Discovery and Characterization of the Phospholemman/SIMP/Viroporin Superfamily

Daniel Tyler · Kevin J. Hendargo · Arturo Medrano-Soto · Milton H. Saier

Department of Molecular Biology, Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA

Keywords

Superfamily · Viroporin · Phospholemman · Endoregulin · Myoregulin · Sarcolipin

Abstract

Using bioinformatic approaches, we present evidence of distant relatedness among the Ephemerovirus Viroporin family, the Rhabdoviridae Putative Viroporin U5 family, the Phospholemman family, and the Small Integral Membrane Protein family. Our approach is based on the transitivity property of homology complemented with five validation criteria: (1) significant sequence similarity and alignment coverage, (2) compatibility of topology of transmembrane segments, (3) overlap of hydropathy profiles, (4) conservation of protein domains, and (5) conservation of sequence motifs. Our results indicate that Pfam protein domains PF02038 and PF15831 can be found in or projected onto members of all four families. In addition, we identified a 26-residue motif conserved across the superfamily. This motif is characterized by hydrophobic residues that help anchor the protein to the membrane and charged residues that constitute phosphorylation sites. In addition, all members of the four families with annotated function are either responsible

for or affect the transport of ions into and/or out of the cell. Taken together, these results justify the creation of the novel Phospholemman/SIMP/Viroporin superfamily. Given that transport proteins can be found not just in cells, but also in viruses, the ability to relate viroporin protein families with their eukaryotic and bacterial counterparts is an important development in this superfamily.

© 2022 The Author(s).
Published by S. Karger AG, Basel

Introduction

As of July 16, 2021, the RefSeq database [O'Leary et al., 2016] at the National Center for Biotechnology Information (NCBI) contains more than 209 million non-redundant proteins from 112,462 organisms. Of these proteins, more than 21,400 can be found in the Transporter Classification Database (TCDB; <https://tcdb.org>), a database of representative transport-related proteins maintained by the Saier research group [Saier et al., 2021]. The proteins found in TCDB can be further classified into more than 1,600 transporter families, which use the IUBMB-approved Transporter Classification (TC) system [Chang et al., 2004] to provide curated annotations. Given that

several pandemics such as the Spanish flu, HIV/AIDS, swine flu, and most recently COVID-19 are all viral in nature, the study of transport proteins (vioporins), which are essential in several steps on the viral life cycle [Sze and Tan, 2015; To and Torres, 2018; Ketter and Randall, 2019; Wong and Saier, 2021] might shed more light onto how current and future candidate viruses responsible for pandemics operate.

Vioaporins are small proteins, usually ranging in size from about 60 to 120 amino acid residues (aas), which can be found in cell membranes and help facilitate the passing of viruses both into and out of the cell [Gonzalez and Carrasco, 2003]. They are also responsible, at least in part, for cell membrane leakiness following viral infection due to their ability to alter membrane permeability [Carrasco, 1978; Lama and Carrasco, 1992]. Associated with membrane leakiness is the ability of viruses to disrupt normal ion flow across membranes and induce permeability to small molecules which would otherwise be kept out of the cell [Hyser and Estes, 2015]. Currently, there are 21 families of vioporins in TCDB while several other families display vioporin-like features. A TCDB-wide search revealed that only the four families in this study satisfied all of our criteria to infer homology.

Bovine ephemeral fever rhabdovirus (BEFV) $\alpha 1$ proteins belong to the Ephemerovirus Vioaporin (EVVP) family (TC: 1.A.95) and can be found in ephemeroviruses such as the Koolpinyah and Yata viruses, which can infect cattle [Blasdel et al., 2014]. Proteins found in the EVVP family range in size from 85 to 120 aas and have one transmembrane segment (TMS). TMSs are composed mainly of hydrophobic aas that anchor proteins into cellular membranes and are interspersed with functional residues that can interact with substrate(s) and maintain specificity. TMSs are characterized by their α -helical structures and rely on amino acid composition, rather than translocation recognition [De Marothy and Elofsson, 2015]. Evidence for the vioporin characteristics of EVVP family members comes from studies showing that BEFV-infected cells expressing $\alpha 1$ show increased membrane permeability in mammals and both compromised membrane permeability and inhibited cell growth in *Escherichia coli* [Joubert et al., 2014].

Members of the Rhabdoviridae Putative Vioaporin U5 (RV-U5) family (TC: 1.A.100) also have one TMS; their lengths range from 90 to 127 amino acids, but not much is currently known about their functions. The Wongabel virus has five genes N, P, M, G, and L, which are typically found in rhabdoviruses. Genes N and G show overlap with two open reading frames known as U4 and U5, de-

noted as such due to their unidentified functions. U5 gene products are predicted to have structures similar to the $\alpha 1$ proteins of the EVVP family, providing evidence for their status as vioporins [Gubala et al., 2008].

