UNIVERSITY OF CALIFORNIA, SAN DIEGO

Defense Against Cannibalism: the SdpI Family of Bacterial Immunity/Signal Transduction Proteins

A Thesis submitted in partial satisfaction of the requirements

for the degree Master of Science

in

Biology

by

Tatyana Leonidovna Povolotsky

Committee in Charge:

Professor Milton Saier, Jr., Chair
Professor Trey Ideker
Professor Kit Pagliano

2009

The Thesis of Tatyana Leonidovna Povolotsky is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____
Chair

University of California, San Diego

2009

DEDICATION

To my beloved family and friends, who have been there to help me up upon each one of my stumbles on my course through life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Milton Saier, Jr for his mentorship and guidance in preparing this work. I would like to also thank my co-authors: Ekaterina Orlova, Rachna Pandey, and Dorjee G. Tamang for their contribution to this work: Rachna Pandey who initiated this project, Dorjee G. Tamang who was invaluable and with limitless patience in helping with the numerous technical aspects, and Ekaterina Orlova through whose instrumental effort and devotion the completion of this research was achieved.

This thesis, in full, is a reprint of the material as it will appear in The SdpI Family of Antibiotic Peptide Killer Factor Immunity Proteins. Povolotsky, Tatyana Leonidovna; Orlova, Ekaterina; Pandey, Rachna; Tamang, Dorjee G.; Saier, Milton H., Jr. The thesis author is the primary investigator and author of this paper.

ABSTRACT OF THE THESIS

Defense Against Cannibalism: the SdpI Family of Bacterial Immunity/Signal Transduction Proteins

by

Tatyana Leonidovna Povolotsky

Master of Science in Biology

University of California, San Diego, 2009

Professor Milton Saier, Jr., Chair

The SdpI family consists of putative bacterial toxin immunity and signal transduction proteins. One member of the family in *Bacillus subtilis*, SdpI, provides immunity to cells from cannibalism in times of nutrient limitation. SdpI family members are transmembrane proteins with 3, 4, 5, 6, 7, 8 or 12 putative transmembrane α-helical segments (TMSs). These varied topologies appear to be genuine rather than artifactual due to sequencing or annotation errors. Bioinformatic methods were used to show that

the basic and most frequently occurring element of the SdpI family has 6 TMSs. Homologues of all topological types were aligned to determine the homologous TMSs and loop regions, and the Positive-Inside Rule was used to determine sidedness. The two most conserved motifs were identified between TMSs 1 and 2 and TMSs 4 and 5 of the 6 TMS proteins. These showed significant sequence similarity, leading us to suggest that the primordial precursor of these proteins was a 3 TMS-encoding genetic element that underwent intragenic duplication. Various fusional, insertional and deletion events, as well as intragenic duplications and inversions, are proposed to have yielded SdpI homologues with topologies of varying numbers of TMSs. We propose a specific evolutionary pathway that could have given rise to these distantly related bacterial immunity proteins. Our analyses allow us to propose structure-function relationships that may be applicable to most or all family members.

INTRODUCTION

Inhospitable environmental conditions prompt microbes to respond to stress by inducing the expression of stress response genes (Barak & Wilkinson, 2005; Hecker & Volker, 2001). In certain microbes such as *Bacillus subtilis*, a more elaborate response is induced under conditions of nutrient limitation: endospore formation (Aguilar *et al.*, 2007; Errington, 2003). Endospores are able to withstand environmental extremes and have the capacity to lie dormant for thousands if not millions of years (Vreeland *et al.*, 2000). The process of endospore formation is time and energy intensive, involving the expression of more than 500 genes over a 6-8 hour period (Britton *et al.*, 2002; Eichenberger *et al.*, 2004; Fujita & Losick, 2002; Molle *et al.*, 2003; Steil *et al.*, 2003). Since this process becomes irreversible after approximately 2 hours (Dworkin & Losick, 2005; Parker *et al.*, 1996) mechanisms exist that delay commitment to this process through cannibalism (Claverys & Havarstein, 2007). The SdpI family of proteins is involved in orchestrating one such delay (Ellermeier *et al.*, 2006). Members of the SdpI family are putative transmembrane proteins involved in both signal transduction and immunity to the cannibalistic process (Ellermeier *et al.*, 2006).

Under the conditions of nutrient limitation and high population density, the response regulator Spo0A is turned on in about half of the cells in the population (Chung *et al.*, 1994; Fujita & Losick, 2002; Gonzalez-Pastor *et al.*, 2003). Spo0A-ON cells switch on transcription of two operons; *sdpABC* and *skfA-H (Ellermeier et al., 2006).* The *skfA-H* operon contains genes for the production of a peptide-like antibiotic killing factor and an export pump that transports the killing factor out of the producing cells

thereby avoiding death of Spo0A-ON cells (Gonzalez-Pastor *et al.*, 2003). The *sdpABC* operon contains three genes that produce and export the SdpC toxin. The toxin and the killing factor lyse Spo0A-OFF cells and Spo0A-ON cells are able to delay or prevent commitment to endospore formation by feeding off of nutrients released from the dead cells (Ellermeier *et al.*, 2006). They may also use the released DNA for natural transformation (Grossman, 1995).

Spo0A-ON *B. subtilis* cells are immune to both the toxin and the killing factor they produce. The same operon that contains genes for the killing factor also contains genes for an export pump that removes it from the Spo0A-ON cells to avoid self-killing (Gonzalez-Pastor *et al.*, 2003). However, the operon that contains the toxin SdpC does not confer immunity. SdpC is, in fact, an extracellular signaling protein, as through its interaction with the SdpI protein the transcription of an adjacent convergently transcribed immunity operon, *sdpRI, is* induced. SdpI is a transmembrane immunity and signal transduction protein, while SdpR is the autorepressor. In Spo0A-ON cells, external SdpC acts as a ligand to existing SdpI in cell membranes. It alters the conformation of SdpI, inducing sequestration of the autorepressor, internal SdpR. Thus, the *sdpRI* operon is de-repressed so that more SdpI is transcribed and translated. Thus, a mechanism has evolved that confers immunity against the SdpC toxin only when SdpC is present.

In Spo0A-OFF cells, the AbrB repressor prevents expression of the *sdpRI* operon, and the cells, unable to promote immunity, die in the presence of external SdpC (Ellermeier *et al.*, 2006). It is thus likely that SdpI exhibits two distinct functions: immunity conferral and signal transduction; these two functions are localized to different parts of the protein. Localized mutagenesis of the first half of *Bacillus subtilis* SdpI

hinders its immunity function, while substitutions in the second half of the protein compromise the signal transduction function of SdpI (Butcher & Helmann, 2006). Other forms of resistance to SdpC have been identified: *yknWXYZ* and *yfhL* σ^w-dependent operons confer immunity to SdpC (Butcher & Helmann, 2006). *yknWXYZ* encodes an ABC transporter and is speculated to export the SdpC toxin, while *yfhL* encodes a paralogue of SdpI (Butcher & Helmann, 2006).

In this paper, we use established bioinformatic methodologies to provide evidence that the basic element of the SdpI family is a 6 TMS protein. This basic structure probably underwent duplication, deletion, inversion and fusion events to give rise to homologous proteins of 3, 4, 5, 7, 8 and 12 putative TMS topologies. The driving force for generation of this unusual degree of topological diversity may have been the bifunctional nature of SdpI where the first half of this proteins serves one function (binding of SdpC and immunity) while the second half serves another (binding of SdpR and signal transduction (Ellermeier *et al.*, 2006)). It is possible that the 6 TMS segment arose by intragenic duplication of a primordial 3 TMS segment. We provide presumptive, but extensive evidence for this postulate.

This section, in full, is a reprint of the material as it will appear in The SdpI Family of Antibiotic Peptide Killer Factor Immunity Proteins. Povolotsky, Tatyana Leonidovna; Orlova, Ekaterina; Pandey, Rachna; Tamang, Dorjee G.; Saier, Milton H., Jr. The thesis author is the primary investigator and author of this paper.

METHODS

Selection of protein sequences

A BLAST search (Altschul *et al.*, 1990) was performed in October 2007 using the SdpI protein of *Bacillus subtilis* [gi # 16080431] as the query sequence with two iterations and the default cut-off. More than one hundred homologous proteins were retrieved from the NCBI database. Eighty-two proteins were retained for topological analysis after redundancies and proteins with greater than 90% identity were eliminated using a modified CD-Hit program (Li *et al.*, 2001, 2002). The proteins were further reduced in number to 76 after translating the DNA in all 6 reading frames and seeking sequence similarities with full-length close homologues of the three translated co-directional reading frames.

The program BCM Search Launcher (Smith *et al.*, 1996) was used to translate the DNA coding for the query protein in the 6 reading frames at both ends flanking the existent sequence. The amino acid sequences at both the N- and C-termini were examined in all three reading frames for potential fragments, premature truncations, and incorrect initiation codon assignments. This was done for all proteins of the 5 TMS topology and smaller, as well as the inverted 6 TMS protein, Afu2, to establish the legitimacy of their topological deviations from the standard majority of 6 TMSs. If translation of any one of the reading frames preceding or following the reported sequence revealed significant similarity to another member of the SdpI family, the sequence was reconstituted or excluded from further studies. If not, it was retained and analyzed. In these procedures, any sequence of 20 aas or greater with 0, 1 or 2 stop codons was searched using the

BLAST search tool against the NCBI database to gain evidence for or against the possibility that the assigned initiation or termination codon was incorrect. If the BLAST search yielded significant similarity of the segment with a corresponding position of an established member of the SdpI family, the extended portion of the query protein was added to the original protein, and a new BLAST search was performed. If the results brought up a close homologue or a match for this new full-length protein, this protein was excluded from our analysis as its abbreviated topology was most likely artificial. When such procedures did not yield significant hits, the topology of the smaller protein was assumed to be accurate and was retained for further study.

A second BLAST search was performed on May 21, 2009, using the SdpI protein of *Bacillus cereus*, Bce2 [gi # 42784033] as the query sequence with two iterations. This was done to update the family, where new members with unexpected topologies were sought. The BLAST search with a cut-off of e^{-4} for the first iteration and a cut-off of e^{-5} for the second iteration yielded 316 homologues. All 316 homologues were analyzed, and their topologies were mapped manually. Proteins with new topologies, or topologies with only one previous example, were then added to the already existing family. Nine proteins were added to the original list. The previously described procedure employing BCM Search Launcher was preformed on these proteins.

Phylogenetic, hydropathy, and sequence analyses

Homologous sequences were multiply aligned using the ClustalX program (Thompson *et al.*, 1997), and phylogenetic trees were visualized using the TreeView program (Zhai *et al.*, 2002). Default parameters of ClustalX were used to align the

sequences. Topological analyses of the individual proteins and the multiply aligned homologues were performed using the WHAT (Zhai & Saier, 2001b) and AveHAS (Zhai & Saier, 2001a) programs, respectively. For the latter program, the ClustalX alignment was used as input to calculate average hydrophobicity and average similarity as a function of alignment position. The window size used was 19 residues. Statistical sequence similarity comparisons between proteins, and between internal regions of these proteins, were conducted using the IC (Zhai & Saier, 2002) and GAP (Devereux *et al.*, 1984) programs. These programs randomly shuffle the desired amino acid sequences and compare these shuffled sequences with the original sequences. In effect, they correct for unusual protein compositions such as those that occur in integral membrane proteins. Default settings and five hundred random shuffles have been shown to be satisfactory for obtaining statistically significant values (Yen *et al.*, 2009). A value of 10 standard deviations (S.D.) for comparable regions of two proteins of at least 60 amino acyl residues (aas) in length, corresponding to a probability of 10^{-24} that the observed degree of sequence similarity arose by chance (Dayhoff *et al.*, 1983; Saier *et al.*, 2009; Yen *et al.*, 2009) is considered sufficient to establish homology. These proteins were then analyzed topologically and phylogenetically. Reference to TMSs refers throughout to putative transmembrane spanners (TMSs), based on hydropathy analyses, since none of the proteins in this family have been characterized topologically.

Motif analyses

All of the SdpI proteins within our study were analyzed for motifs using the MEME program (Bailey & Elkan, 1995). Default settings were used, except that the

condition “any number of repetitions” was selected for the prediction of how single motifs were distributed among the sequences. The consensus sequences generated by the program guided the determination of the consensus sequences of the phylogenetic clusters through analysis of the ClustalX alignments of the individual clusters. The locations of the motifs were determined for individual proteins relative to the locations of the TMSs using the hydropathy plots generated by the WHAT program.

Determination of protein orientation within the cell membrane

The orientations of the SdpI homologues in the cell membrane were determined using the HMMTop (Tusnady & Simon, 2001) and TMHMM (Krogh *et al.*, 2001) programs. If and only if the two programs provided contradictory results were the proteins examined manually. The positively charged amino acyl residues (Arginine and Lysine) were counted in the first and last 20 residues of the primary sequence (unless otherwise specified – see Table 1 for exceptions), as well as in the loop regions between the TMSs. The inter-TMS loops were located using the TMHMM program and confirmed with the WHAT program (Zhai & Saier, 2001b). The Positive-Inside Rule was then applied to determine orientation of the proteins within the cell membrane (von Heijne & Gavel, 1988). Table S1 lists the proteins analyzed manually and includes the regions of the primary sequences that were examined for positively charged amino acyl residues. The numbers of positively charged residues (Rs and Ks) that were counted in the above mentioned regions are also recorded in Table S1. The regions with the largest numbers of positively charged residues were assumed to be located inside the cell. This process estimated orientation in the cell membrane. For proteins Bcl2 and Cte1, the WHAT program was also

used to determine the N- and C-terminal and loop regions, as the TMHMM program did not recognize all of the putative TMSs.

This section, in full, is a reprint of the material as it will appear in The SdpI Family of Antibiotic Peptide Killer Factor Immunity Proteins. Povolotsky, Tatyana Leonidovna; Orlova, Ekaterina; Pandey, Rachna; Tamang, Dorjee G.; Saier, Milton H., Jr. The thesis author is the primary investigator and author of this paper.

