40 amino acids of all SAR endolysins (**a**) and their respective non-SAR endolysin (**b**). The color scale ranges from the most (red) to the least hydrophobic (blue) amino acid. Structural pairwise alignment of representative SAR endolysin with their respective non-SAR endolysin. **c** Glycoside hydrolase family 25 (PF001183) superimposed with the non-SAR representative Lys_ TP901-1 (NP_112716). **d** Glycoside hydrolase family 19 putative SAR endolysins (PF00182) superimposed with the non-SAR representative Lys_vB_AbaP_AS12 (YP_009599231). **e** Glycoside hydrolase family 24 (PF00959) superimposed with the non-SAR representative Lys_YMC-13–01-C62 (YP_009055463). **f** Glycoside hydrolase family 108 (PF05838) superimposed with the non-SAR representative Lys_AbaM-IME-AB2 (YP_009592219). SAR domains are indicated in the red arrows

SPN9CC SAR endolysin is active against Gram-negative and Gram-positive bacteria. In addition, Lim et al. [19] also constructed the SPN9CC endolysin with deletions for the predicted SAR domain. The engineered proteins did not act against *E. coli* even after an EDTA pre-treatment. Thus, the author suggested that the deletion of the SAR influences protein folding and the catalytic activity of the endolysin. In the case of SAR endolysins R^{21}, the SAR domain is also essential to the enzyme's catalytic activity [45].

Nonetheless, Sun et al. [45] demonstrated that for Lyz^{P1}, one SAR endolysin, the SAR domain is not essential for maintaining protein structure or enzymatic activity. However, the results found in the literature are still too scarce to imply the antimicrobial spectra of SAR endolysins. We can only infer that SAR endolysins act against Gram-negative bacteria. Further studies are required to determine the actual action antimicrobial spectra of SAR endolysins, even those encoded by bacteriophages that infect Gram-positive bacteria, and to determine the essential role of the SAR domain in the outer membrane permeation of SAR endolysins.

Conclusion

Our results add 53 putative SAR endolysins to a list that previously contained 66 non-redundant candidates. These additional candidates complete a previously published metagenomic study [22]. The list of putative SAR lysis identified in the literature currently contains 119 non-redundant proteins, including 14 multimodular SAR endolysins. These proteins are clustered into eight groups. Further experimental lab studies must be conducted to assess the efficacy of each of these groups to determine the fundamental role of the SAR domain in the activity of SAR endolysins against Gram-negative bacteria.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12602-022-09948-y.

Acknowledgements MTPG is supported by the Brazilian funding agency *Fundação de Amparo a Pesquisa do Estado de São Paulo* (FAPESP) through an M.Sc. scholarship (grant 2020/01535-9). MPT is also supported by FAPESP through a scientific initiation scholarship (grant 2020/09815-0). Research in our laboratory is supported by FAPESP (grant 2021/00465-0). MB is a research fellow of the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq – process number: 309380/2019-7).

Author Contribution MTPG, MPT, and PMPV conceived and designed the study, performed the analysis, and interpreted the data. MTPG wrote the manuscript. MB contributed to finalizing the manuscript and coordinated the study. All authors read and approved the final text.

Data Availability All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Declarations

Competing Interests The authors declare no competing interests.

References

1. Chanishvili N (2012) Phage therapy–history from Twort and d'Herelle through Soviet experience to current approaches. Adv Virus Res 83:3–40. https://doi.org/10.1016/B978-0-12-394438-2.00001-3
2. Du Toit A (2017) The language of phages. Nat Rev Microbiol 15:135–135. https://doi.org/10.1038/nrmicro.2017.8
3. Chan BK, Abedon ST (2015) Bacteriophages and their enzymes in biofilm control. Curr Pharm Des 21:85–99. https://doi.org/10.2174/1381612820666140905112311
4. Lin DM, Koskella B, Lin HC (2017) Phage therapy: an alternative to antibiotics in the age of multi-drug resistance. World J Gastrointest Pharmacol Ther 8:162–173. https://doi.org/10.4292/wjgpt.v8.i3.162
5. Principi N, Silvestri E, Esposito S (2019) Advantages and limitations of bacteriophages for the treatment of bacterial infections. Front Pharmacol. https://doi.org/10.3389/fphar.2019.00513