Sharing a similar range of about 65–190 amino acids and one TMS, characteristic of EVVP and RV-U5 family members are the Small Integral Membrane Protein (SIMP) Family (TC: 1.A.113), found mainly in eukaryotes (systems 1.A.113.1-5), but also in bacteria (e.g., systems 1.A.113.5.4 and 1.A.113.5.5). While some members remain uncharacterized, systems in subfamilies 1 and 5 have been identified as being either endoregulin (ELN; TC: 1.A.113.5.1) or ELN homologs [Anderson et al., 2016]. ELN is a small integral membrane protein designated as a “micropeptide,” that plays an important role in the regulation of calcium-dependent signaling by inhibiting the sarco/endoplasmic reticulum Ca^{2+} -ATPase (SERCA), the membrane pump responsible for muscle relaxation via Ca^{2+} uptake into the sarcoplasmic reticulum [Anderson et al., 2016]. Three micropeptides, myoregulin (MLN), phospholamban (PLN), and sarcolipin, are muscle-specific regulators of SERCA, while ELN and another-regulin (ALN) are responsible for regulation of nonmuscle cell SERCA. In experiments where ELN was expressed with the muscle cell micropeptide, PLN, PLN competes with ELN for binding to a single site on the surface of SERCA [Anderson et al., 2016]. ELN is physiologically important, given SERCA’s significant role in metabolism, cell growth, and cell death pathways in multiple cell types [Anderson et al., 2016].

The Phospholemman (PLM) family (TC: 1.A.27) contains at least seven FXYD proteins split into three subfamilies [Jespersen et al., 2006]. These proteins range in size from 60 to 180 aas and similarly, each contains a single TMS. The primary purpose of all PLM family proteins is regulation of ion channels in eukaryotes [Zhang et al., 2015]. In addition to evoking Cl^- and Na^+ conductances in most tissues [Garty and Karlish, 2006], and altering the V_{max} of the Na,K-ATPase [Lubarski et al., 2005], members of the PLM family can function as cation-selective channels. Mutations in these channels are linked to dominant renal hypomagnesemia in humans [Sha et al., 2008].

The identification of a novel superfamily (a group of transport families with a common evolutionary origin) allows for a better understanding of the structures and functions of their less well-studied members. Here, we report the evidence supporting the relationships among these 4 families and thus their incorporation into the new Phospholemman/SIMP/Vioaporin (PSV) superfamily.

Fig. 1. Flowchart of the strategy to infer distant evolutionary relationships between two families. The figure summarizes the main steps in our strategy. See text for discussion.

We applied the transitivity property of homology (if A is a homolog of B, B is a homolog of C, and C is a homolog of D, then A is a homolog of D), which has been successfully used for the identification of distant evolutionary relationships [Medrano-Soto et al., 2018; Medrano-Soto et al., 2020; Wang et al., 2020]. Briefly, our criteria consist of (1) significant sequence similarity and alignment coverage across the homology transitivity path, (2) compatible TMS topologies, (3) similar hydropathy profiles, (4) shared sequence motifs, and (5) conserved domains. Unfortunately, comparison of 3D structures was not possible because at the time of this study, there were no structures available in the Protein Data Bank (PDB) for members of any of these families.

Results

Inference of Homology

Following previously published methodologies [Medrano-Soto et al., 2018; Medrano-Soto et al., 2020; Wang et al., 2020], we used the transitivity principle of homology to infer distant relationships between members of the four families in this study. In this strategy, two proteins A and D, with no obvious sequence similarity, are candidate homologs if two additional proteins B (homolog of A) and C (homolog of D) can be identified such that a clear path of significant sequence similarity and other functional/structural properties can be traced connecting proteins A and D ($A \rightarrow B \rightarrow C \rightarrow D$). Homology is then inferred by association between the two families to which proteins

Fig. 2. Evidence of homology between families RV-U5 and SIMP. Hydropathy plots are presented across the homology transitivity path between members of families RV-U5 (TC: 1.A.100) and SIMP (TC: 1.A.113). Hydropathy values are computed and plotted as described in Methods. Panels **a–c** depict relationships within family RV-U5. Panels **d–f** depict relationships within the SIMP family, and panel **g** shows the evidence supporting homology between these two families. Orange and cyan bars denote hydrophobic peaks (i.e., inferred TMSs). Direct Pfam hits or projected Pfam domains (see methods) are shown as solid or dashed colored horizontal bars, respectively. Thin vertical black lines with wedges delimit the region of a protein involved in an alignment. The wedges in panels **a** and **d** delimit the regions covered by the alignments shown in panels **b** and **e**, relative to the full-length proteins in panels **a** and **d**, respectively. Proteins in panels **c** and **f** have two sets of delimiting wedges. Wedges plotted for positive hydropathy values delimit regions covered by the alignments in panel **b** and **e** relative to the full-length proteins in panels **c** and **f**, respectively. Wedges plotted for negative hydropathy values delimit regions covered by