RESULTS

Table 1 lists the proteins of the SdpI family analyzed in this study alphabetically within each phylogenetic cluster (Figure 1). A multiple alignment of these proteins may be found on our website (http://biology.ucsd.edu/~msaier/supmat/SdpI-family) (Figure S1).

Classification of organisms represented in the SdpI family

Organisms represented include Firmicutes, with 52 of the 87 homologues derived from this bacterial kingdom. Euryarchaeota and Actinobacteria were equally represented (11 homologues each). There were also representatives from γ-proteobacteria (1), α-proteobacteria (3), *Bacteroidetes* (3), *Chlorobi* (2), *Chloroflexi* (2), Acidobacteria (1), Actinobacteria (11) and *Deinococcus* (1). The proteins vary widely in size, with sequences as short as 137 residues (Hma1 from *Haloarcula marismortui*) and as long as 404 residues (Dge1 from *Deinococcus geothermalis*). The majority of the proteins are of a size near 200 (170-230) residues in length and exhibit a putative 6 TMS topology. The SdpI family appears to be topologically heterogeneous; it includes four proteins predicted to have 3 TMSs, nine proteins with 4 TMSs, six proteins with 5 TMSs, fifty-eight proteins with 6 TMSs, four proteins with 7 TMSs, five proteins with 8 TMSs and one protein with 12 TMSs.

SdpI homologues

Figure 1 shows the phylogenetic tree for the SdpI family proteins included in this

study. These proteins cluster primarily in accordance with topology, and to a lesser degree with organismal type. Cluster 1 is made up of only 4 TMS proteins with the majority being from Firmicutes with two exceptions - Afu1 from *Archaeoglobus fulgidus*, a euryarchaeon, and Csp1 from *Cellulophaga sp.* MED134, a member of the *Bacteroidetes*. Cluster 2 is composed of eight proteins, a 4 TMS homologue from *Staphylococcus aureus* (a Firmicute), two 5 TMS proteins (both from Actinobacteria) and five 8 TMS homologues, of which four are from Firmicutes and one is from an Actinobacterium. Cluster 3 contains all of the 3 TMS proteins, four corynebacterial (Actinobacterial) orthologues.

Cluster 4 contains five proteins, Afu2 from *Archaeoglobus fulgidus* (6 TMSs), Dge1 from *Deinococcus geothermalis* (a 12 TMS homologue), and three 7 TMS homologues: Tko1 from *Thermococcus kodakarensis*, Ton1 from *Thermococcus onnurineus*, Tsp3 from *Thermococcus* sp. AM4. The proteins in this cluster are all from *Euryarchaota* except for Dge1. Surprisingly, they were found to have an inverted order of their two 3 TMS segments relative to the majority type. Accordingly, the first 3 TMSs in these proteins show a high degree of sequence similarity with the last 3 TMSs in the standard 6 TMS homologues, while the last 3 TMSs more closely resemble the first 3 TMSs in the standard 6 TMS homologues.

Cluster 5 contains three proteins of varying topologies. Aba1 from *Acidobacteria bacterium* (an *Acidobacterium*) has 6 TMSs; Cte1 from *Chlorobium tepidum* (a *Chlorobi*) has 5 TMSs, and Pae1 from *Prosthecochloris aestuarii* (a *Chlorobi*) has 4 TMSs. Cluster 6 is comprised predominantly of 6 TMS proteins from Firmicutes with the exception of the 4 TMS Hma1 homologue from *Haloarcula marismortui*, a member of the

Euryarchaeota. Cluster 7 is composed of four proteins, all from Firmicutes; two are 6 TMS homologues, and two are 5 TMS homologues.

Cluster 8 is made up of only 6 TMS homologues derived exclusively from Firmicutes. Cluster 9 is also derived from Firmicutes, and is comprised of 6 TMS proteins with just two exceptions: a 5 TMS protein from *Bacillus clausii* (Bcl2) and a 7 TMS homologue from *Dorea longicatena* (Dlo1). Cluster 10 contains only 6 TMS homologues of varying types, predominantly from Firmicutes, although five other phyla are represented (Table 1). It is interesting to note that most of the 6 TMS proteins cluster loosely together (clusters 8-10) while proteins of other topologies are phylogenetically more distant.

Search for internal repeats within the 6 TMS proteins

All of the 6 TMS proteins were analyzed for internal duplication of a 3 TMS segment and triplication of a 2 TMS segment, the two principal routes by which 6 TMS proteins have been shown to arise in other families (Kimball *et al.*, 2003; Lee *et al.*, 2007; Saier, 2003). However, we could not demonstrate homology of repeat segments, as both pathways gave comparable results far below the threshold comparison score needed for proof of homology, 10 S.D. (Saier, 1994; Saier *et al.*, 2009).

Sequence and topological analyses

The archaeal SdpI proteins, Afu2 (6 TMSs), Tko1 (7 TMSs), Ton1 (7 TMSs) and Tsp3 (7 TMSs), proved to have inverted segements of 3 TMSs relative to the standard 6 TMS homologues; TMSs 1-3 of the standard 6 TMS proteins are homologous to TMSs 4-

6 of the inverted proteins, and TMSs 4-6 of the standard 6 TMS proteins are homologous to TMSs 1-3 of the inverted proteins. All of the inverted 7 TMS proteins aligned throughout with each other and with TMSs 1-6 of the inverted 6 TMS protein, Afu2 (Figure 2). The seventh peak of the inverted 7 TMS proteins did not show statistically significant similarity to any of the peaks from the other proteins within the SdpI family, but the 7 TMS proteins all exhibited homology with each other throughout their lengths. They may have arisen by gene fusion following the inversion event.

To demonstrate the inversion, a representative of the standard 6 TMS topology, Bce2 of *Bacillus cereus,* was chosen arbitrarily for comparison with Afu2, one of the inverted proteins. Figure 3 shows the hydropathy plots for Afu2 and Bce2 where this inversion may be visualized. With respect to the relative positions of hydrophobic peaks in their WHAT-generated hydrophobicity plots (Zhai & Saier, 2001b) the first half of Afu2 resembles the second half of Bce2, and the first half of Bce2 resembles the second half of Afu2. Figure 4A shows the GAP analysis between TMSs 1-3 of Afu2 and TMSs 4-6 of Bce2, with a comparison score of 16.6 S.D. Figure 4B shows the GAP analysis between TMSs 4-6 of Afu2 and TMSs 1-3 of Bce2, with a comparison score of 15.5 S.D. These values are substantially in excess of what is required to establish homology (Saier, 1994; Saier *et al.*, 2009).

Excluding the four archaeal proteins with inverted 3 TMS segments noted above, all of the 6 TMS proteins aligned with each other throughout their lengths. We then analyzed proteins with other topologies to determine the regions of homology with the standard 6 TMS homologues. In the corynebacterial proteins with 3 TMSs (Cluster 3), the 3 TMSs correspond only to TMSs 4-6 in the 6 TMS proteins (Figure 5). The 4 TMS

proteins align with each other and correspond to TMSs 1-4 in the 6 TMS proteins. Figure 6 presents a GAP analysis of the 4 TMS Hma1 homologue with the 6 TMS Gka1 protein; it demonstrates the afore mentioned alignment with a comparison score of 15.3 S.D. Proteins with 4 TMSs are found predominantly in Cluster 1, the three exceptions being Pae1 from *Prosthecochloris aestuarii*, found in Cluster 5, Sau1 from *Staphylococcus aureus*, located in Cluster 2, and Hma1 from *Haloarcula marismortui*, located in Cluster 6. Although Hma1 is found in Cluster 6, based on the branching pattern of the tree, it is distantly related to all of the 6 TMS proteins. This, in turn, leads to the supposition that the 4 TMS topology arose at least twice from the 6 TMS proteins, once by truncation of a Cluster 6 homologue, leading to the formation of Hma1, and once by truncation of a Cluster 1 6 TMS homologue. Pae1 is associated with Cte1 from *Chlorobium tepidum*, a 5 TMS protein whose hydrophobic peaks 2-5 correspond to peaks 1-4 in Pae1 and 1-4 in any of the standard 6 TMS proteins. The first peak of Cte1 does not align with anything else in these proteins, leading to the suggestion that this unique 5 TMS topology arose from the 4 TMS proteins through a gene fusion event at the N-terminus or by extensive sequence divergence over evolutionary time. Pae1 and Cte1 are found in Cluster 5 along with Aba1. Aba1 is the longest 6 TMS protein with 303 residues. Only the first 210 residues code for the membrane-integrated portion of the protein.

The 5 TMS proteins proved to have the most varied topologies. There are four unique 5 TMS topologies, each aligning slightly differently with the standard 6 TMS proteins. Cte1 (Cluster 5) is the only protein within the SdpI family to have its TMSs 2-5 aligning with TMSs 1-4 in the standard 6 TMS proteins (Figure 7). The first peak of Cte1 does not align with any of the peaks within the family and has been given the designation

of "A." Bcl2, with a differing 5 TMS topology, has peaks 1-5 aligning with peaks 2-6 of the standard 6 TMS proteins (Figure 8). It is found within Cluster 9, clustering mainly with 6 TMS proteins, suggesting that it evolved by deletion of a TMS from the N-terminus of a 6 TMS protein. The third variation in the 5 TMS topology is exemplified by two proteins: Sgo1 and Ssa2. These two proteins align with each other, and their peaks, numbered 1-5, correspond to peaks 1-5 of the standard 6 TMS proteins (Figure 9). They appear in Cluster 7 with 6 TMS proteins and seem to have arisen by deletion of a TMS from the C-terminus of a 6 TMS protein. The final 5 TMS topological variant type is illustrated by proteins Rsa1 and Cgl2. Peaks 1-4 in these two proteins align with peaks 1-4 of the standard 6 TMS proteins (Figure 10). Their 5th peak corresponds best to the 8th peak of the 8 TMS proteins. Rsa1 and Cgl2 align with the 8 TMS proteins throughout their lengths, with their TMSs 1-5 aligning with TMSs 4-8 in the 8 TMS homologues. The two 5 TMS proteins align with each other throughout and align extremely well with the 8 TMS proteins, as revealed by a comparison score of 35.4 S.D. between proteins Rsa1 from *Renibacterium salmoninarum*, a 5 TMS protein, and Lsp1 from *Lysinibacillus sphaericus*, an 8 TMS homologue (Figure 11).

The 8 TMS homologues, though aligning well with themselves, align only partially with the standard 6 TMS proteins. Peaks 4-7 of the 8 TMS proteins align with peaks 1-4 of the standard 6 TMS proteins (Figure 12). The eighth peak of the 8 TMS homologues and the fifth peak of Rsa1 and Cgl2, are designated "B" and do not match any of the TMSs within other members of the family. The first three TMSs of the 8 TMS homologues also do not have matches within the SdpI family and were designated "E," "F," and "G," respectively. The 8 TMS proteins and the two 5 TMS proteins (Rsa1 and

Cgl2) are found in Cluster 2 along with a 4 TMS protein, Sau1. It is possible that the 5 TMS proteins arose by addition of one TMS at the C-terminus of a 4 TMS protein. The 8 TMS topology may then have arisen from the 5 TMSs by the addition of three TMSs at the N-terminus of a 5TMS protein. Other possibilities can be considered.

There are two variations of the 7 TMS topology. The first is an inverted topology as previously discussed. The second is observed in Dlo1 with TMSs 1-6 aligning with TMSs 1-6 of the standard 6 TMS proteins (Figure 13). The seventh peak of Dlo1 does not align with any other peak within the SdpI family and is designated "C." This protein is found in Cluster 9 with 6 TMS proteins and Bcl2 of 5 TMSs. This clustering leads to the supposition that Dlo1 originated from a 6 TMS protein by addition of a C-terminal TMS.

An internal duplication within Dge1

The final topology is that of Dge1, a 12 TMS protein. Dge1 was cut in half to test for an internal duplication. A GAP analysis of the first 6 TMSs against the second 6 TMSs yielded a comparison score that was insufficient to establish homology. However, when the two halves were compared to the 6 TMS proteins, statistically significant similarity was found between several 6 TMS proteins and both halves of Dge1, clearly implying by the Superfamily Principle (Doolittle, 1981; Saier, 1994) that an intragenic duplication event of the basic 6 TMS element had led to the formation of the 12 TMS protein. The best comparison score was 19.3 S.D., generated by the comparison of the first half of Dge1 with Bcl1 (Figure 14), with TMSs 4-6 of Dge1 corresponding to TMSs 4-6 of Bcl1. The second half of Dge1 aligned with Mma2 (Figure 15), giving a

comparison score of 11.5 S.D.. The 6 TMS protein, Dha1, aligned with both halves of Dge1. Alignment with the first half of Dge1gave a comparison score of 15.4 S.D. (Figure 16), while alignment of the second half of Dge1 with Dha1 gave a comparison score of 14.6 S.D. (Figure 17).

The duplication event that led to the appearance of Dge1 was evidently followed by extensive sequence divergence within both halves of Dge1. The middle region of Dge1, spanning approximately 6 TMSs in length (TMSs 4-9) is better conserved than the end regions spanning TMSs 1-3 and TMSs 10-12. This is evident in the alignment of the inverted 6 TMS protein, Afu2, with TMSs 4-9 in Dge1, yielding a comparison score of 20.9 S.D. (Figure 18). The appearance of the hydropathy plot (WHAT program) for Dge1 also supports the conclusion of an internal duplication (Figure 19). The evidence supports the proposal that the 6 TMS proteins represent the basic element for the SdpI family from which other family members evolved.