6. Hobbs Z, Abedon ST (2016) Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic.' FEMS Microbiol Lett. https://doi.org/10.1093/femsle/fnw047
7. Baliga P, Goolappa PT, Shekar M, Kallappa GS (2022) Cloning, characterization, and antibacterial properties of endolysin LysE against planktonic cells and biofilms of Aeromonas hydrophila. Probiotics Antimicrob Proteins. https://doi.org/10.1007/s12602-021-09880-7
8. Gontijo MTP, Vidigal PMP, Lopez MES, Brocchi M (2021) Bacteriophages that infect Gram-negative bacteria as source of signal-arrest-release motif lysins. Res Microbiol 172:103794. https://doi.org/10.1016/j.resmic.2020.103794
9. Young R (2014) Phage lysis: three steps, three choices, one outcome. J Microbiol 52:243–258. https://doi.org/10.1007/s12275-014-4087-z
10. Xu M, Struck DK, Deaton J et al (2004) A signal-arrest-release sequence mediates export and control of the phage P1 endolysin. Proc Natl Acad Sci USA 101:6415–6420. https://doi.org/10.1073/pnas.0400957101
11. Gontijo MTP, Jorge GP, Brocchi M (2021) Current status of endolysin-based treatments against Gram-negative bacteria. Antibiotics 10:1143. https://doi.org/10.3390/antibiotics10101143
12. Briers Y, Schmelcher M, Loessner MJ et al (2009) The high-affinity peptidoglycan binding domain of Pseudomonas phage endolysin KZ144. Biochem Biophys Res Commun 383:187–191. https://doi.org/10.1016/j.bbrc.2009.03.161
13. Sekiya H, Kamitori S, Nariya H et al (2021) Structural and biochemical characterization of the Clostridium perfringens-specific Zn2+-dependent amidase endolysin, Psa, catalytic domain. Biochem Biophys Res Commun 576:66–72. https://doi.org/10.1016/j.bbrc.2021.08.085
14. Simmons M, Morales CA, Oakley BB, Seal BS (2012) Recombinant expression of a putative amidase cloned from the genome of Listeria monocytogenes that lyses the bacterium and its monolayer in conjunction with a protease. Probiotics & Antimicro Prot 4:1–10. https://doi.org/10.1007/s12602-011-9084-5
15. Hosseini ES, Moniri R, Goli YD, Kashani HH (2016) Purification of antibacterial CHAPK protein using a self-cleaving fusion tag and its activity against methicillin-resistant Staphylococcus aureus. Probio Antimicro Prot 8:202–210. https://doi.org/10.1007/s12602-016-9236-8
16. Xu M, Arulandu A, Struck DK et al (2005) Disulfide isomerization after membrane release of its SAR domain activates P1 lysozyme. Science. https://doi.org/10.1126/science.1105143
17. Park T, Struck DK, Dankenbring CA, Young R (2007) The pinholin of lambdoid phage 21: control of lysis by membrane depolarization. J Bacteriol 189:9135–9139. https://doi.org/10.1128/JB.00847-07
18. Tran TAT, Struck DK, Young R (2007) The T4 RI antiholin has an N-terminal signal anchor release domain that targets it for degradation by DegP. J Bacteriol. https://doi.org/10.1128/JB.00854-07
19. Lim J-A, Shin H, Heu S, Ryu S (2014) Exogenous lytic activity of SPN9CC endolysin against gram-negative bacteria. J Microbiol Biotechnol 24:803–811. https://doi.org/10.4014/jmb.1403.03035
20. Oliveira H, Melo LDR, Santos SB et al (2013) Molecular aspects and comparative genomics of bacteriophage endolysins. J Virol 87:4558–4570. https://doi.org/10.1128/JVI.03277-12
21. Valero-Rello A (2019) Diversity, specificity and molecular evolution of the lytic arsenal of Pseudomonas phages: in silico perspective. Environ Microbiol 21:4136–4150. https://doi.org/10.1111/1462-2920.14767
22. Fernández-Ruiz I, Coutinho FH, Rodriguez-Valera F (2018) Thousands of novel endolysins discovered in uncultured phage genomes. Front Microbiol. https://doi.org/10.3389/fmicb.2018.01033
23. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R (2013) Expanding the marine virosphere using metagenomics. PLoS Genet 9:e1003987. https://doi.org/10.1371/journal.pgen.1003987
24. Mizuno CM, Ghai R, Saghaï A et al (2016) Genomes of abundant and widespread viruses from the deep ocean. mBio 7. https://doi.org/10.1128/mBio.00805-16
25. Roux S, Brum JR, Dutilh BE et al (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689–693. https://doi.org/10.1038/nature19366
26. Coutinho FH, Silveira CB, Gregoracci GB et al (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun 8:15955. https://doi.org/10.1038/ncomms15955
27. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA et al (2016) Uncovering Earth's virome. Nature 536:425–430. https://doi.org/10.1038/nature19094
28. Roux S, Hallam SJ, Woyke T, Sullivan MB (2015) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. eLife 4:e08490. https://doi.org/10.7554/eLife.08490
29. Hirokawa T, Boon-Chieng S, Mitaku S (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14:378–379. https://doi.org/10.1093/bioinformatics/14.4.378
30. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580. https://doi.org/10.1006/jmbi.2000.4315
31. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016
32. Hennerdal A, Elofsson A (2011) Rapid membrane protein topology prediction. Bioinformatics 27:1322–1323. https://doi.org/10.1093/bioinformatics/btr119
33. Kozlowski LP (2021) IPC 2.0: prediction of isoelectric point and pKa dissociation constants. Nucleic Acids Res 49:W285–W292. https://doi.org/10.1093/nar/gkab295
34. Zimmermann L, Stephens A, Nam S-Z et al (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J Mol Biol 430:2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007
35. Mistry J, Chuguransky S, Williams L et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419. https://doi.org/10.1093/nar/gkaa913
36. Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160–1166. https://doi.org/10.1093/bib/bbx108
37. Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190. https://doi.org/10.1101/gr.849004
38. Metsalu T, Vilo J (2015) ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic Acids Res 43:W566–W570. https://doi.org/10.1093/nar/gkv468
39. Han MV, Zmasek CM (2009) phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics 10:356. https://doi.org/10.1186/1471-2105-10-356
40. Takahashi D, Fujiwara I, Miyata M (2020) Phylogenetic origin and sequence features of MreB from the wall-less swimming bacteria Spiroplasma. Biochem Biophys Res Commun 533:638–644. https://doi.org/10.1016/j.bbrc.2020.09.060
41. Baek M, DiMaio F, Anishchenko I et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876. https://doi.org/10.1126/science.abj8754