the alignment in panel **g** relative to the full proteins in panels **c** and **f**, respectively. Interruptions in the hydropathy curves of panels **b**, **e**, and **g**, indicate gaps in the corresponding sequence alignments. **a** Hydropathy plot of RV-U5 member B2X7D9 (TC: 1.A.100.1.1). **b** Hydropathy plot of the alignment (E-value: 9.4×10^{-31}) between RV-U5 member B2X7D9 and its homolog ASM90778. **c** Hydropathy plot of RV-U5 homolog ASM90778. **d** Hydropathy plot of SIMP member XP_012787199 (TC: 1.A.113.1.4). **e** Hydropathy plots of the alignment (E-value: 1.5×10^{-16}) between SIMP member XP_012787199 and its homolog XP_006204489. **f** Hydropathy plot of SIMP homolog XP_006204489. **g** Hydropathy plots of the alignment (E-value: 3.0×10^{-7}) between RV-U5 homolog ASM90778 and SIMP homolog XP_006204489. Only the regions where hydrophobic peaks overlap are highlighted in the alignments. The Pfam domain PF15831, characteristic of the SIMP family, can be projected (E-value: 8.1×10^{-7}) to RV-U5 homolog ASM90778 (panel **c**; see Methods), providing further support for homology between these families.

Fig. 3. Evidence of homology between families EVVP and PLM. Hydropathy plots show the relationship between members of families EVVP (TC: 1.A.95) and PLM (TC: 1.A.27). Hydropathy values are computed and plotted as described in Methods. Although this figure contains just three panels, refer to the legend of Figure 2 for a description of the general format of hydropathy plots. In this case, thin vertical black lines with wedges in panels **a** and **b** delimit the region of each protein involved in the alignment shown in panel **c**. **a** Hydropathy plot of EVVP member AJR28579 (TC: 1.A.95.2.3). **b** Hydropathy plot of PLM homolog XP_027999365.

Note that the alignment between PLM member P58549 (TC: 1.A.27.1.7) and its homolog XP_027999365 is highly significant (E-value: 2.1×10^{-37}) and has 90% coverage (not shown). **c** Hydropathy plot of the alignment (E-value: 9.1×10^{-6}) between EVVP member AJR28579 and PLM homolog XP_027999365. Only the region where hydrophobic peaks overlap is highlighted in the alignment. In support of this relationship, the Pfam domain PF02038 characteristic of family PLM can be projected with an E-value of 2×10^{-3} to the EVVP member AJR28579 (dashed blue line in panel **a**; see Methods).

A and D belong. Given that sequence similarity alone may not be enough to conclude homology when dealing with integral membrane proteins [Wong et al., 2010; Wong et al., 2011; Wong et al., 2012], the following 4 additional criteria were applied to support the inference of homology: (a) similar hydropathy profiles, which alone constitutes weak evidence of homology; (b) similar TMS topology; (c) shared motifs; and (d) shared domains. Normally, this methodology also includes the comparison of 3D structures, but no structures were available for these families at the time of this study. A summary of this strategy can be found in Figure 1.

Relationship between the RV-U5 and the SIMP Families

Figure 2 shows comparisons across the homology transitivity path between families RV-U5 (TC: 1.A.100; panel a) and SIMP (TC: 1.A.113; panel d). Evidence of homology is presented in panel g with the alignment (E-value of 3.0×10^{-7}) between RV-U5 homolog ASM90778

(panel c) and SIMP homolog XP_006204489 (panel f). The transmembrane regions of the two homologs show good hydropathy overlap (panels b, e, and g). The Pfam domain PF15831 of unknown function found in the SIMP family can be projected (see Methods) onto the RV-U5 homolog ASM90778 with an E-value of 8.1×10^{-7} further supporting the relationship. Conserved motifs across all families in this study will be discussed in section Motif Analysis of the PSV superfamily below.

Relationship between the EVVP and the PLM Families

Evidence of homology between families EVVP (TC: 1.A.95) and PLM (TC: 1.A.27) is presented in Figure 3. In this case, no homolog of the EVVP family is necessary to detect the relationship, homology can be inferred directly by aligning EVVP member 1.A.95.2.3 (panel a) with PLM homolog XP_027999365 (panel b) as shown in panel c (E-value: 9.1×10^{-6}). The alignment shows hydropathy overlap between the transmembrane regions of these two proteins. The Pfam domain PF02038 characteristic of

Fig. 4. Evidence of homology between members of families EVVP and RV-U5. Hydropathy plots show the relationship between families EVVP (TC: 1.A.95) and RV-U5 (TC: 1.A.100). Hydropathy values are computed and plotted as described in Methods. Refer to the legend of Figure 3 for a description of the format. **a** Hydropathy plot of EVVP member AJR28579 (TC: 1.A.95.2.3). **b** Hydropathy

plot of RV-U5 homolog ASM90778. Note that the alignment between RV-U5 member B2X7D9 (TC: 1.A.100.1.1) and its homolog ASM90778 is appreciable (E-value: 1.4×10^{-31}) with 99% coverage (not shown). **c** Hydropathy plot of the alignment (E-value: 2.9×10^{-7}) between EVVP member AJR28579 and RV-U5 homolog ASM90778. These families have no matches with Pfam domains.

the PLM family can be projected onto the EVVP member 1.A.95.2.3 with a marginal E-value of 2×10^{-3} (see Methods). Proteins containing this domain may act as ion channels, modulators of ion channels, or as tissue-specific regulators of the Na,K-ATPase [Crambert and Geering, 2003; Teriete et al., 2007; Zhang et al., 2015], a trait also found in viroporins [Hyser and Estes, 2015]. The projection of this domain agrees with the viroporin characteristics of EVVP and suggests an ion-channel (regulatory) role.