Figure 20 shows the average hydropathy plot (top) and average similarity plot (bottom) for the SdpI family of proteins excluding the four internally inverted proteins, Afu2, Tsp3, Ton1 and Tko1, and with the 12 TMS protein, Dge1, cut into two 6 TMS segments. The plots were generated from the multiple alignment shown in Figure S2. Alignment of the proteins is shown according to their topologies (Figure 20) as summarized in Figure 21. Proteins of the 6 TMS topology, with the exception of the four inverted proteins, all align with TMSs 1-6 of all of the others. The 4 TMS proteins align with each other as well as with TMSs 1-4 of the 6 TMS proteins. The 3 TMS proteins also align with each other and with TMSs 4-6 of the 6 TMS proteins. The four varying 5 TMS topologies partially align with each other; TMSs 2-5 of Cte1 align with TMSs 1-4

of the 6 TMS proteins. In Bcl2, TMSs 1-5 align with TMSs 2-6 of the 6 TMS proteins. TMSs 1-5 of Sgo1 and Ssa2 align with each other and with TMSs 1-5 of the 6 TMS proteins. Rsa1 and Cgl2 align with each other, and their TMSs 1-4 align with TMSs 1-4 of the 6 TMS proteins. TMSs 1-6 of Dlo1 (7 TMS topology) align with TMSs 1-6 of the 6 TMS proteins. TMSs 4-7 of the 8 TMS proteins align with TMSs 1-4 of the 6 TMS proteins. Finally, TMSs 1-6 and TMSs 7-12 of the 12 TMS protein, Dge1, align with TMSs 1-6 of the 6 TMS proteins as noted above.

Motif Analyses

Proteins of the SdpI family have two well conserved motifs that were recognized by the MEME program (Bailey & Elkan, 1995). The best conserved motif, Motif 1 ([IV]G[LI]L[FL]I[VG][LI]GNY[LM][PG]KX[KR]PN[YW]F[VI]GIRTPWTLS[SN][ED]EVW[RN]KT[HN]R[LF][GA]G[KR][LV][FW]V[IAV][GA]G) (alternative residues at a single position are in brackets; X = any residue) is well conserved in the majority of the members of the family. It spans the hydrophilic region between the fourth and fifth TMSs in the standard 6 TMS proteins. It was also identified in the expected locations of most of the other topological variants that include TMSs 4 and 5. Using the 3 TMS proteins as an example, Motif 1 is found between the first and second TMSs as expected since these proteins align with TMSs 4-6 of the standard 6 TMS proteins. Figures 11 and S13 depict the locations of the recognized Motif 1 variants in all of the proteins displaying this motif within the SdpI family. All members of clusters 3, 4, 8, 9 and 10 have this motif, but Lpl1 from *Lactobacillus plantarum* is the only protein in Cluster 7 for which the MEME program recognized Motif 1. Likewise, Cac2 from

Corynebacterium accolens and Swo1 from *Syntrophomonas wolfei* were the only proteins in cluster 2 for which MEME identified this motif. It is possible that this motif deviates in sequence in some clusters. Such differences may have functional significance (see Discussion).

The second best conserved motif, Motif 2 (AL[YW]PXLP[ED]R[VI][PA][VI]H [WF][NG]ASGE[VP][DN][GR][YF][GM]SKF[EV][GL]) is also found in most members of the family that include TMSs 1 and 2. Based on results obtained with the MEME and WHAT programs, Motif 2 spans the hydrophilic region between the first and second TMSs of the standard 6 TMS proteins. Clusters 1, 2, 4, 5, 6, 7, 8 and 10 contain variants of Motif 2. The absence of this motif in Cluster 3 is logical because Cluster 3 contains the 3 TMS proteins homologous to TMSs 4-6 of the standard 6 TMS proteins, while Motifs 2 are found in the region between TMSs 1 and 2. Therefore, Motif 2 would not be expected to appear in these proteins. Lsa1 from *Lactobacillus salivarius* and Bsu1 from *Bacillus subtilis* are the only members of Cluster 9 to have Motif 2.

The majority of the proteins with the standard 6 TMS topology have one of three combinations of these two motifs. 6 TMS proteins from clusters 8 and 10 contain both motifs, with Motif 2 upstream of Motif 1. The four inverted proteins were also found to contain the same combination of motifs albeit in an inverted manner.

The 6 TMS proteins of clusters 5, 6 and 7 contain only Motif 2 with the exception of Lpl1 of Cluster 7, which displays both motifs. Finally, Cluster 9 contains 6 TMS proteins in which only Motif 1 was recognized by MEME except for the afore mentioned proteins, Lsa1 and Bsu1.

All of the standard 6 TMS proteins align throughout their lengths and have high

comparison scores with one another despite variations in the sequences displayed by these two motifs. The cluster differences for these two motifs are summarized in Table 2A and B, as are the sequence similarities between the consensus motifs 1 and 2 (Table 2C).

Proposed pathway for the evolution of varying topologies

Figure 22 diagrams the proposed pathway for the evolution of proteins of the SdpI family and shows their differing topologies. The primary 6 TMS proteins, assumed to correspond to the basic element from which all other topological types derived, may have arisen through intragenic duplication of a primordial 3 TMS-encoding DNA segment. Deletions in this basic element lead to the formation of the 4 TMS, 3 TMS and two of the 5 TMS variant proteins. Deletion and fusion events appear to have led to the evolution of the two other 5 TMS variants as well as to the 8 TMS proteins. A fusion event led to the appearance of the non-inverted 7 TMS protein (Dlo1). An inversion of the two 3 TMS portions of the 6 TMS proteins led to the Afu2 protein (6 TMS), and this same inversion event also produced the 7 TMS proteins, Tko1, Tsp3 and Ton1, but with a fusional event at the C- terminus generating the extra TMS. Finally, the 12 TMS protein undoubtedly arose by intragenic duplication of the basic 6 TMS element followed by extensive sequence divergence of both halves.

This section, in full, is a reprint of the material as it will appear in The SdpI Family of Antibiotic Peptide Killer Factor Immunity Proteins. Povolotsky, Tatyana Leonidovna; Orlova, Ekaterina; Pandey, Rachna; Tamang, Dorjee G.; Saier, Milton H., Jr. The thesis author is the primary investigator and author of this paper.

FIGURES

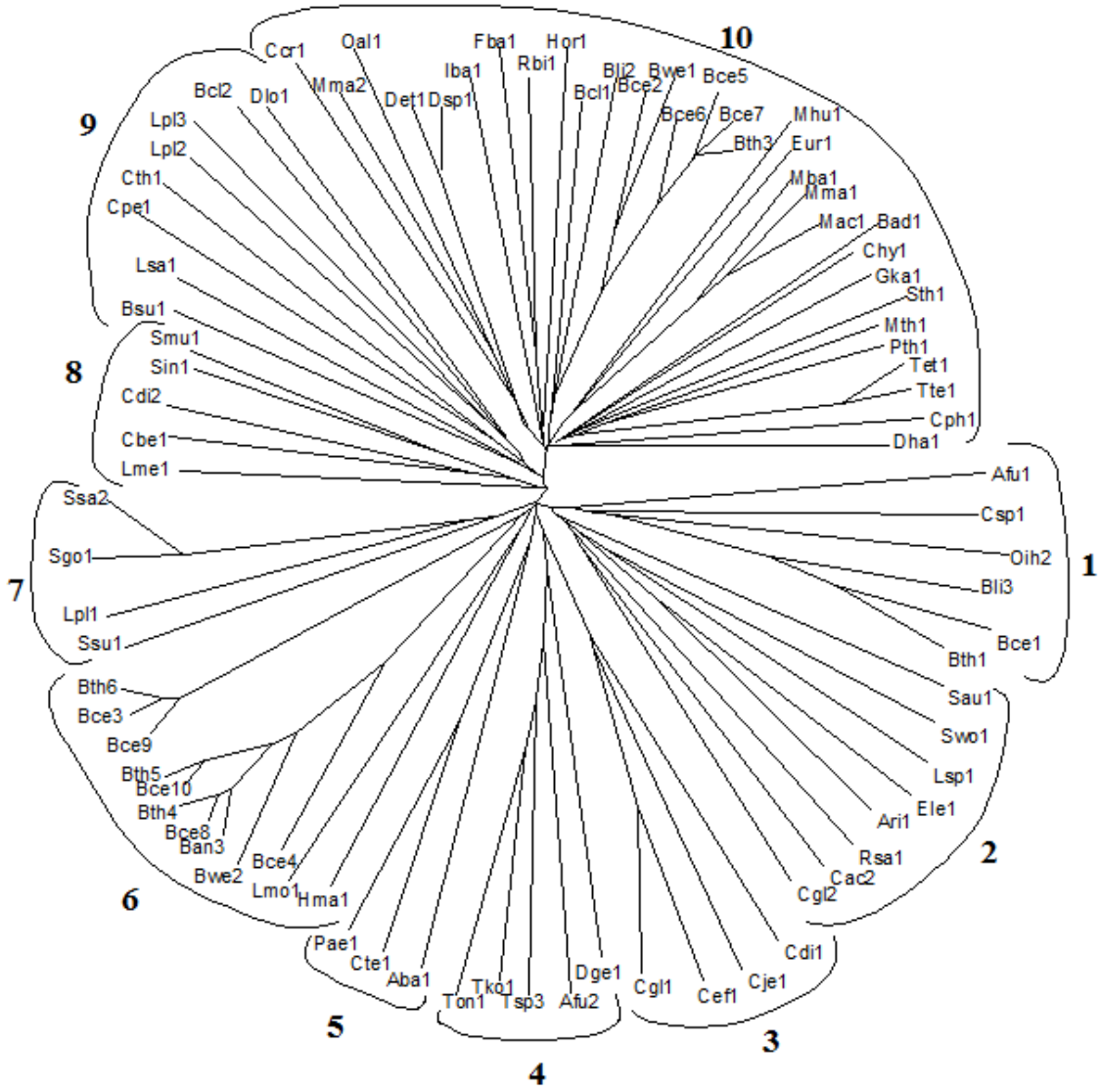

Figure 1: Phylogenetic tree of the SdpI family with labeled phylogenetic clusters. The tree is based on the ClustalX multiple alignment shown in Fig. S1 and drawn with the TreeView program. Protein abbreviations are listed in Table 2.

Figure 2: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Afu2 (residues 1 to 228), an inverted 6 TMS protein, is aligned with Tko1 (residues 1 to 237), an inverted 7 TMS representative. Quality: 413; Length: 240; Gaps: 8; Percent similarity: 54.2; Percent identity: 41.8. Average quality based on 100 randomizations for 5 runs: 13.3+/-5.8, 14.1 +/-6.0, 13.6 +/-5.0, 13.4 +/-5.8, 13.6 +/-4.9. The average comparison score was 73.1 S.D.

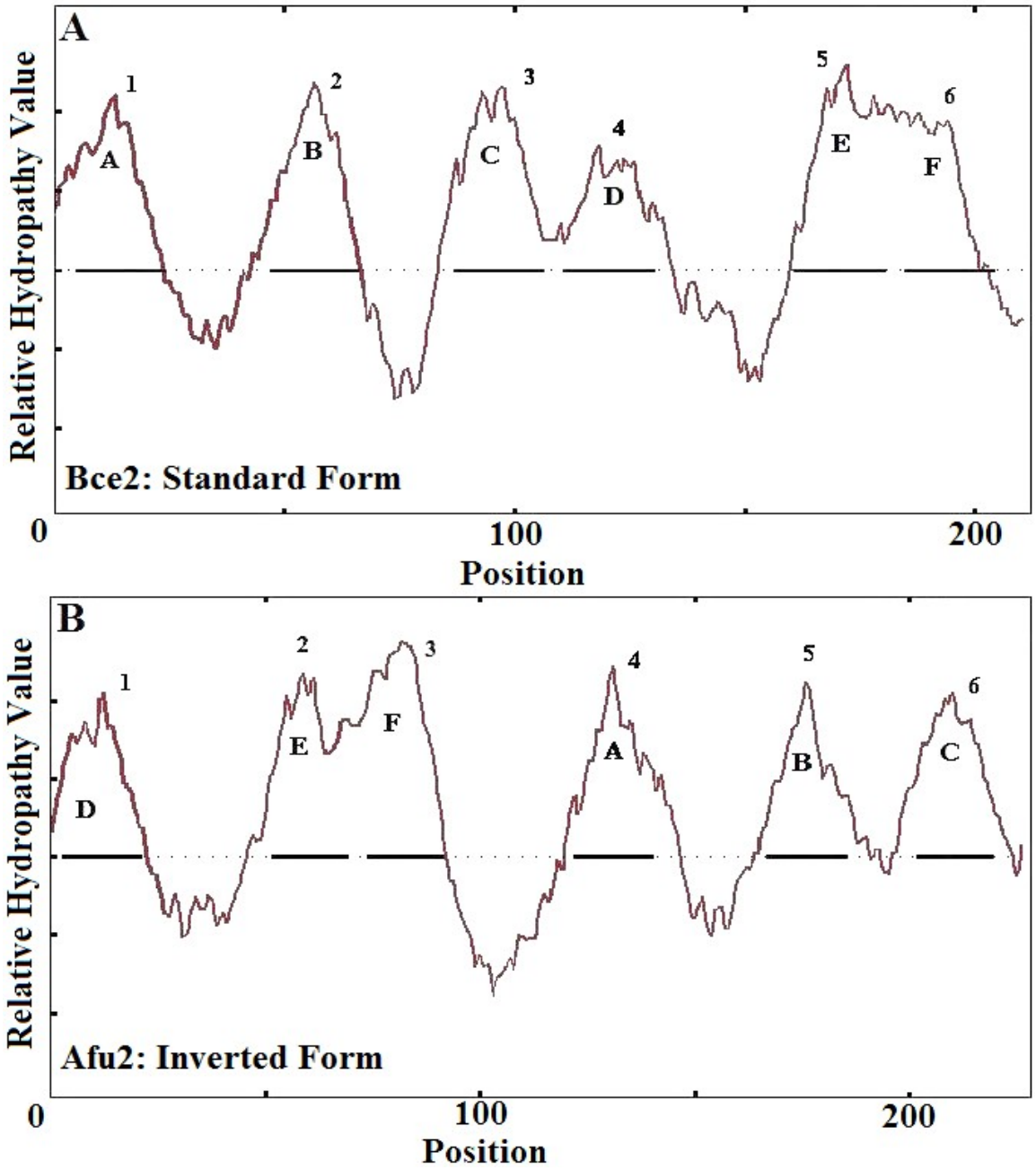

Figure 3: (A) Hydropathy plot of the SdpI protein from *Bacillus cereus* (Bce2) with numbered peaks of hydropathy corresponding to putative TMSs. (B) Hydropathy plot of the SdpI homologue from *Archaeoglobus fulgidus* (Afu2) with numbered TMSs. The letters correspond to the homologous TMSs between the 2 proteins, demonstrating the inversion within Afu2 relative to the standard 6 TMS proteins, represented here by Bce2.

