UC San Diego
UC San Diego Previously Published Works

Title

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

Permalink

Journal

ISSN

Authors

Povolotsky, Tatyana Leonidovna
Orlova, Ekaterina
Tamang, Dorjee G.
et al.

Publication Date

DOI

Peer reviewed

ORIGINAL ARTICLES

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

Tatyana Leonidovna Povolotsky · Ekaterina Orlova ·
Dorjee G. Tamang · Milton H. Saier Jr.

Abstract 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 artifacts due to sequencing or annotation errors. 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 deletional and fusional events, as well as intragenic duplications and inversions, may have yielded SdpI homologues with topologies of varying numbers and positions of TMSs. We propose a specific evolutionary pathway that could have given rise to these distantly related bacterial immunity proteins. We further show that genes encoding SdpI homologues often appear in operons with genes for homologues of SdpR, SdpI's autorepressor. Our analyses allow us to propose structure–function relationships that may be applicable to most family members.

Introduction

Inhospitable environmental conditions prompt microbes to respond to stress by inducing the expression of stress response genes (Barak and Wilkinson 2005; Hecker and 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-h period (Britton et al. 2002; Eichenberger et al. 2004; Fujita and Losick 2002; Molle et al. 2003; Steil et al. 2003). Because this process becomes irreversible after approximately 2 h (Dworkin and Losick 2005; Parker et al. 1996), mechanisms exist that delay commitment to this process through cannibalism (Claverys and Havarstein 2007). The SdpI protein of *Bacillus subtilis* is involved in orchestrating one such delay (Ellermeier et al. 2006). It is a transmembrane protein 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

T. L. Povolotsky and E. Orlova contributed equally to this work.

T. L. Povolotsky · E. Orlova · D. G. Tamang ·
M. H. Saier Jr. (✉)
Division of Biological Sciences, University of California,
San Diego, La Jolla, CA 92093-0116, USA
e-mail: msaier@ucsd.edu

on in about half of the cells in the population (Chung et al. 1994; Fujita and Losick 2002; Gonzalez-Pastor et al. 2003). Spo0A-ON cells switch on transcription of two operons, *skfA-H* and *sdpABC*. 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).

B. subtilis Spo0A-ON 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. Through its interaction with SdpI, the transcription of an adjacent convergently transcribed immunity operon, *sdpRI,* is induced. Thus, 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 derepressed so that more *sdpI* is transcribed and translated. This mechanism 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 associated with 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 (Ellermeier et al. 2006). Other forms of resistance to SdpC have been identified: σ^w-dependent operons *yknWXYZ* and *yfhL* confer immunity to SdpC. *yknWXYZ* encodes an ABC transporter and is speculated to export the SdpC toxin, while *yfhL* encodes a paralogue of SdpI (Butcher and Helmann 2006).

Here, we use established bioinformatic methodologies to show that the basic element of the SdpI family is a 6 transmembrane α-helical segment (TMS) protein. This basic structure probably derived from a primordial 3 TMS

element by intragenic duplication. The resultant 6 TMS protein then underwent another duplication, followed or occasionally preceded by deletion and possible 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 the *B. subtilis* SdpI where the first 3 TMS half of this protein serves one function (external binding of SdpC and immunity), while the second 3 TMS half serves another (internal binding of SdpR and signal transduction) (Ellermeier et al. 2006). In addition, we demonstrate the existence of homologues of SdpR and other transcriptional regulators within the operons that code for SdpI homologues.

Methods

Selection of Protein Sequences

A Basic Local Alignment Search Tool (BLAST) search (Altschul et al. 1990) was performed in October 2007 using the SdpI protein of *Bacillus subtilis* (gi no. 16080431) as the query sequence with two iterations and the default cutoff. More than 100 homologous proteins were retrieved from the National Center for Biotechnology Information (NCBI) database. Eighty-two proteins were retained for topological analysis after redundancies and proteins with greater than 90% identity were eliminated by a modified CD-Hit program (Li et al. 2001, 2002; Saier et al. 2009; Yen et al. 2009). The proteins were further reduced in number to 76 after translating the DNA in all six reading frames and seeking sequence similarities with full-length close homologues of the three translated codirectional reading frames.