Relationship between the EVVP and RV-U5 Families

Figure 4 shows the comparison between members of families EVVP (TC: 1.A.95) and RV-U5 (TC: 1.A.100). As in the case of the relationship between families EVVP and PLM (see Fig. 2), the evidence of homology is identified in the alignment (panel c) between EVVP family member 1.A.95.2.3 (panel a) and RV-U5 homolog ASM90778 (panel b). The E-value of the alignment is 2.9×10^{-7} , and good hydropathy overlap between the transmembrane regions of the two proteins can be observed. No Pfam domains have been identified for these two families to date.

Other comparisons among the four families that provide further support for the creation of the VSP Superfamily can be observed in supplementary Figures S1–S3. Table 1 presents the E-values across the homology transitivity paths for all possible pairwise comparisons between the 4 families (Fig. 2–4 and supplementary Fig. S1–S3). Shared motifs are discussed below in section Motif Analysis of the PSV Superfamily.

Motif Analysis of the PSV Superfamily

Following the strategy described in Methods and based on MEME [Bailey et al., 2015], the analysis of members of the four PLM, EVVP, RV-U5, and SIMP families led to the identification of a motif 26 aa long, hereafter referred to as the PSV motif (Fig. 5). As a benchmark, the PSV motif was scanned against a negative control set consisting of members of the PLB family (TC: 1.A.50) and all other families in TCDB. We consider family PLB a good negative control because despite having no detectable sequence similarity to the 4 families in the PSV superfamily, it also consists of small proteins with 1 or 2 TMSs that are involved in ion channel activity and in regulation of

Fig. 5. Sequence logo of the PSV superfamily motif. The model of the motif was generated using the MEME program (see Methods) based on sequences from members of the PLM, EVVP, RV-U5, and SIMP families. The level of conservation of each amino acid within the motif is proportional to the size of the one-letter amino acid abbreviation code at any one position. The MEME model of the PSV motif is available in supplementary File S1 (FigShare: <https://doi.org/10.6084/m9.figshare.16570752>).

Table 1. Summary of alignment scores across the homology transitivity path between pairs of families in the new PSV superfamily

A (TCID)	B (accession)	C (accession)	D (TCID)	AB	BC	CD
1.A.100.1.1	ASM90778	XP_006204489	1.A.113.1.4	9.4×10^{-31}	3.0×10^{-07}	1.5×10^{-16}
1.A.95.2.3	–	XP_027999365	1.A.27.1.7	–	9.1×10^{-06}	2.1×10^{-37}
1.A.95.2.3	–	ASM90778	1.A.100.1.1	–	2.9×10^{-07}	1.4×10^{-31}
1.A.100.1.3	YP_009177198	XP_023395639	1.A.27.1.7	7.0×10^{-30}	8.1×10^{-06}	$1.3 \times 10^{-30*}$
1.A.113.1.4	XP_019515853	XP_013860682	1.A.27.1.3	1.4×10^{-16}	7.3×10^{-07}	$2.8 \times 10^{-30†}$
1.A.95.1.2	YP_009094430	XP_020664264	1.A.113.1.1	8.8×10^{-34}	1.8×10^{-05}	$3.2 \times 10^{-22‡}$

Columns A and D indicate the established members of the families being compared. Columns B and C show the homologs of families in columns A and D, respectively. Columns AB, BC, and CD show the E-value scores across the transitivity path A→B→C→D. Column BC represents the link between the two families being compared. Double dashes indicate that TCDB family member 1.A.95.2.3 was used directly in the comparisons (i.e., proteins A and B are the same), and thus scores for self-alignments are not shown. These sequences also have similar topologies, hydrophathy profiles, conserved Pfam domains that can be projected among one another, and as discussed in section Motif Analysis of the Phospholemman/SIMP/Viroporin (PSV) superfamily, a conserved sequence motif across the superfamily. * See supplementary Figure S1. † See supplementary Figure S2. ‡ See supplementary Figure S3.