Figure 4: (A) GAP comparison of the first 3 TMS segment of Afu2 (residues 1 to 105) with the second 3 TMS segment of Bce2 (residues 111 to 212) proteins using the GAP program. Quality: 102; Length: 108; Gaps: 5; Percent similarity: 44.4; Percent identity: 33.3. Average quality based on 100 randomizations for 5 runs: 9.9 +/-5.3, 10.5 +/-5.9, 10.8 +/-5.8, 9.6 +/-5.4, 10.0 +/-5.3. The average comparison score was 16.6 S.D. The average comparison score was 16.6 S.D. (B) GAP comparison of the second 3 TMS segment of Afu2 (residues 106 to 228) with the first 3 TMS segment of Bce2 (residues 1 to 110) proteins using the GAP program. Quality: 87; Length: 125; Gaps: 3; Percent similarity: 38.9; Percent identity: 21.3. Average quality based on 100 randomizations for 5 runs: 10.7 +/-5.0, 9.6 +/-4.9, 10.5 +/-4.7, 10.0 +/-4.4, 10.4 +/-6.0. The average comparison score was 15.5 S.D.

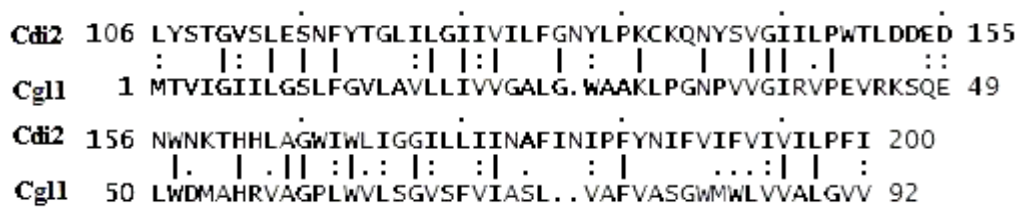

Figure 5: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Cdi2 (residues 106 to 200), a 6 TMS representative, is aligned with Cgl1 (residues 1 to 92), a 3 TMS representative. Quality: 79; Length: 95; Gaps: 2; Percent similarity: 41.3; Percent identity: 26.1. Average quality based on 100 randomizations for 5 runs: 15.5 +/-6.0, 13.4 +/-5.5, 14.8 +/-6.4, 14.3 +/-5.9, 14.0 +/-5.4. The average comparison score was 11.1 S.D.

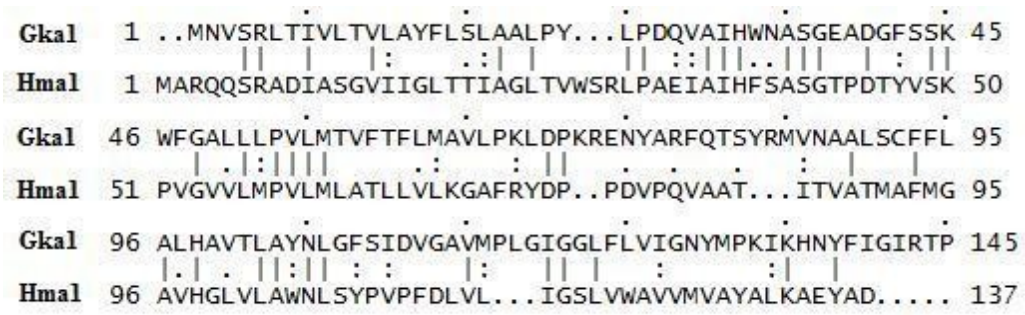

Figure 6: GAP alignment demonstrating the regions of homology between varying topological types within the SdpI family. Gka1 (residues 1 to 214), a 6 TMS representative, is compared with Hma1 (from residues 1 to 137), a 4 TMS representative. Quality: 106; Length: 219; Gaps: 4; Percent similarity: 40.9; Percent identity: 29.5. Average quality based on 100 randomizations for 5 runs: 12.9 +/-6.1, 12.4 +/- 6.2, 11.5 +/- 5.8, 12.8 +/- 6.0, 12.3 +/- 6.6. The average comparison score was 15.3 S.D.

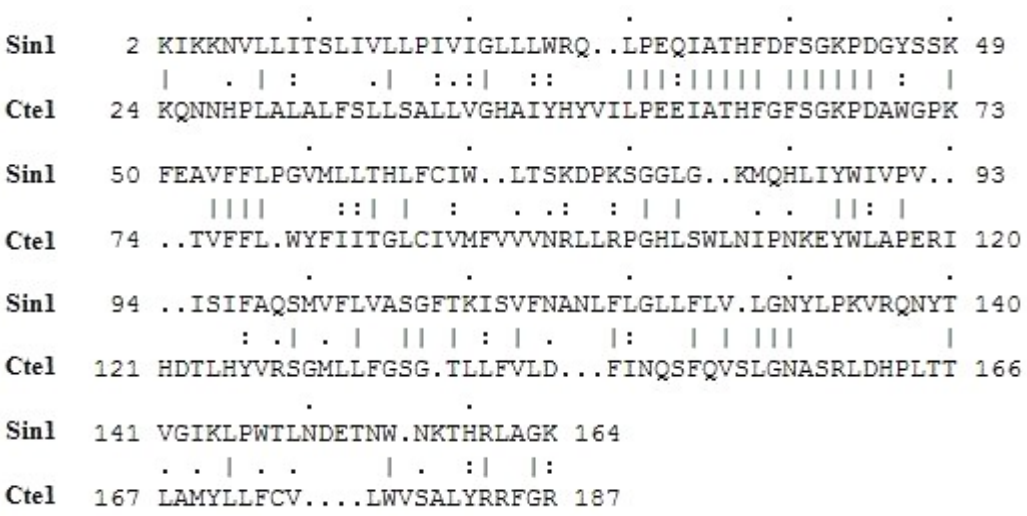

Figure 7: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Sin1 (residues 2 to 164), a 6 TMS representative, is aligned with Cte1 (residues 24 to 187), a 5 TMS Representative. Quality: 76; Length: 175; Gaps: 11; Percent similarity: 42.8; Percent identity: 30.9. Average quality based on 100 randomizations for 5 runs: 9.2 +/-5.2, 9.3 +/-5.1, 9.6 +/-5.1, 10.0 +/-5.3, 10.0 +/-6.2. The average comparison score was 12.4 S.D.

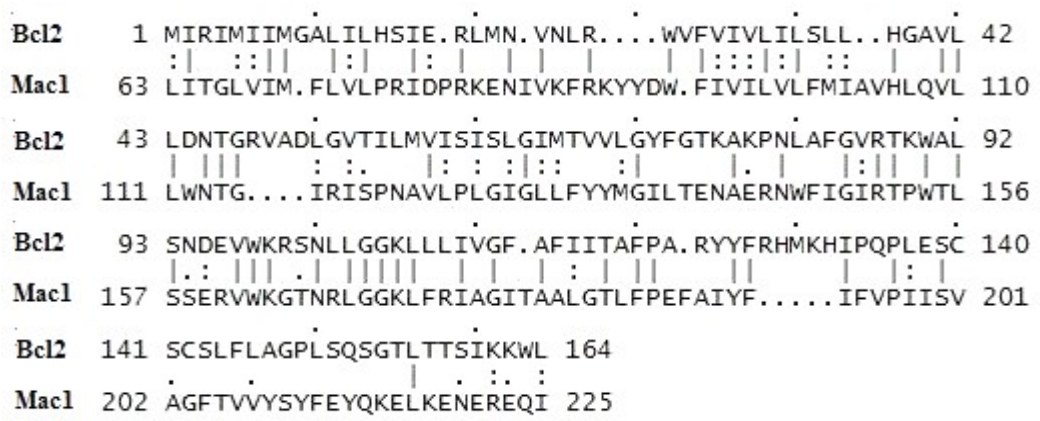

Figure 8: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Bcl2 (from residues 1 to 164), a 5 TMS representative, is aligned with Mac1 (from 63 to 225 aa), a 6 TMS representative. Quality: 128; Length: 174; Gaps: 10; Percent similarity: 51.0; Percent identity: 34.6. Average quality based on 100 randomizations for 5 runs: 12.3+/-6.0, 12.3 +/-4.7, 12.4 +/-4.9, 12.5 +/-6.1, 12.5 +/-5.8. The average comparison score was 21.3 S.D.

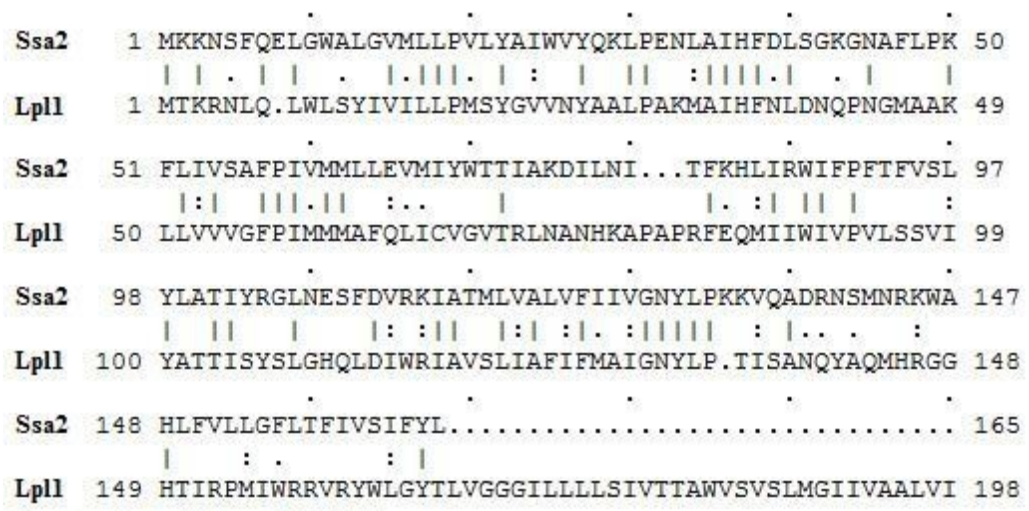

Figure 9: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Ssa2 (residues 1 to 165), a 5 TMS representative, is aligned with Lpl1 (residues 1 to 198), a 6 TMS representative. Quality: 124; Length: 413; Gaps: 12; Percent similarity: 43.4; Percent identity: 33.8. Average quality based on 100 randomizations for 5 runs: 14.4+/-6.1, 14.4 +/-6.1, 14.3 +/-6.4, 14.2 +/-7.0, 14.8 +/-6.4. The average comparison score was 17.2 S.D.

Figure 10: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Rsa1 (residues 4 to 292), a 5 TMS representative, is aligned with Mac1 (residues 2 to 227), a 6 TMS representative. Quality: 58; Length: 299; Gaps: 7; Percent similarity: 33.8; Percent identity: 22.2. Average quality based on 100 randomizations for 5 runs: 7.0+/-4.2, 7.0 +/-4.2, 7.5 +/-4.5, 6.7 +/-3.8, 7.6 +/-4.7. The average comparison score was 11.9 S.D.

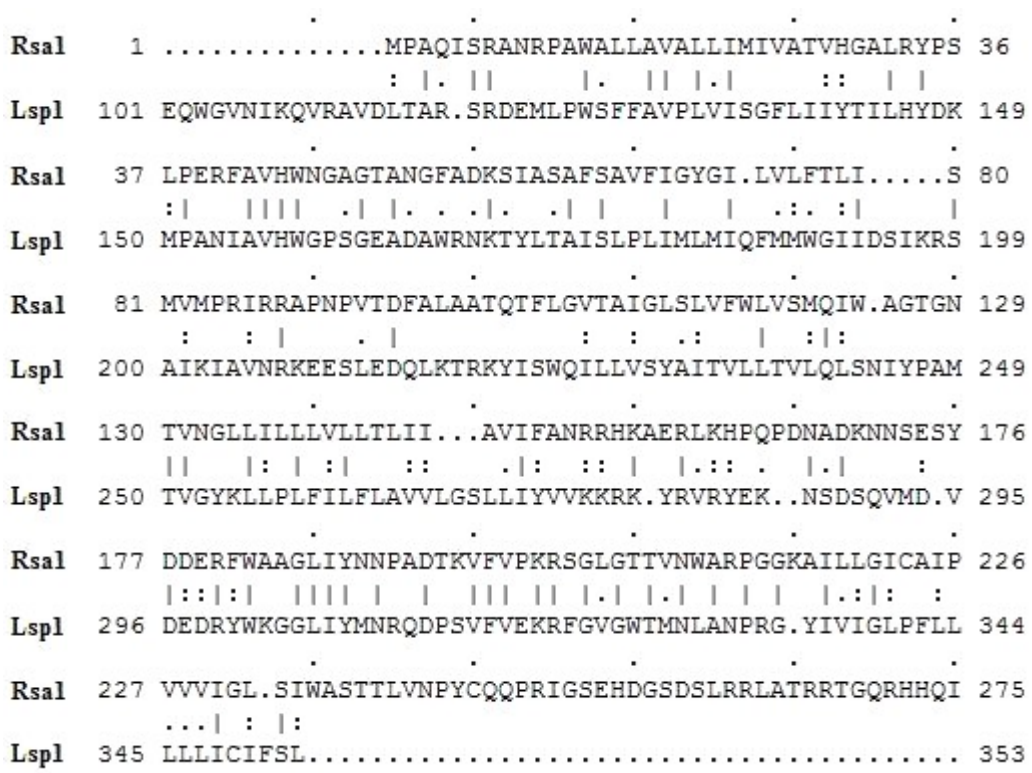

Figure 11: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Rsa1 (residues 1 to 292), a 5 TMS representative, is aligned with Lsp1 (residues 1 to 353), an 8 TMS representative. Quality: 174; Length: 417; Gaps: 10; Percent similarity: 41.2; Percent identity: 27.6. Average quality based on 100 randomizations for 5 runs: 7.7 +/-4.8, 7.9 +/-4.9, 7.0 +/-4.5, 8.0 +/-5.2, 8.2 +/-4.2. The average comparison score was 35.4 S.D.