The program BCM Search Launcher (Smith et al. 1996) was used to translate the DNA coding for the query protein in the six reading frames at both ends flanking the existing 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 before or after 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 amino acyl residues (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 analyses 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. The occurrence of multiple such homologues of a specific topology provided confirmation of this assumption.

A second BLAST search was performed on May 21, 2009, using the SdpI protein of *Bacillus cereus*, Bce2 (gi no. 42784033), a close orthologue of the *B. subtilis* protein, 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 cutoff of e^{-4} for the first iteration and a cutoff 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 previously existing family. Nine proteins were added to the original list. The DNA translating procedure which used the program BCM Search Launcher was also applied to the nine added 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 and Saier 2002; 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 by the WHAT (Zhai and Saier 2001a) and AveHAS (Zhai and Saier 2001b) 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 aas. Statistical sequence similarity comparisons between proteins, and between internal regions of these proteins, were conducted by the IC (Zhai and 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 500 random shuffles have been shown to be satisfactory for

obtaining statistically significant values (Yen et al. 2009). A value of 10 standard deviations (SD) for comparable regions of two proteins of at least 60 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 based on hydropathy analyses, because none of the proteins in this family has been characterized topologically.

Motif Analyses

All of the SdpI proteins within our study were analyzed for motifs using the MEME program (Bailey and 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 estimated by the HMMTop (Tusnady and 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 aas of the primary sequence (unless otherwise specified; see Table S1 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 and Saier 2001a). The positive-inside rule was then applied to determine orientation of the proteins within the cell membrane (von Heijne and 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 because the TMHMM program did not recognize all of the putative TMSs.

Operon Analyses

Three representative proteins were chosen from every topological group of proteins (i.e., three from the proteins with 3 TMSs, three from the 4 TMS proteins, etc.). The encoding operons were examined using the Genome Browser feature of the National Microbial Pathogen Data Resource (NMPDR) database (McNeil et al. 2007). Some proteins were excluded from the analysis if their genomes were not yet represented in the NMPDR genome database or if a genome, although represented, was not yet populated with genes in the vicinity of the purported locations of the *sdpI* homologues included in the analysis. *sdpI* homologues were considered to be in the same operon with other cistrons if the distance between them was between −8 and 30 bp. Elements suspected of being in the same operon as the *sdpI* homologue were identified using BLAST searches, and the results are tabulated in Table S2. Within the BCM Search Launcher's Gene Feature Search, the Prokaryotic Promoter Prediction by Neural Network (Smith et al. 1996) was used to find promoters with a score cutoff of 0.80 upstream of the alleged operons to verify their legitimacy.

Results

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

Classification of Organisms Possessing SdpI Family Members

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), and *Deinococcus* (1). The proteins vary widely in size, with sequences as short as 137 aas (Hma1 from *Haloarcula marismortui*) and as long as 404 aas (Dge1 from *Deinococcus geothermalis*). The majority of the proteins are of a size near 200 (170–230) aas in length and exhibit putative 6 TMS topologies. 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 organism type. Cluster 1 is made up only of 4 TMS proteins, with the majority being from Firmicutes. The two exceptions are the most distant members of the cluster, 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,* and Tsp3 from *Thermococcus* sp. AM4. The proteins in this cluster are all from Euryarchaeota except for Dge1. Surprisingly, they were found to have an inverted order of their two 3 TMS segments relative to the standard 6 TMS 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 comprises 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, predominantly from Firmicutes, although five other phyla are

Table 1 Proteins of the SdpI family included in this study, listed alphabetically according to cluster

Abbreviation	GenBank no.	Organism source	Protein size (amino acyl residues)	No. of TMSs	Organism 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	257792477	*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

Table 1 continued

Abbreviation	GenBank no.	Organism source	Protein size (amino acyl residues)	No. of TMSs	Organism group
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
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	α-Proteobacteria
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	γ-Proteobacteria
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	α-Proteobacteria
Mth1	83590912	*Moorella thermoacetica* ATCC 39073	223	6	Firmicutes
Oal1	83859055	*Oceanicaulis alexandrii* HTCC2633	228	6	α-Proteobacteria
Pth1	98659796	*Pelotomaculum thermopropionicum* SI	229	6	Firmicutes
Rbi1	88804820	*Robiginitalea biformata* HTCC2501	216	6	Bacteroidetes