P-type ATPases (TC: 3.A.3). No significant matches were identified within these negative controls. The PSV motif is contained within Pfam domain PF02038, characteristic of the PLM family, recognizable by the glycyl residues at motif residue positions 2 and 13, the serine at position 19, and the series of charged residues from positions 20 to 25 [Delprat et al., 2006]. In domain PF02038, a mutation at PSV motif position 13 from glycine to arginine results in primary hypomagnesaemia in humans, wherein Ca^{2+} excretion is disrupted [Meij et al., 2000], which suggests that

the PSV motif denotes ion transport or regulation. Lysine, a functionally basic amino acid similar to arginine, which can also be found at motif position 13, may contribute to EVVP's role as a viroporin, disrupting not-yet-specified ion transport, given the similarity to class IA viroporins [Joubert et al., 2014; Hyser and Estes, 2015]. The nonpolar residues from motif positions 3 to 18 have been identified as membrane anchors for members of the SIMP [Anderson et al., 2016] and EVVP [McWilliam et al., 1997] families. The series of charged residues at the

Fig. 6. Conservation of the PSV motif across the superfamily. Results from scanning the PSV motif (Fig. 5) across the superfamily using MAST are shown (see Methods). For simplicity, the figure shows matches and their scores for a sample of 20 sequences (5 per family). However, MAST results throughout the superfamily, including hundreds of homologs, are available in supplementary File S1 (FigShare: <https://doi.org/10.6084/m9.figshare.16570752>).

end of the motif is also indicative of phosphorylation sites in both PLM and EVVP [Delprat et al., 2006; Joubert et al., 2014]. MAST [Bailey et al., 2015] results, identifying the PSV motif in members of all four families, are shown in Figure 6.

Protein Tree of Candidate Superfamily Members

We first attempted to build a phylogenetic tree of the four families using a Bayesian Approach (see Methods). Although the topology of the tree (suppl. Fig. S4) supports the integrity of the families and reveals their overall relationships, the sequence divergence among families

yielded weak statistical support (posterior probability <0.8) for branches that describe within-family relationships for the SIMP and PLM families, as well as the branch shared by the EVVP and RV-U5 families. To test the reliability of the topology observed in the phylogenetic tree, we applied an independent method using the program mkProteinClusters [Medrano-Soto et al., 2018] that clusters sequences based on Smith-Waterman bit scores between pairs of sequences (see Methods). The radial tree in Figure 7 has strong clustering structure (agglomerative coefficient: 0.949) and shows the same overall topology and family relationships as the phylogenetic tree. That is,

Fig. 7. Protein tree of the PLM (TC: 1.A.27), EVVP (TC: 1.A.95), RV-U5 (TC: 1.A.100), and SIMP (TC: 1.A.113) families. The leaves of the tree show the TCID for each system. Family PLM is highlighted in red, EVVP in green, RV-U5 in brown, and SIMP in blue. As described in Methods, the scale bar is not shown because only the topology of the tree is meaningful. Notice that this tree shows the same family groupings than the phylogenetic tree in supplementary Figure S4.

members of families PLM and SIMP form monophyletic groups, whereas families EVVP and RV-U5 share a major branch. Sequence alignments show that apart from recognizing each other, systems 1.A.95.2.1, 1.A.95.2.2, and

1.A.95.2.3 also match members of the RV-U5 with marginal scores, thus providing an explanation for the closer relationship shown in both trees between these two families.

Discussion

In accordance with the methodology presented in Figure 1, our analysis of the families PLM (TC: 1.A.27), EVVP (TC: 1.A.95), RV-U5 (TC: 1.A.100), and SIMP (TC: 1.A.113) has led to the formation of the PSV superfamily. One of the first superfamilies to contain both eukaryotic viral and cellular members, the PSV superfamily members are generally single-TMS transport proteins (Fig. 2–4 and suppl. Fig. S1–S3). Currently, all members of the PSV superfamily with described function are either responsible for or affect the transport of ions into and out of the cell, or both. As such, proteins without defined function may be assumed to play similar roles in regulation of and/or transport of ions.

Further supporting the idea of being related to ion transport, protein Pfam domains PF02038 and PF15831 can either be found in or projected onto members of all families. PF02038, the characteristic domain of family PLM [Delprat et al., 2006], is associated with ion channel regulation [Crambert and Geering, 2003; Zhang et al., 2015], and its overlap with PF15831 (suppl. fig. s2) strengthens the argument that members of the EVVP, RV-U5, and SIMP families serve similar roles. Additionally, the PSV motif (Fig. 5) is contained within PF02038, with the same or chemically similar residues at positions relevant for function.

Given that the criteria of sequence similarity, compatibility of TMS topologies, overlap of hydropathy profiles, conservation of domains and shared motifs can be met for all pairs of families, we concluded that the creation of the PSV Superfamily is justified. Currently, no other family in TCDB meets the requirements laid out in Figure 1 to be a member of the superfamily. All in all, the identification of the PSV superfamily is an exciting addition to transport protein biology, as it contributes to the discussion regarding the origins of proteins both vital to human/animal life and virus survival, even if the two purposes can be at odds with each other.

The existence of a superfamily including pore-forming proteins from both animal cells and viruses is not without precedence. For example, bacterial holin proteins are encoded within bacteriophage and bacterial genomes [Reddy and Saier, 2013; Saier and Reddy, 2015], and both voltage-gated ion channels and aquaporins are found encoded within eukaryotic viruses as well as the genomes of the cells which they infect [Gazzarrini et al., 2006; Thiel et al., 2011; Sze and Tan, 2015]. We anticipate that such relationships will prove to be much more common than is currently recognized.