Figure 12: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Mac1 (residues 2 to 212), a 6 TMS representative, is aligned with Swo1 (residues 134 to 373), an 8 TMS representative. Quality: 98; Length: 243; Gaps: 13; Percent similarity: 42.3; Percent identity: 30.3. Average quality based on 100 randomizations for 5 runs: 12.6+/-6.4, 12.3 +/-6.0, 12.3 +/-6.0, 11.8 +/-6.0, 12.6 +/-6.1. The average comparison score was 14.1 S.D.

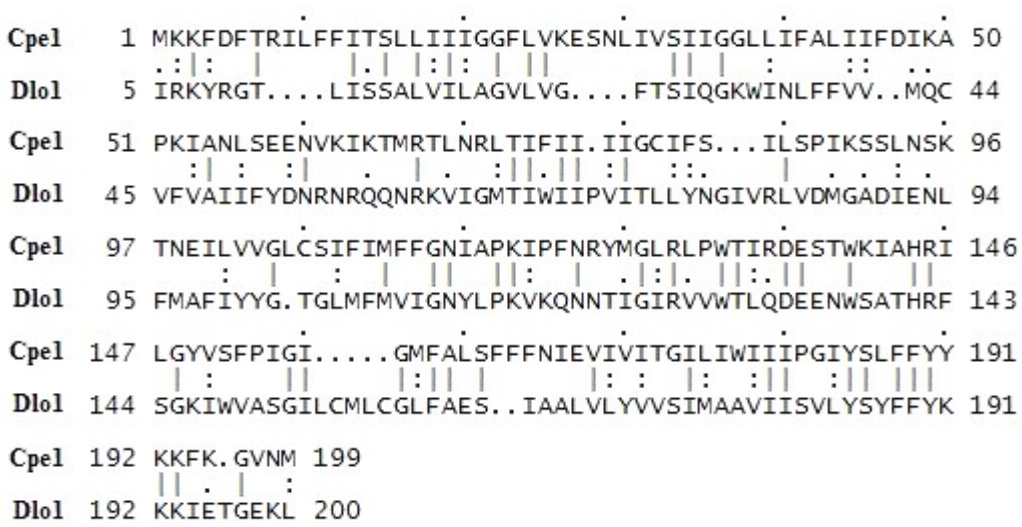

Figure 13: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Cpe1 (residues 1 to 199), a 6 TMS representative, is aligned with Dlo1 (residues 5 to 200), a 7 TMS representative. Quality: 144; Length: 209; Gaps: 9; Percent similarity: 45.2; Percent identity: 30.1. Average quality based on 100 randomizations for 5 runs: 17.0 +/-6.1, 16.1 +/-6.5, 17.2 +/-6.5, 17.0 +/-5.8, 16.5 +/-6.4. The average comparison score was 20.4 S.D.

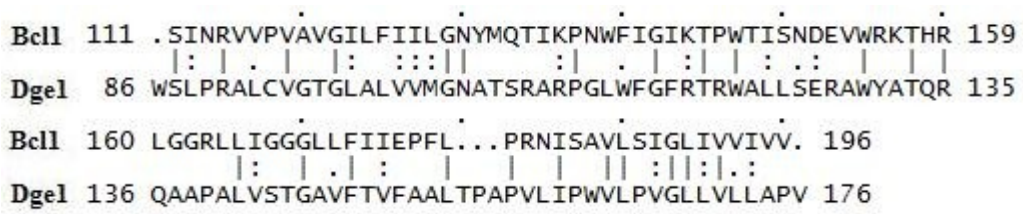

Figure 14: GAP comparison of Dge1 (residues 86 to 176), a 12 TMS protein, with Bcl1 (residues 111 to 196), a 6 TMS protein using the GAP program. Quality: 110; Length: 91; Gaps: 1; Percent similarity: 44.2; Percent identity: 27.9. Average quality based on 100 randomizations for 5 runs: 12.4 +/-5.3, 11.8 +/-5.1, 11.2 +/-4.7, 11.5 +/-4.5, 12.1 +/- 6.1. The average comparison score was 19.3 S.D.

Figure 15: GAP comparison of Mma2 (residues 1 to 207), a 6 TMS protein, with Dge1 (residues 214 to 404), a 12 TMS protein using the GAP program.. Quality: 88; Length: 210; Gaps: 8; Percent similarity: 37.2; Percent identity: 33.5. Average quality based on 100 randomizations for 5 runs: 18.2 +/- 6.0, 17.4 +/- 5.9, 18.2 +/- 7.1, 18.3 +/- 6.7, 14.0 +/-5.4. The average comparison score was 11.5 S.D.

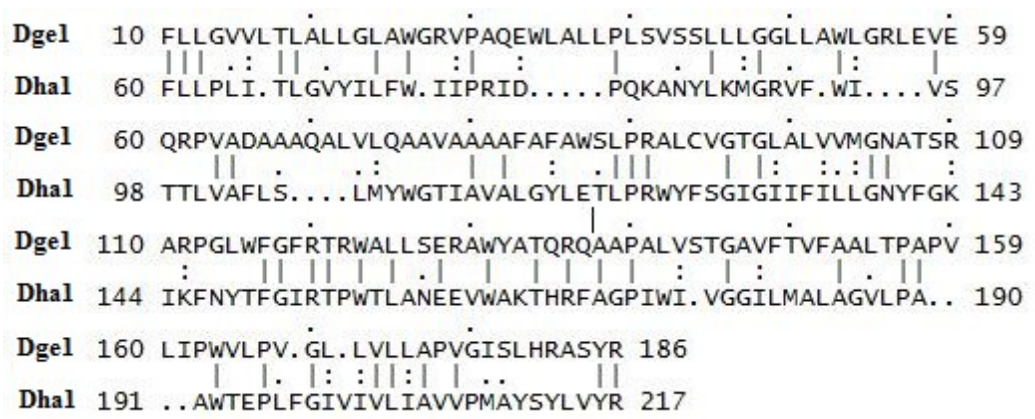

Figure 16: GAP comparison of Dge1(residues 10 to 186), a 12 TMS protein, with Dha1 (residues 60 to 217), a 6 TMS protein using the GAP program.. Quality: 114; Length: 179; Gaps: 10; Percent similarity: 42.3; Percent identity: 31.4. Average quality based on 100 randomizations for 5 runs: 16.4 +/-5.5, 17.4 +/-7.0, 16.7 +/-6.2, 16.8 +/-6.1, 17.0 +/-6.9. The average comparison score was 15.4 S.D.

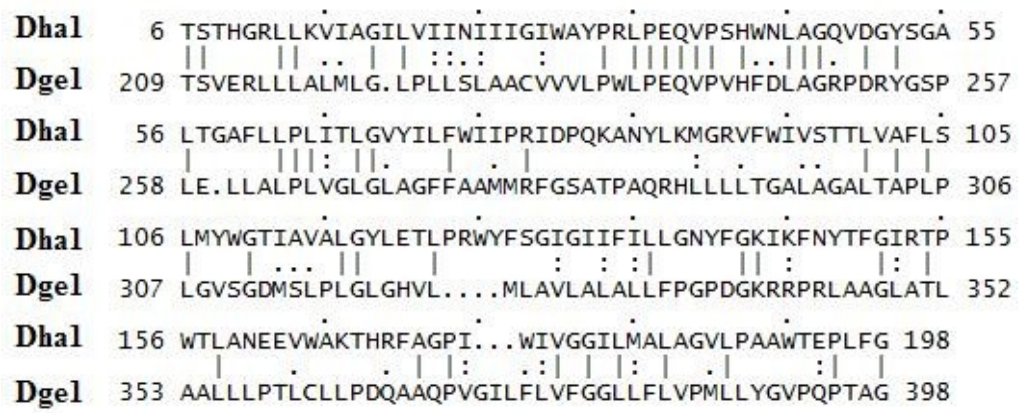

Figure 17: GAP comparison of Dha1(residues 6 to 198), a 6 TMS protein, with Dge1 (residues 209 to 404), a 12 TMS protein using the GAP program.. Quality: 94; Length: 202; Gaps: 4; Percent similarity: 34.8; Percent identity: 26.7. Average quality based on 100 randomizations for 5 runs: 12.6+/-5.3, 13.1 +/-5.6, 13.1 +/-5.7, 13.4 +/-5.6, 13.2 +/-5.6. The average comparison score was 14.6 S.D.

Figure 18: GAP alignment demonstrating the region of homology between varying topological types within the SdpI family. Afu2 (residues 4 to 226), an inverted 6 TMS protein, is aligned with Dge1 (residues 92 to 338), a 12 TMS protein using the GAP program. Quality: 133; Length: 256; Gaps: 12; Percent similarity: 44.4; Percent identity: 34.6. Average quality based on 100 randomizations for 5 runs: 12.6+/-6.2, 13.5 +/-5.7, 13.8 +/-5.6, 12.7 +/-5.5, 13.3 +/-5.7. The average comparison score was 20.9 S.D.

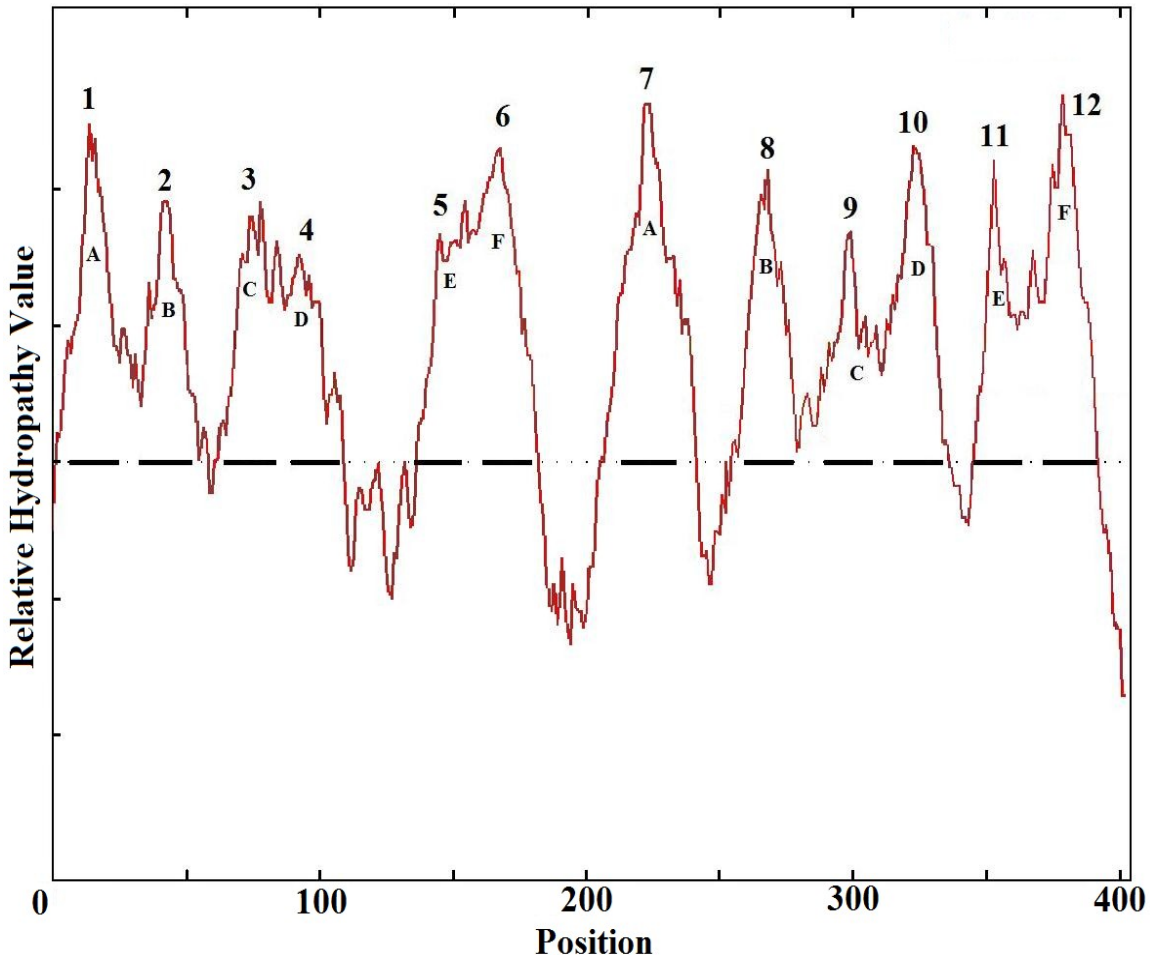

Figure 19: Hydropathy plot of the SdpI homologue from *Deinococcus geothermalis* (Dge1) with numbered TMSs. Letters correspond to the homologous TMSs within the protein that arose through intragenic duplication. The plot was generated using the WHAT program.