42. Baek M, Baker D (2022) Deep learning and protein structure modeling. Nat Methods 19:13–14. https://doi.org/10.1038/s41592-021-01360-8
43. Li Z, Jaroszewski L, Iyer M et al (2020) FATCAT 2.0: towards a better understanding of the structural diversity of proteins. Nucleic Acids Res 48:W60–W64. https://doi.org/10.1093/nar/gkaa443
44. Sehnal D, Bittrich S, Deshpande M et al (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res 49:W431–W437. https://doi.org/10.1093/nar/gkab314
45. Sun Q, Kuty GF, Arockiasamy A et al (2009) Regulation of a muralytic enzyme by dynamic membrane topology. Nat Struct Mol Biol 16:1192–1194. https://doi.org/10.1038/nsmb.1681
46. Huang Y, Yang H, Yu J, Wei H (2015) Molecular dissection of phage lysin PlySs2: integrity of the catalytic and cell wall binding domains is essential for its broad lytic activity. Virol Sin 30:45–51. https://doi.org/10.1007/s12250-014-3535-6
47. Hu S, Kong J, Kong W et al (2010) Characterization of a novel LysM domain from Lactobacillus fermentum bacteriophage endolysin and its use as an anchor to display heterologous proteins on the surfaces of lactic acid bacteria. Appl Environ Microbiol 76:2410–2418. https://doi.org/10.1128/AEM.01752-09