Materials and Methods/Experimental Procedures

All programs developed by the Saier laboratory are available in the public software repository: <https://github.com/SaierLaboratory>.

Obtaining Candidate Homologs

Candidate homologs for each member of the families involved in this study were retrieved from NCBI using the program famXpander [Medrano-Soto et al., 2018]. This program relies on BLAST [Altschul et al., 1997] to extract homologs from the NCBI non-redundant database. Up to 10,000 matches per query sequence were extracted showing E-values $<10^{-6}$ and a minimal alignment coverage of 80% of the smaller proteins to allow for potential fusions. Redundant sequences were removed with the program CD-HIT [Fu et al., 2012] using an identity threshold of 90%.

Comparison of Homologs between Pairs of Families

Sets of candidate homologs, retrieved with famXpander from the NCBI database, were compared between pairs of families with the program areFamiliesHomologous [Medrano-Soto et al., 2018; Medrano-Soto et al., 2020], which integrates into a pipeline several of our other methods including Protocol2 [Reddy and Saier, 2012], GSAT [Reddy and Saier, 2012], QUOD [Medrano-Soto et al., 2020], and HVORDAN [Medrano-Soto et al., 2020] in order to apply the transitivity principle of homology between families and infer distant evolutionary relationships. Final alignments were generated with the Smith-Waterman algorithm as implemented in SSEARCH [Pearson, 1991]. Hydropathy profiles of proteins and alignments across the homology transitivity path were generated with the programs QUOD and HVORDAN [Medrano-Soto et al., 2020]. These programs are based on and extend the functionality of the Web-based Hydropathy Amphipathicity and Topology (WHAT) program [Zhai and Saier, 2001]. Hydropathy values are computed as a moving average using a sliding window of 19 residues. Average hydropathy values are plotted at the central positions of their respective windows. Hydropathy profiles of aligned sequences were compared visually to assess the quality of the highest scoring matches between pairs of families.

Conservation of Protein Domains

To identify conserved domains, proteins were compared against Pfam [Finn et al., 2016] with the program HMMSCAN from the HMMER suite [Eddy, 2011] using the gathering threshold. The identification of characteristic domains (present in at least 80% of members) within a family and the projection of domains from one family to individual candidate homologs without direct Pfam matches were carried out using our in-house program get-DomainTopology [Medrano-Soto et al., 2020]. This program collects the sequence regions with direct Pfam hits in the family and aligns them to query proteins with no hits using SSEARCH [Pearson, 1991]. If significant alignments were detected (E-value $<10^{-2}$ and coverage $\geq 50\%$ of the domain sequences to account for repeats), then the domain was considered to be present in the query protein.

Conservation of Sequence Motifs

Identification of shared sequence motifs within and across families was carried out using the MEME suite [Bailey et al., 2015] as previously reported [Medrano-Soto et al., 2018]. To avoid intro-

ducing biases toward overrepresented families, given that family RV-U5 (TC: 1.A.100) has very few homologs in NCBI, we randomly selected 11 homologs from each family to search for conserved motifs. MEME was set to generate motif models for widths between 8 and 40 residues long using the mode One motif Occurrence Per Sequence (OOPS), at most 1,000 iterations of the expectation-maximization algorithm to calculate the maximum likelihood of models, a distance cutoff of 10^{-5} between frequency matrices, any model width allowed, and a maximal E-value of 10^{-5} . The resulting MEME models were benchmarked against two negative controls: (1) the PLB family (TC: 1.A.50), which also consists of small proteins with one or two TMSs; and (2) all families in TCDB except for the 4 families in the PSV superfamily. The motif of length 26 aas was selected as it retrieved the most members from all families while failing to generate matches in the negative controls (Fig. 5). Finally, the 26 aas motif was scanned against all four families and their homologs with MAST using a maximal E-value cutoff of 10^{-3} (Fig. 6).

Clustering Tree of Protein Sequence Similarities

In our initial approach, we attempted to build a phylogenetic tree of the 4 families included in this study. Protein sequences for all systems in these families were downloaded from TCDB. Multiple alignments were generated with MAFFT using the L-INS-i algorithm [Katoh and Standley, 2013]. Uninformative positions in the multiple alignment were removed with the program TrimAl [Capella-Gutierrez et al., 2009] to keep positions with less than 30% gaps. Phylogenetic analysis was carried out using a Bayesian approach as implemented in MrBayes [Ronquist and Huelsenbeck, 2003]. We assumed different substitution rates among sites that followed a gamma distribution with 4 rate categories based on the Jones-Taylor-Thornton rate matrix for amino acids. Posterior probabilities were estimated with Metropolis coupling (1 cold and 3 heated chains), and 2 million Markov Chain Monte Carlo (MCMC) generations were used to lower the average standard deviation of split frequencies below 0.01. Unfortunately, given the amount of sequence variation between these families, the tree obtained lacked significant statistical support (posterior probability <0.8) in three key branches (suppl. Fig. S4). Note that the topology of the tree and the statistical support for all branches did not change when running the analysis for 5 million and 10 million MCMC generations. Therefore, we complemented and compared the analysis with a different approach. We clustered protein sequences with the program mkProteinClusters [Medrano-Soto et al., 2018], which groups sequences based on pairwise Smith-Waterman bit scores calculated with the program SSEARCH36 [Pearson, 1991], and using the Ward agglomerative method as implemented in the R statistical computing environment (<https://www.R-project.org>). The clustering tree is shown in Figure 7. Because this method does not require multiple alignments, selection of evolutionary models, and consideration of different substitution rates, branch lengths are not directly indicative of evolutionary distance, and thus, the scale bar was omitted. However, the overall topologies of the trees (family groupings) produced by this method have shown excellent agreement with the topologies generated by phylogenetic trees [Medrano-Soto et al., 2018], as can also be observed by comparing Figure 7 and supplementary Figure S4.