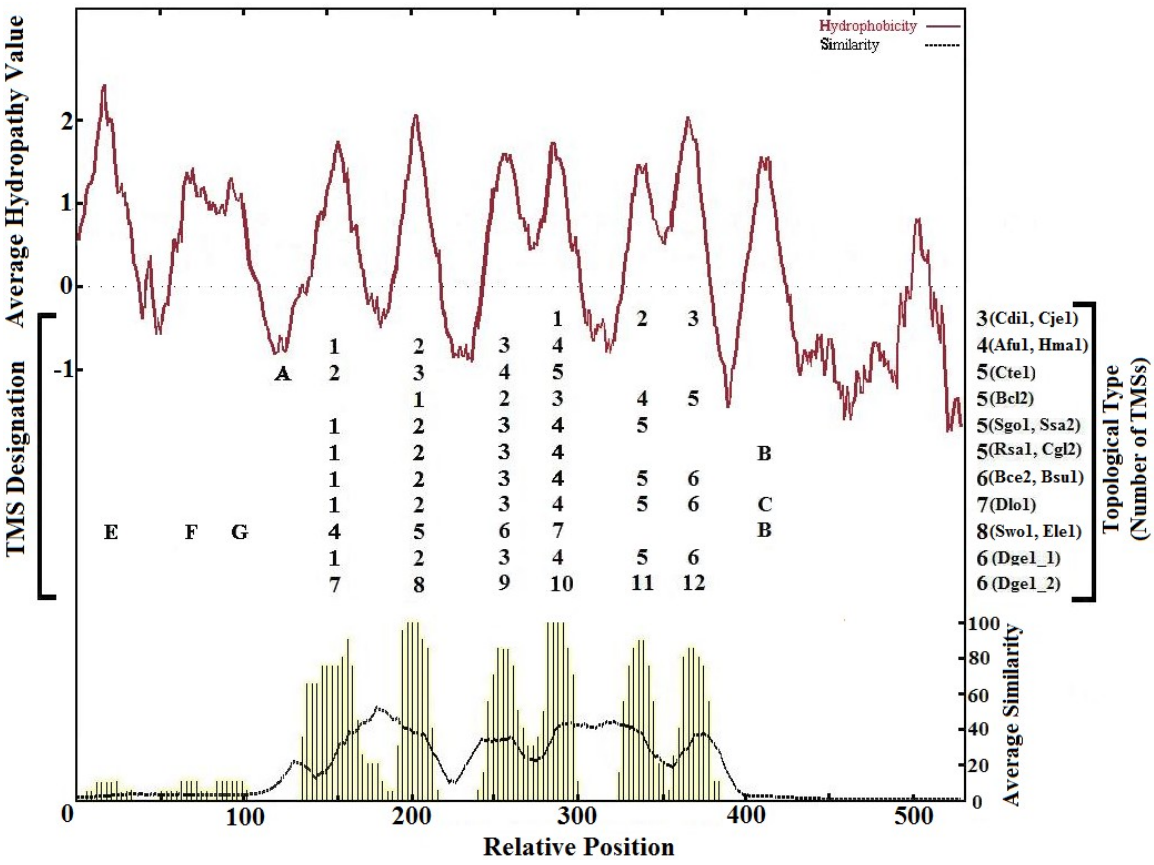

Figure 20: Average hydropathy (top) and similarity (bottom) plots for the SdpI family excluding the 4 inverted proteins Afu2, Tsp3, Ton1 and Tko1 and with the 12 TMS protein Dge1 spliced into two 6 TMS long halves. These plots were generated using the AveHAS program based on the ClustalX multiple alignment shown in Fig. S2 on our website. Between the two plots are the designations of the TMSs which are indicated either by a number (1-12) if conserved between the different groups, or by a letter (A-F) if not conserved among the groups of proteins. At the right, the total numbers of putative TMSs of each topological type are presented. All TMSs in a single vertical column are homologous regardless of the number designations used except for TMSs indicated by letters. The lettered TMSs are not demonstrably homologous to each other or to TMSs in the other homologues within the SdpI family. Note: the letter A marks the region where the first peak of Cte1 aligned, and due to it being the only representative within the SdpI family to have this region, it is poorly displayed in the AveHAS plot. In this alignment, non-conserved regions B and C overlap but are distinct from each other and are not homologous.

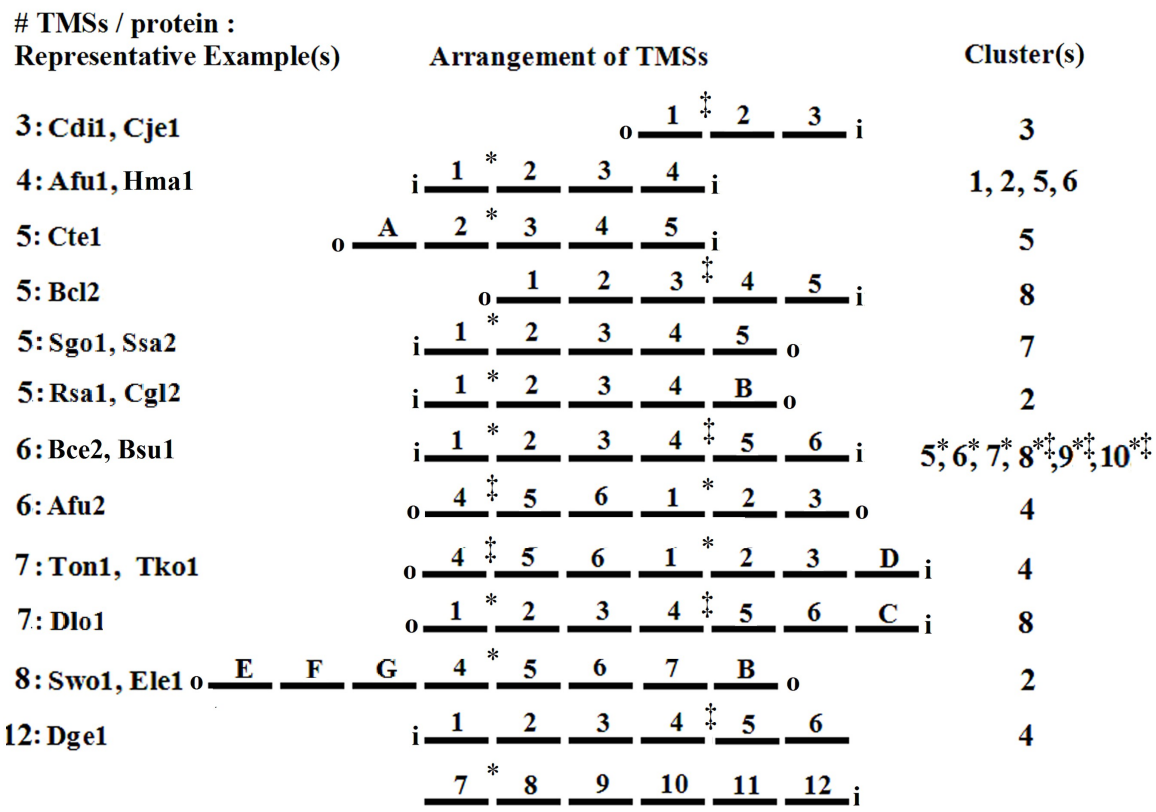

Figure 21: Topological types of proteins of the SdpI family analyzed in this work. The left column lists the number of TMSs in each topological type of protein analyzed together with representative proteins. The centeral column shows the arrangement of the TMSs. The topological types are aligned by regions of homology; that is, TMSs found in the same column are homologous to each other unless they are designated by letter. TMSs indicated by number are conserved throughout the family while TMSs indicated by letter are not conserved. The location of Motif 1 is denoted by ‡. The location of Motif 2 is denoted by *. The right column lists the cluster numbers assigned in the phylogenetic tree (Figure 1) in which proteins of the topological type of the same row are found. i denotes inside of the cell; o denotes outside of the cell.

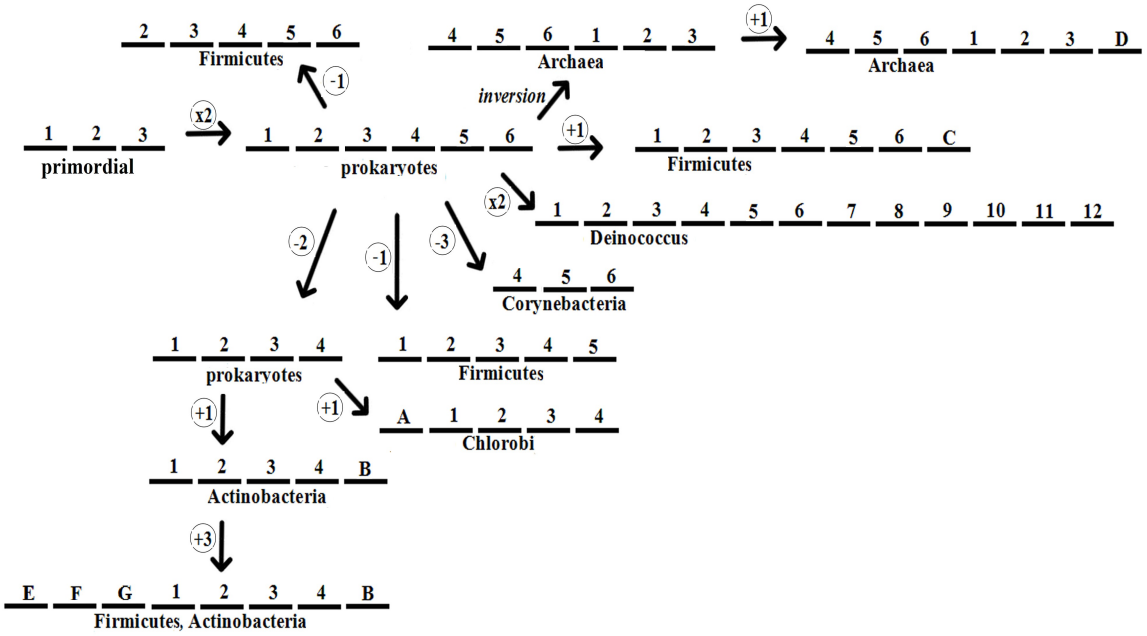

Figure 22: Proposed pathway for the evolution of the proteins of differing topologies within the SdpI family.

TABLES

Table 1: The proteins examined manually for probable orientation within the membrane for which HMMTOP and TMHMM [28-29] gave conflicting results. The top row lists the proteins examined and their respective numbers of TMSs. For each protein, the numbers of K and R residues found in the N- and C- terminus regions and in the loops between TMSs are presented. The numbers of the residues corresponding to the loop or terminal regions examined are also indicated for each protein. Twenty amino acyl residues were examined at each terminus unless TMHMM predicted that fewer residues were found at the N-terminus before the start of the first TMS, as is the case for Cgl1 (6 residues examined), Cje3 (1 residue), and Bcl2, Cte1, and Afu2 whose N-terminal regions started with a TMS and therefore had zero residues examine ("n/a").

Protein Abr.	Cdi1 (3 TMSs)		Cef1 (3 TMSs)		Cgl1 (3 TMSs)		Cje1 (3 TMSs)		Bcl2 (5 TMSs)		Cte1 (5 TMSs)		Afu2 (6 TMSs)	
	Residues examined	# of K + R residues	Residues examined	# of K + R residues	Residues examined	# of K + R residues	Residue examined	# of K + R residues	Residues examined	# of K + R residues	Residues examined	# of K + R residues	Residues examined	# of K + R residues
N-terminus	1-20	2	1-20	4	1-6	0	1	0	N/A	N/A	N/A	N/A	1-6	1
Loop 1 Region	57-89	5	123-151	6	30-56	5	25-53	5	18-25	2	15-29	2	25-46	5
Loop 2 Region	113-116	1	175-178	0	80-83	0	74-79	1	44-52	1	53-75	2	70-75	0
Loop 3 Region									76-107	7	99-125	3	96-121	8
Loop 4 Region									120-135	2	149-162	1	145-164	3
Loop 5 Region													188-201	4
C-terminus	170-190	2	260-280	3	107-170	4	176-196	1	165-175	6	183-189	4	225-228	1

Table 2: Proteins of the SdpI family included in this study, listed alphabetically according to cluster.