Table 1 continued

Abbreviation	GenBank no.	Organism source	Protein size (amino acyl residues)	No. of TMSs	Organism group
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

TMS transmembrane α-helical segments

Fig. 1 Phylogenetic tree of the SdpI family. Phylogenetic clusters are labeled 1–10. The tree is based on the ClustalX multiple alignment shown in Fig. S1 (http://www.biology.ucsd.edu/~msaier/supmat/SdpI/mat/FigureS1.html) and drawn with the TreeView program. Protein abbreviations are listed in Table 1
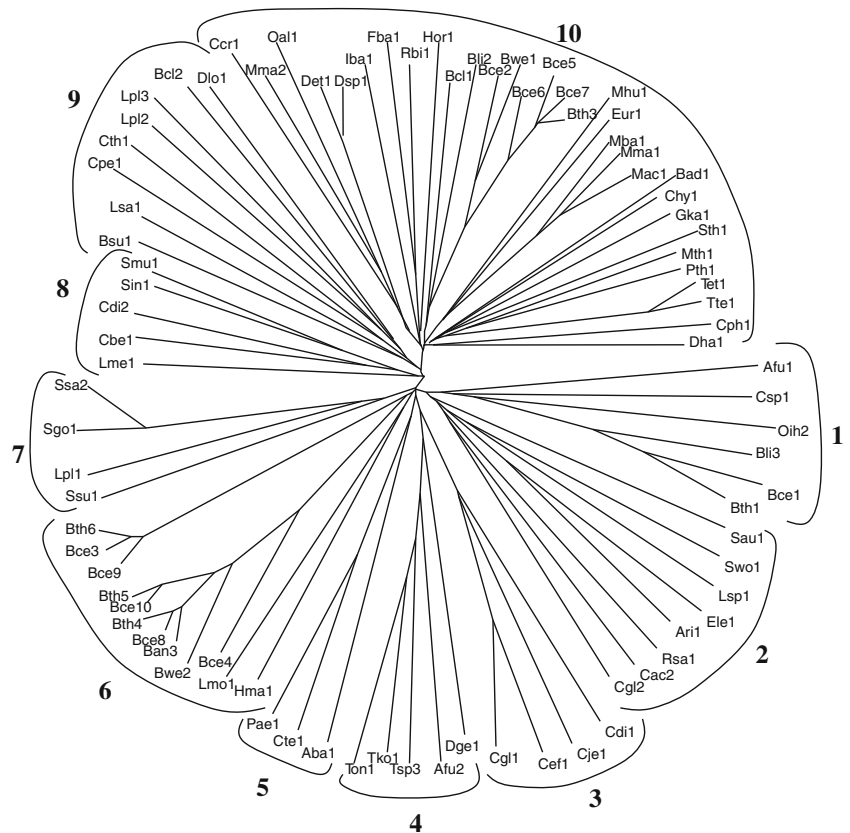
represented (Table 1). It is interesting to note that most of the 6 TMS proteins cluster loosely together (clusters 8–10), while proteins of other topologies are phylogenetically more distant.

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 because both pathways gave comparable results below the threshold comparison score needed for proof of homology, 10 SD (Saier 1994; Saier et al. 2009).

Several lines of indirect evidence support the suggestion that a 3 TMS primordial precursor duplicated internally to give rise to the standard 6 TMS topologies, as discussed below.