Acknowledgements

We thank the members of the Saier laboratory for their valuable discussions during the development of this project.

Statement of Ethics

An ethics statement was not required for this study type, no human or animal subjects or materials were used.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This work was supported by grant GM077402 to MHS from the National Institutes of Health (<https://www.nih.gov/>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

D.T. performed most of the work for this project and participated in writing the manuscript. K.J.H. improved methodologies, performed analyses and developed programs for this project. A.M.S. supervised, designed strategies, performed analyses, developed programs for the project and participated in writing the manuscript. M.H.S. defined the project, designed strategies, supervised the project, obtained funding, and participated in writing the manuscript.

Data Availability Statement

All data used in this study are available as supplementary material in this article, the TCDB website (<https://tcdb.org/>), and FigShare (<https://doi.org/10.6084/m9.figshare.c.5603658>).

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.
- Anderson DM, Makarewich CA, Anderson KM, Shelton JM, Bezprozvannaya S, Bassel-Duby R, et al. Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci Signal.* 2016;9(457):ra119.
- Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. *Nucleic Acids Res.* 2015; 43(W1):W39–49.

- Blasdel KR, Widen SG, Diviney SM, Firth C, Wood TG, Guzman H, et al. Koolpinyah and Yata viruses: two newly recognised ephemeroviruses from tropical regions of Australia and Africa. *Vet Microbiol*. 2014;174(3-4): 547–53.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009;25(15):1972–3.
- Carrasco L. Membrane leakiness after viral infection and a new approach to the development of antiviral agents. *Nature*. 1978;272(5655): 694–9.
- Chang AB, Lin R, Keith Studley W, Tran CV, Saier MH Jr. Phylogeny as a guide to structure and function of membrane transport proteins. *Mol Membr Biol*. 2004;21(3):171–81.
- Crambert G, Geering K. FXYD proteins: new tissue-specific regulators of the ubiquitous Na,K-ATPase. *Sci STKE*. 2003;2003(166): RE1.
- De Marothy MT, Elofsson A. Marginally hydrophobic transmembrane α -helices shaping membrane protein folding. *Protein Sci*. 2015; 24(7):1057–74.
- Delpat B, Bibert S, Geering K. [FXYD proteins: novel regulators of Na,K-ATPase]. *Med Sci (Paris)*. 2006;22(6-7):633–8.
- Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 2011;7(10):e1002195.
- Finn RD, Coghill P, Eberhardt RF, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016;44(D1):D279–85.
- Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23): 3150–2.
- Garty H, Karlsh SJ. Role of FXYD proteins in ion transport. *Annu Rev Physiol*. 2006;68:431–59.
- Gazzarrini S, Kang M, Epimashko S, Van Etten JL, Dainty J, Thiel G, et al. Chlorella virus MT325 encodes water and potassium channels that interact synergistically. *Proc Natl Acad Sci U S A*. 2006;103(14):5355–60.
- Gonzalez ME, Carrasco L. Viroporins. *FEBS Lett*. 2003;552(1):28–34.
- Gubala AJ, Proll DF, Barnard RT, Cowled CJ, Cramer SG, Hyatt AD, et al. Genomic characterisation of Wongabel virus reveals novel genes within the Rhabdoviridae. *Virology*. 2008;376(1):13–23.
- Hyser JM, Estes MK. Pathophysiological consequences of calcium-conducting viroporins. *Annu Rev Virol*. 2015;2(1):473–96.
- Jespersen T, Grunnet M, Rasmussen HB, Jorgensen NB, Jensen HS, Angelo K, et al. The corticosteroid hormone induced factor: a new modulator of KCNQ1 channels? *Biochem Biophys Res Commun*. 2006;341(4):979–88.
- Joubert DA, Blasdel KR, Audsley MD, Trinidad L, Monaghan P, Dave KA, et al. Bovine ephemeral fever rhabdovirus alpha1 protein has viroporin-like properties and binds importin beta1 and importin 7. *J Virol*. 2014; 88(3):1591–603.
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013;30(4):772–80.
- Ketter E, Randall G. Virus impact on lipids and membranes. *Annu Rev Virol*. 2019;6(1):319–40.