Abb.	GenBank Index #	Organismal Source	Protein Size (# aas)	# TMS	Organismal Group
Cluster 1					
Afu1	11497780	*Archaeoglobus fulgidus* DSM 4304	183	4	*Euryarchaeota*
Bce1	89200654	*Bacillus cereus* subsp. cytotoxis NVH 391-98	173	4	Firmicutes
Bli3	52784069	*Bacillus licheniformis* ATCC 14580	168	4	Firmicutes
Bth1	49478191	*Bacillus thuringiensis serovar konkukian* str. 97-27	141	4	Firmicutes
Csp1	86132642	*Cellulophaga* sp. MED134	153	4	*Bacteroidetes*
Oih2	23099993	*Oceanobacillus iheyensis* HTE831	167	4	Firmicutes
Cluster 2					
Ari1	221195540	*Atopobium rimae* ATCC 49626	373	8	Actinobacteria
Cac2	227502806	*Corynebacterium accolens* ATCC 49725	374	8	Actinobacteria
Cgl2	145296541	*Corynebacterium glutamicum* R	238	5	Actinobacteria
Ele1	227411139	*Eggerthella lenta* DSM 2243	371	8	Actinobacteria
Lsp1	169826230	*Lysinibacillus sphaericus* C3-41	353	8	Firmicutes
Rsa1	163839709	*Renibacterium salmoninarum* ATCC 33209	292	5	Actinobacteria
Sau1	57652456	*Staphylococcus aureus* subsp. aureus COL	157	4	Firmicutes
Swo1	114566915	*Syntrophomonas wolfei* subsp. wolfei str. Goettingen	378	8	Firmicutes
Cluster 3					
Cdi1	38234884	*Corynebacterium diphtheriae* NCTC 13129	190	3	Actinobacteria
Cef1	25029421	*Corynebacterium efficiens* YS-314	280	3	Actinobacteria
Cgl1	19554220	*Corynebacterium glutamicum* ATCC 13032	170	3	Actinobacteria
Cje1	68537171	*Corynebacterium jeikeium* K411	196	3	Actinobacteria
Cluster 4					
Afu2	11499784	*Archaeoglobus fulgidus* DSM 4304	228	6	*Euryarchaeota*
Dge1	94985414	*Deinococcus geothermalis* DSM 11300	404	12	*Deinococci*
Tko1	57641858	*Thermococcus kodakarensis* KOD1	264	7	*Euryarchaeota*
Ton1	212225082	*Thermococcus onnurineus* NA1	258	7	*Euryarchaeota*
Tsp3	223478533	*Thermococcus* sp. AM4	267	7	*Euryarchaeota*
Cluster 5					
Aba1	94968429	*Acidobacteria bacterium* Ellin345	303	6	Acidobacteria
Cte1	21674060	*Chlorobium tepidum* TLS	189	5	*Chlorobi*
Pae1	68552512	*Prosthecochloris aestuarii* DSM 271	170	4	*Chlorobi*
Cluster 6					
Ban3	30261395	*Bacillus anthracis* str. Ames	201	6	Firmicutes
Bce3	30020208	*Bacillus cereus* ATCC 14579	205	6	Firmicutes
Bce4	89200937	*Bacillus cereus* subsp. cytotoxis NVH 391-98	194	6	Firmicutes
Bce8	47566179	*Bacillus cereus* G9241	201	6	Firmicutes
Bce9	52143342	*Bacillus cereus* E33L	205	6	Firmicutes
Bce10	30019445	*Bacillus cereus* ATCC 14579	205	6	Firmicutes
Bth4	49479775	*Bacillus thuringiensis serovar konkukian* str. 97-27	201	6	Firmicutes
Bth5	75764858	*Bacillus thuringiensis serovar israelensis* ATCC 35646	201	6	Firmicutes
Bth6	75761225	*Bacillus thuringiensis serovar israelensis* ATCC 35646	208	6	Firmicutes
Bwe2	89204480	*Bacillus weihenstephanensis* KBAB4	201	6	Firmicutes
Hma1	55378946	*Haloarcula marismortui* ATCC 43049	137	4	*Euryarchaeota*
Lmo1	16804608	*Listeria monocytogenes* EGD-e	204	6	Firmicutes
Cluster 7					
Lpl1	28378914	*Lactobacillus plantarum* WCFS1	208	6	Firmicutes
Sgo1	157149986	*Streptococcus gordonii* str. Challis substr. CH1	165	5	Firmicutes
Ssa2	125717586	*Streptococcus sanguinis* SK36	165	5	Firmicutes
Ssu1	81097456	*Streptococcus suis* 89/1591	200	6	Firmicutes
Cluster 8					
Cbe1	82746983	*Clostridium beijerincki* NCIMB 8052	210	6	Firmicutes
Cdi2	90574392	*Clostridium difficile* QCD-32g58	213	6	Firmicutes
Lme1	116617456	*Leuconostoc mesenteroides* subsp. mesenteroides ATCC 8293	211	6	Firmicutes
Sin1	2239172	*Streptococcus iniae*	210	6	Firmicutes
Smu1	24380024	*Streptococcus mutans* UA159	212	6	Firmicutes

Table 2: Continued.

Cluster 9					
Bcl2	56965759	*Bacillus clausii* KSM-K16	175	5	Firmicutes
Bsu1	16080431	*Bacillus subtilis* subsp. subtilis str. 168	207	6	Firmicutes
Cpe1	110802548	*Clostridium perfringens* SM101	199	6	Firmicutes
Cth1	67875454	*Clostridium thermocellum* ATCC 27405	199	6	Firmicutes
Dlo1	153853119	*Dorea longicatena* DSM 13814	339	7	Firmicutes
Lpl2	28378259	*Lactobacillus plantarum* WCFS1	192	6	Firmicutes
Lpl3	28379444	*Lactobacillus plantarum* WCFS1	200	6	Firmicutes
Lsa1	90962640	*Lactobacillus salivarius* subsp. salivarius UCC118	197	6	Firmicutes
Cluster 10					
Bad1	85667575	*Bifidobacterium adolescentis*	240	6	Actinobacteria
Bce2	42784033	*Bacillus cereus* ATCC 10987	212	6	Firmicutes
Bce5	30022902	*Bacillus cereus* ATCC 14579	211	6	Firmicutes
Bce6	47568007	*Bacillus cereus* G9241	211	6	Firmicutes
Bce7	52140669	*Bacillus cereus* E33L	211	6	Firmicutes
Bcl1	56965474	*Bacillus clausii* KSM-K16	212	6	Firmicutes
Bli2	52079220	*Bacillus licheniformis* ATCC 14580	212	6	Firmicutes
Bth3	75759285	*Bacillus thuringiensis serovar israelensis* ATCC 35646	211	6	Firmicutes
Bwe1	89204331	*Bacillus weihenstephanensis* KBAB4	211	6	Firmicutes
Ccr1	16127257	*Caulobacter crescentus* CB15	225	6	Alphaproteobacteria
Chy1	78044771	*Carboxydothermus hydrogenoformans* Z-2901	222	6	Firmicutes
Cph1	106885445	*Clostridium phytofermentans* ISDg	217	6	Firmicutes
Det1	57233995	*Dehalococcoides ethenogenes* 195	221	6	*Chloroflexi*
Dha1	89896096	*Desulfitobacterium hafniense* Y51	221	6	Firmicutes
Dsp1	88933845	*Dehalococcoides* sp. BAV1	221	6	*Chloroflexi*
eur1	71394057	uncultured *euryarchaeote* Alv-FOS5	206	6	*Euryarchaeota*
Fba1	89890638	*Flavobacteria bacterium* BBFL7	217	6	*Bacteroidetes*
Gka1	56420668	*Geobacillus kaustophilus* HTA426	214	6	Firmicutes
Hor1	89210783	*Halothermothrix orenii* H 168	222	6	Firmicutes
Iba1	85712133	*Idiomarina baltica* OS145	220	6	Gammaproteobacteria
Mac1	20091953	*Methanosarcina acetivorans* C2A	227	6	*Euryarchaeota*
Mhu1	88603182	*Methanospirillum hungatei* JF-1	212	6	*Euryarchaeota*
Mba1	73669446	*Methanosarcina barkeri* str. Fusaro	219	6	*Euryarchaeota*
Mma1	21226485	*Methanosarcina mazei* Go1	213	6	*Euryarchaeota*
Mma2	114571457	*Maricaulis maris* MCS10	230	6	Alphaproteobacteria
Mth1	83590912	*Moorella thermoacetica* ATCC 39073	223	6	Firmicutes
Oal1	83859055	*Oceanicaulis alexandrii* HTCC2633	228	6	Alphaproteobacteria
Pth1	98659796	*Pelotomaculum thermopropionicum* SI	229	6	Firmicutes
Rbi1	88804820	*Robiginitalea biformata* HTCC2501	216	6	*Bacteroidetes*
Sth1	51892521	*Symbiobacterium thermophilum* IAM 14863	225	6	Actinobacteria
Tet1	76795994	*Thermoanaerobacter ethanolicus* ATCC 33223	220	6	Firmicutes
Tte1	20807164	*Thermoanaerobacter tengcongensis* MB4	220	6	Firmicutes

Table 3: Summary of the similarities and differences within and between the sequence of Motifs 1 (A) and 2 (B) among clusters. (C) shows an alignment of Motif 1 (M1) with Motif 2 (M2)1

A

Cluster	Motif 1
3	[VL]L L L V XGXL[GA]W X AKL P GNPVVG I RV PE VR K S QELWDMAHRVAGPL WVL S V
4	L[TL]L L A AGL L T X A FRN R RNXXI G FRXGY TY X S DRA WRKT NTFAGI F XVAX S
7	V A L V F[IM]I I G N Y P KKV Q XNRXXMNRKLA XL F V GIG F L X FI VAI FYL
8	L G I[ML][FLI]IV I G N Y[LM]PKX[KR]QNYTVG I RL PWTL[ND][ND]EXNWNKTHRLAGKLWV LGG
9	L G I M X IV I G N I X PKX K QNRXI G I RT PWTL S S EEVW KKTNXLXGKL LV I VG
10	[IV]G L L F IV L G N Y L PKC K PNYFVG I RT PWTL S S EEVW RKTHRFGGKV FV I LG

B

Cluster	Motif 2
1	I I[YW]G X L P EE V PAH Y N A XGE V D R W G S KW EL
2	[VL]H Y P X[ML]PDX V PVHW N[GF]SGE[AV]D X W X X KA DK
4	L L W D K L P EX V AXH F [ND]A SGE P D X Y X S KF AG
5	Y H Y X X L PXX I AXH F G F SGX P D A W G P KG PK
6	C L Y P Y L P NR I AVHW[ND] F NGX P N E F M S KQ VV
7	W V Y X X L PXN[LM]A IH F D L SGK[GP] N[AG]F L P K LP K
8	X L W X[KQ]L PEE [VI] A IH F D X SGX P D[GN]Y[AS][PS]KV XA
10	A A Y P H L PDT I PTH W N S SGE V D G Y G S S K FE

C

```
M1  1 VLLILIGLLFIVLGNY..LPKAKPNYFVGIRTPWTLSSE.EVWRKTHRLGGKLFVIGGILML 59
M2  4 LLTLLLGPLILTIALYPALPERVPVH.......WNASGEVDGYGSKFEGGSKL...EGVFLL 55
```

1 A and B: *, an identified; :, a close similarity; ., a more distant similarity within the cluster as defined by the MEME program. Note: for Cluster 7, Motif 1 was only found for the Lpl1 protein by the MEME program, but the motif above is conserved within the cluster based on the alignment obtained with the ClustalX program. C: |, and identity; :, a close similarity; ., a more distant similarity as defined by the GAP program. The residues indicated in Motifs 1 and 2 are the dominant residues at the various aligned positions.

DISCUSSION

Evolutionary origins of varying topological types

It is likely that the standard 6 TMS proteins represent the basic element of the SdpI family. Several other membrane protein families with members possessing 6 TMSs per polypeptide chain are known to have arisen through either internal triplication of a primordial 2 TMS element (CytC (Lee *et al.*, 2007), MC (Kuan & Saier, 1993), and ABC1 (Wang *et al.*, 2009) or by duplication of a primordial 3 TMS element (MIP (Pao *et al.*, 1991), DsbD (Kimball *et al.*, 2003) and ABC2 (Wang *et al.*, 2009). We suggest that other topological types within the SdpI family arose from this basic 6 TMS element. We further suggest that deletions in this basic element led to the formation of the proteins of 4 and 3 TMSs as well as two of the four 5 TMS topological variants. Deletion and fusion events led to evolution of the two other 5 TMS protein variants and to the 8 TMS proteins, respectively. A fusion event possibly led to the creation of the non-inverted 7 TMS protein, and an inversion of the two 3 TMS halves of the 6 TMS proteins led to the appearance of the inverted 6 TMS protein, Afu2, as well as the inverted 7 TMS proteins, Tko1, Tsp3 and Ton1. The inverted 7 TMS proteins may have also undergone a C-terminal fusion event generating an extra TMS. Finally, the single 12 TMS protein (Dge1) undoubtedly arose by intragenic duplication followed by extensive sequence divergence within both halves.

Protein orientation within the cell membrane

All of the proteins of the SdpI family included in our study proved to be oriented

within the cell membrane (Figure 21) in such a way that Motif 1, between TMSs 4 and 5 in the standard 6 TMS proteins, is always located on the inside, facing the cytoplasm, while Motif 2 is always found to be externally localized. The N-termini of the four 3 TMS homologues, all of the inverted 7 TMS proteins, Bcl2 (5 TMSs) and Cte1 (5 TMSs) were predicted to be localized to the external surface of the cell membrane, and the C-termini were predicted by both programs to be on the inside. Both the N- and C-termini of the 4 TMS proteins, the standard 6 TMS proteins and the duplicated 12 TMS protein were predicted to be located on the inside. Both the N- and C-termini of the inverted 6 TMS and 8 TMS proteins appeared to be located on the outside. The N-termini of the standard 7 TMS homologue (Dlo1) and four of the 5 TMS variants (Rsa1/Cgl2 and Ssa2/ Sgo1 – see Figure 21) were predicted to be localized to the inside of the cell, while the C-termini were predicted to be on the outside. Based on all of these predicted orientations, which were in surprising agreement with each other, Motif 1 is always in the cytoplasm, while Motif 2 is always on the external surface to the membrane. As we postulate that Motif 1 is responsible for promoting expression of the *sdpRI* operon by sequestering the autorepressor, SdpR, it would follow that this process occurs on the inside of the membrane. By contrast, since Motif 2 is predicted to be responsible for neutralizing the SdpC toxin by forming an SdpI-SdpC complex in the membrane, Motif 2 should be localized to the outer surface of the cellular membrane. The predicted topologies therefore fully support the functional predictions.

Conserved motifs confirm homology of SdpI family members

Analysis of the motifs present in the proteins of the SdpI family confirmed homology of most family members despite variations in their topologies. Figure 21 illustrates the alignment of the proteins according to their topologies with the locations of the two conserved motifs denoted. Motif 1, when present, is always found between TMSs 4 and 5 in the standard 6 TMS homologues, while Motif 2, when present, is always found between TMSs 1 and 2 of the standard 6 TMS proteins. Thus, when these motifs are found in the other topologically variant proteins, they are always located in the region that would be expected to exhibit the motif in question within the standard 6 TMS proteins. These hydrophilic loops proved to be the best conserved regions of these proteins as revealed by the average similarity plots generated with AveHAS program (Figure 20).

Motif analysis of the four inverted proteins confirmed the proposed inversion. Motif 1, located in the hydrophilic region between TMSs 1 and 2 of the inverted proteins, is homologous to the hydrophilic region between TMSs 4 and 5 of the standard 6 TMS proteins. Further, Motif 2 is found in the region between TMSs 4 and 5 in the inverted proteins which is homologous to the hydrophilic region between the first and second TMSs of the standard 6 TMS proteins. This occurrence provides further evidence for the inversion proposed initially on the basis of primary sequence similarity alone.

The clustering of the single 4 TMS protein, Hma1 (Cluster 6), with all of the 6 TMS proteins in cluster 6 can be rationalized based on our motif analyses. Cluster 6 contains 6 TMS proteins which only exhibit Motif 2, and Hma1 also contains only Motif 2. This is expected as Hma1 is homologous to TMSs 1-4 of the standard 6 TMS proteins

and lacks the hydrophilic region between TMSs 4 and 5. Possibly it arose independently of the other 4 TMS proteins of the SdpI family by deletion of the C-terminus of a 6 TMS homologue like those with which it clusters.

The same principle can be applied to explain the origins of the 4, 5 and 6 TMS proteins (Pae1, Cte1 and Aba1) within cluster 5. All three proteins contain only Motif 2 and are very closely related, leading to the possibility that a 6 TMS precursor underwent C-terminal deletions, yielding the 4 and 5 TMS proteins.

It is likely that the original 6 TMS proteins contained the equivalent of primordial Motifs 1 and 2. These 6 TMS proteins are highly similar and align with one another throughout their lengths. Consequently, there is no reason to support the idea that convergent evolution led to the appearance of the two motifs. More likely, some of the 6 TMS proteins lost one or the other motif and lost the corresponding function or had the same motif diverge in sequence to an unrecognizable state while gaining a dissimilar function. Lpl1 of Cluster 7 can serve as an example in support of this hypothesis. Both motifs were recognized by MEME in Lpl1, although this program recognized only Motif 2 in the rest of the proteins in this cluster.

The SdpI family is unusual in that it contains proteins of widely varying topologies. Such a situation has rarely been observed, the only other well documented example being the Heme Handling Protein (HHP) Family (TC# 9.B.14; (Lee *et al.*, 2007)). We propose two possible explanations for this phenomenon. First, it is possible that the entirety of the protein is not necessary for function; Motif 1 between TMSs 4-5 or Motif 2 between TMSs 1-2 may alone be adequate for one of the subfunctions currently recognized for the SdpI protein of *Bacillus subtilis*. Second, the truncated versions of the

6 TMS proteins and the 6 TMS proteins containing only one recognizable motif form heterodimers to ensure a complex possessing both of the conserved motifs. In either case, the diverse topological types can be attributed to the two dissimilar functions as, for example, in binding SdpC, and in binding SdpR, respectively, to SdpI as suggested by the work of Ellermeier *et al.* (2006) and as elaborated in the next paragraph.

The NCBI database was searched with Motifs 1 and 2 but no significant matches were found outside of the SdpI family. The work of Ellermeier *et al.* (2006) provides a functional explanation for the topological variants within the members of the SdpI family. The first 3 TMSs of the 6 TMS SdpI protein are responsible for the SdpC immunity function while the second 3 TMSs are responsible for SdpR sequestration. All of the topological variants within the family include at least one of the regions that is potentially responsible for one of the functions. Proteins with 3, 4, 5 and 8 TMSs may be unifuctional because they only contain the first three or second three TMSs of the 6 TMS proteins. Proteins with 6, 7 or 12 TMSs would be predicted to have both functions. Since both functions are needed to ensure regulated immunity to SdpC, it is reasonable to postulate that an organism could have two unifunctional proteins to compensate for not having a protein with both functions in a single polypeptide chain. Alternatively, an organism may have just one or the other function, e.g., unregulated immunity, or regulation of a dissimilar function.

Corynebacterium glutamicum and *C. efficiens* have two SdpI homolougues, a 3 TMS protein (e.g., Cgl1) and a 5 TMS protein (e.g., Cgl2). The 3 TMS protein is homologous to the second half (TMSs 4-6) of the standard 6 TMS proteins, the region that is believed to be responsible for promoting the expression of the *sdpRI* operon by

sequestering the autorepressor, SdpR. The 5 TMS protein is homologous to TMSs 1-4 of the standard 6 TMS proteins, the region in SdpI that is probably responsible for neutralizing the SdpC toxin by forming an SdpI-SdpC complex in the membrane. By having two truncated proteins with complementary functions, possibly in complex with each other, regulated SdpC immunity could therefore involve two related but dissimilar proteins.

The two representatives from *Corynebacterium glutamicum* that are part of this study are from two different strains, with Cgl1 being from *C. glutamicum* ATCC 13032 and Cgl2 being from *C. glutamicum* R. The genomes of both proteins were searched for their potential complementary-functional counterparts, and in both genomes these proteins were located. By BLASTing the genome of *C. glutamicum* ATCC 13032 with the 5 TMS protein, Cgl2, a corresponding 5 TMS protein (gi # 19553748) was found. By BLASTing the genome of *C. glutamicum* R with the 3 TMS protein, Cgl1, a corresponding 3TMS protein (gi # 145297017) was also located. The two proteins had been arbitrarily excluded from this study by use of the CD-Hit program for the elimination of redundancies and very close sequences. The existence of a 3TMS (Motif 1 present) and a 5 TMS (Motif 2 present) protein within the same organism substantiates the postulate that protein complementarity may occur for proteins with only one of the two motifs.

Evidence that the 6 TMS topology arose by duplication of a 3 TMS precursor

Several independent lines of evidence lead us to suggest that duplication of a primordial 3 TMS element, followed by substantial sequence divergence, gave rise to the

major class of 6 TMS proteins. (1) The best-conserved motifs occur between TMSs 1 and 2 and TMSs 4 and 5, equivalent positions in the two halves of the protein. (2) Assuming that these conserved motifs in the two halves of SdpI bind SdpC (the toxin) and SdpR (the regulator), respectively, with the N- and C-termini inside, then SdpC would bind to the external surface of the membrane while SdpR would bind to the cytoplasmic side, as is likely, based on mutational analyses (Ellermeier *et al.*, 2006). Opposite orientation of repeat units in the membrane is always observed when an odd number of TMSs is duplicated (Saier, 2003). (3) Comparison of the sequences of Motif 1 with those of Motif 2 revealed similarities, suggestive of homology, even though the observed similarity was not sufficient to establish common origin (Table 2C). (4) Binding of SdpC and SdpR to the first and second halves of the membrane as suggested by Ellermeier *et al.* (2006), could be explained if the two halves of the 6 TMS SdpI protein arose from a 3 TMS protein binding precursor polypeptide. Sequence divergence would allow the two halves of SdpI to bind two structurally unrelated proteins, SdpC and SdpR. (5) The fact that several SdpI homologues exhibit an inverted topology makes functional sense since these two 3-TMS halves have distinct protein-binding functions. (6) The same argument can be used to explain conservation within the 12 TMS homologue. The second 3 TMS element within the first 6 TMS half of the protein, and the first 3 TMS element within the second 6 TMS half, proved to be better conserved than the first 3 TMS element in the first half and the second 3 TMS element in the second half. This would suggest that only second and third 3-TMS elements in this duplicated 12 TMS protein have retained function. The first and fourth 3-TMS elements may have diverged in sequence with concomitant loss of functionality (Figure 21).

Taken together, these observations suggest an origin of SdpI homologues comparable to those of the MIP (Pao *et al.*, 1991), DsbD (Kimball *et al.*, 2003) and ABC2 families (Wang *et al.*, 2009), namely, duplication of a 3-TMS-encoding genetic element. Further work, including the generation of high resolution 3-dimensional structural data, is likely to provide confirmation or refutation of this proposal.

This section, in full, is a reprint of the material as it will appear in The SdpI Family of Antibiotic Peptide Killer Factor Immunity Proteins. Povolotsky, Tatyana Leonidovna; Orlova, Ekaterina; Pandey, Rachna; Tamang, Dorjee G.; Saier, Milton H., Jr. The thesis author is the primary investigator and author of this paper.

REFERENCES

Aguilar, C., Vlamakis, H., Losick, R. & Kolter, R. (2007). Thinking about *Bacillus subtilis* as a multicellular organism. *Curr Opin Microbiol* **10**, 638-643.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-410.

Bailey, T. L. & Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-29.

Barak, I. & Wilkinson, A. J. (2005). Where asymmetry in gene expression originates. *Mol Microbiol* **57**, 611-620.

Britton, R. A., Eichenberger, P., Gonzalez-Pastor, J. E., Fawcett, P., Monson, R., Losick, R. & Grossman, A. D. (2002). Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of *Bacillus subtilis*. *J Bacteriol* **184**, 4881-4890.

Butcher, B. G. & Helmann, J. D. (2006). Identification of *Bacillus subtilis* sigma-dependent genes that provide intrinsic resistance to antimicrobial compounds produced by Bacilli. *Mol Microbiol* **60**, 765-782.

Chung, J. D., Stephanopoulos, G., Ireton, K. & Grossman, A. D. (1994). Gene expression in single cells of *Bacillus subtilis*: evidence that a threshold mechanism controls the initiation of sporulation. *J Bacteriol* **176**, 1977-1984.

Claverys, J. P. & Havarstein, L. S. (2007). Cannibalism and fratricide: mechanisms and raisons d'etre. *Nat Rev Microbiol* **5**, 219-229.

Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983). Establishing homologies in protein sequences. *Methods Enzymol* **91**, 524-545.

Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* **12**, 387-395.

Doolittle, R. F. (1981). Similar amino acid sequences: chance or common ancestry? *Science* **214**, 149-159.

Dworkin, J. & Losick, R. (2005). Developmental commitment in a bacterium. *Cell* **121**, 401-409.

Eichenberger, P., Fujita, M., Jensen, S. T. & other authors (2004). The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus*

subtilis. *PLoS Biol* **2**, e328.

Ellermeier, C. D., Hobbs, E. C., Gonzalez-Pastor, J. E. & Losick, R. (2006). A three-protein signaling pathway governing immunity to a bacterial cannibalism toxin. *Cell* **124**, 549-559.

Errington, J. (2003). Regulation of endospore formation in *Bacillus subtilis*. *Nat Rev Microbiol* **1**, 117-126.

Fujita, M. & Losick, R. (2002). An investigation into the compartmentalization of the sporulation transcription factor sigmaE in *Bacillus subtilis*. *Mol Microbiol* **43**, 27-38.

Gonzalez-Pastor, J. E., Hobbs, E. C. & Losick, R. (2003). Cannibalism by sporulating bacteria. *Science* **301**, 510-513.

Grossman, A. D. (1995). Genetic networks controlling the initiation of sporulation and the development of genetic competence in *Bacillus subtilis*. *Annu Rev Genet* **29**, 477-508.

Hecker, M. & Volker, U. (2001). General stress response of *Bacillus subtilis* and other bacteria. *Adv Microb Physiol* **44**, 35-91.

Kimball, R. A., Martin, L. & Saier, M. H., Jr. (2003). Reversing transmembrane electron flow: the DsbD and DsbB protein families. *J Mol Microbiol Biotechnol* **5**, 133-149.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567-580.

Kuan, J. & Saier, M. H., Jr. (1993). The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Crit Rev Biochem Mol Biol* **28**, 209-233.

Lee, J. H., Harvat, E. M., Stevens, J. M., Ferguson, S. J. & Saier, M. H., Jr. (2007). Evolutionary origins of members of a superfamily of integral membrane cytochrome c biogenesis proteins. *Biochim Biophys Acta* **1768**, 2164-2181.

Li, W., Jaroszewski, L. & Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282-283.

Li, W., Jaroszewski, L. & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18**, 77-82.

Molle, V., Fujita, M., Jensen, S. T., Eichenberger, P., Gonzalez-Pastor, J. E., Liu, J. S. & Losick, R. (2003). The Spo0A regulon of *Bacillus subtilis*. *Mol Microbiol* **50**, 1683-

1701.

Pao, G. M., Wu, L. F., Johnson, K. D., Hofte, H., Chrispeels, M. J., Sweet, G., Sandal, N. N. & Saier, M. H., Jr. (1991). Evolution of the MIP family of integral membrane transport proteins. *Mol Microbiol* **5**, 33-37.

Parker, G. F., Daniel, R. A. & Errington, J. (1996). Timing and genetic regulation of commitment to sporulation in *Bacillus subtilis*. *Microbiology* **142 (Pt 12)**, 3445-3452.

Saier, M. H., Jr. (1994). Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol Rev* **58**, 71-93.

Saier, M. H., Jr. (2003). Tracing pathways of transport protein evolution. *Mol Microbiol* **48**, 1145-1156.

Saier, M. H., Jr., Yen, M. R., Noto, K., Tamang, D. G. & Elkan, C. (2009). The Transporter Classification Database: recent advances. *Nucleic Acids Res* **37**, D274-278.

Smith, R. F., Wiese, B. A., Wojzynski, M. K., Davison, D. B. & Worley, K. C. (1996). BCM Search Launcher--an integrated interface to molecular biology data base search and analysis services available on the World Wide Web. *Genome Res* **6**, 454-462.

Steil, L., Hoffmann, T., Budde, I., Volker, U. & Bremer, E. (2003). Genome-wide transcriptional profiling analysis of adaptation of *Bacillus subtilis* to high salinity. *J Bacteriol* **185**, 6358-6370.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876-4882.

Tusnady, G. E. & Simon, I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849-850.

von Heijne, G. & Gavel, Y. (1988). Topogenic signals in integral membrane proteins. *Eur J Biochem* **174**, 671-678.

Vreeland, R. H., Rosenzweig, W. D. & Powers, D. W. (2000). Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature* **407**, 897-900.

Wang, B., Dukarevich, M., Sun, E. I., Yen, M. R. & Saier Jr, M. H. (2009). Membrane Porters of ATP-binding Cassette (ABC) Transport Systems are Polyphyletic. *J Membrane Biol (submitted)*.

Yen, M. R., Choi, J. & Saier Jr, M. H. (2009). Bioinformatic Analyses of

Transmembrane transport: Novel Software for Deducing Protein Phylogeny, Topology, and Evolution. *J Mol Microb Biotech (in press)*.

Zhai, Y. & Saier, M. H., Jr. (2001a). A web-based program for the prediction of average hydropathy, average amphipathicity and average similarity of multiply aligned homologous proteins. *J Mol Microbiol Biotechnol* **3**, 285-286.

Zhai, Y. & Saier, M. H., Jr. (2001b). 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* **3**, 501-502.

Zhai, Y. & Saier, M. H., Jr. (2002). A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins. *J Mol Microbiol Biotechnol* **4**, 375-377.

Zhai, Y., Tchieu, J. & Saier, M. H., Jr. (2002). A web-based Tree View (TV) program for the visualization of phylogenetic trees. *J Mol Microbiol Biotechnol* **4**, 69-70.