- Lama J, Carrasco L. Expression of poliovirus non-structural proteins in *Escherichia coli* cells. Modification of membrane permeability induced by 2B and 3A. *J Biol Chem*. 1992; 267(22):15932–7.
- Lubarski I, Pihakski-Maunsbach K, Karlsh SJ, Maunsbach AB, Garty H. Interaction with the Na,K-ATPase and tissue distribution of FXYD5 (related to ion channel). *J Biol Chem*. 2005;280(45):37717–24.
- McWilliam SM, Kongsuwan K, Cowley JA, Byrne KA, Walker PJ. Genome organization and transcription strategy in the complex GNS-L intergenic region of bovine ephemeral fever rhabdovirus. *J Gen Virol*. 1997;78(Pt 6): 1309–17.
- Medrano-Soto A, Ghazi F, Hendargo KJ, Moreno-Hagelsieb G, Myers S, Saier MH Jr. Expansion of the Transporter-Opisn-G protein-coupled receptor superfamily with five new protein families. *PLoS One*. 2020;15(4): e0231085.
- Medrano-Soto A, Moreno-Hagelsieb G, McLaughlin D, Ye ZS, Hendargo KJ, Saier MH Jr. Bioinformatic characterization of the Anoctamin Superfamily of Ca²⁺-activated ion channels and lipid scramblases. *PLoS One*. 2018;13(3):e0192851.
- Meij IC, Koenderink JB, van Bokhoven H, Assink KF, Groenestege WT, de Pont JJ, et al. Dominant isolated renal magnesium loss is caused by misrouting of the Na(+),K(+)-ATPase gamma-subunit. *Nat Genet*. 2000;26(3):265–6.
- O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1): D733–45.
- Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*. 1991;11(3):635–50.
- Reddy BL, Saier MH Jr. Topological and phylogenetic analyses of bacterial holin families and superfamilies. *Biochim Biophys Acta*. 2013; 1828(11):2654–71.
- Reddy VS, Saier MH Jr. BioV Suite--a collection of programs for the study of transport protein evolution. *FEBS J*. 2012;279(11):2036–46.
- Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003;19(12):1572–4.
- Saier MH, Jr, Reddy BL. Holins in bacteria, eukaryotes, and archaea: multifunctional xenologues with potential biotechnological and biomedical applications. *J Bacteriol*. 2015; 197(1):7–17.
- Saier MH, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, et al. The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res*. 2021;49(D1): D461–D67.
- Sha Q, Pearson W, Burcea LC, Wigfall DA, Schlesinger PH, Nichols CG, et al. Human FXYD2 G41R mutation responsible for renal hypomagnesemia behaves as an inward-rectifying cation channel. *Am J Physiol Renal Physiol*. 2008;295(1):F91–9.
- Sze CW, Tan YJ. Viral Membrane Channels: Role and Function in the Virus Life Cycle. *Viruses*. 2015;7(6):3261–84.
- Teriete P, Franzin CM, Choi J, Marassi FM. Structure of the Na,K-ATPase regulatory protein FXYD1 in micelles. *Biochemistry*. 2007; 46(23):6774–83.
- Thiel G, Baumeister D, Schroeder I, Kast SM, Van Etten JL, Moroni A. Minimal art: or why small viral K(+) channels are good tools for understanding basic structure and function relations. *Biochim Biophys Acta*. 2011;1808(2):580–8.
- To J, Torres J. Beyond Channel Activity: Protein-Protein Interactions Involving Viroporins. *Subcell Biochem*. 2018;88:329–77.
- Wang SC, Davejan P, Hendargo KJ, Javadi-Razaz I, Chou A, Yee DC, et al. Expansion of the Major Facilitator Superfamily (MFS) to include novel transporters as well as transmembrane-acting enzymes. *Biochim Biophys Acta Biomembr*. 2020;1862(9):183277.
- Wong NA, Saier MH Jr. The SARS-Coronavirus Infection Cycle: A Survey of Viral Membrane Proteins, Their Functional Interactions and Pathogenesis. *Int J Mol Sci*. 2021;22(3).
- Wong WC, Maurer-Stroh S, Eisenhaber F. More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput Biol*. 2010;6(7):e1000867.
- Wong WC, Maurer-Stroh S, Eisenhaber F. Not all transmembrane helices are born equal: towards the extension of the sequence homology concept to membrane proteins. *Biol Direct*. 2011;6:57.
- Wong WC, Maurer-Stroh S, Schneider G, Eisenhaber F. Transmembrane helix: simple or complex. *Nucleic Acids Res*. 2012;40(Web Server issue):W370–5.
- Zhai Y, Saier MH Jr. A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J Mol Microbiol Biotechnol*. 2001;3(4):501–2.
- Zhang XQ, Wang J, Song J, Rabinowitz J, Chen X, Houser SR, et al. Regulation of L-type calcium channel by phospholemman in cardiac myocytes. *J Mol Cell Cardiol*. 2015;84:104–11.