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 segments 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 (Fig. S2).
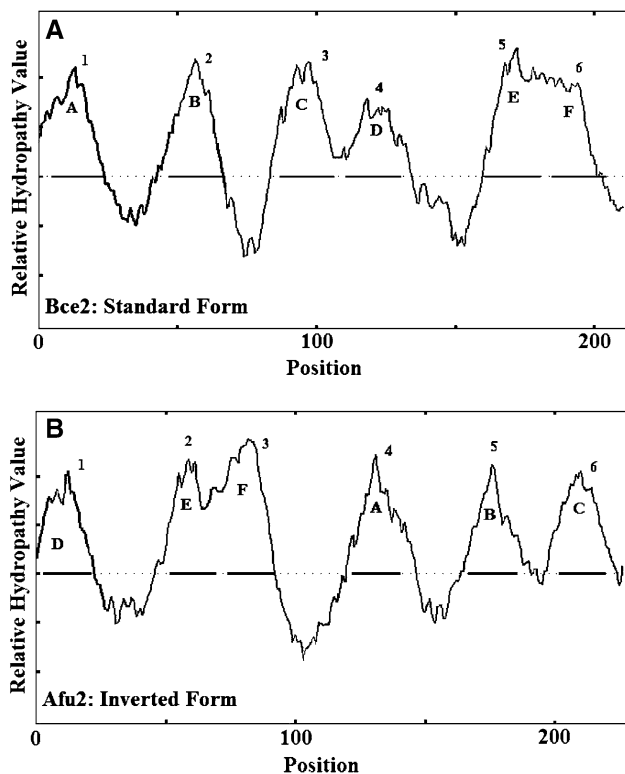
Fig. 2 a Hydropathy plot of the SdpI protein from *Bacillus cereus* (Bce2) with numbered peaks of hydropathy corresponding to putative TMSs. **b** Hydropathy plot of the SdpI homologue from *Archaeoglobus fulgidus* (Afu2) with numbered TMSs. The letters correspond to the homologous TMSs between the two proteins, demonstrating the inversion within Afu2 relative to the standard 6 TMS proteins, represented here by Bce2

The seventh peak of the inverted 7 TMS proteins showed statistically significant similarity to peak 10 of the Dge1 12 TMS protein (Fig. S3). It seems likely that the inverted 6 TMS protein arose from the inverted 7 TMS proteins with a one-TMS deletion event at the C-terminus (see "Discussion" and Figs. 8 and 9).

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 2 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 and Saier 2001a), the first half of Afu2 resembles the second half of Bce2, and the first half of Bce2 resembles the second half of Afu2. Figure 3a shows the GAP analysis between TMSs 1–3 of Afu2 and TMSs 4–6 of Bce2, with a comparison score of 16.6 SD. Figure 3b shows the GAP analysis between TMSs 4–6 of Afu2 and TMSs 1–3 of Bce2, with a comparison score of 15.5 SD. These values are substantially in excess of what is required to establish homology (Saier 1994; Saier et al. 2009).
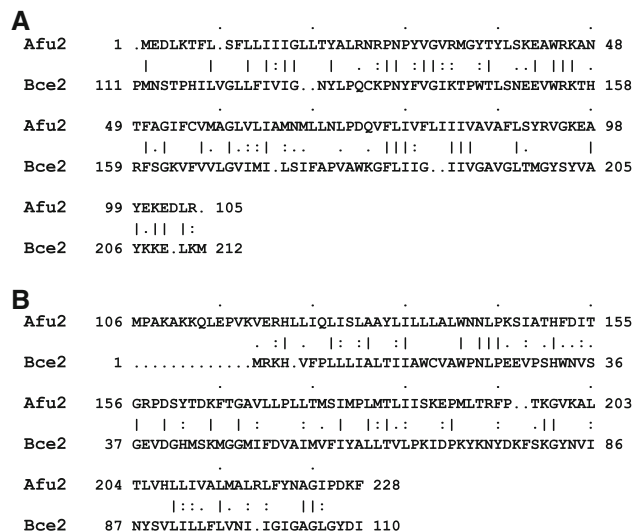

Fig. 3 a GAP comparison of the first 3 TMS segment of Afu2 (residues 1 to 105) with the second 3 TMS segment of Bce2 (residues 111 to 212) using the GAP program. Quality: 102; gaps: 5; percentage similarity: 44.4; percentage identity: 33.3. The average comparison score was 16.6 SD. **b** GAP comparison of the second 3 TMS segment of Afu2 (residues 106 to 228) with the first 3 TMS segment of Bce2 (residues 1 to 110) using the GAP program. Quality: 87; length: 125; gaps: 3; percent similarity: 38.9; percent identity: 21.3. The average comparison score was 15.5 SD

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 (Fig. S4). The 4 TMS proteins align with each other and correspond to TMSs 1–4 in the 6 TMS proteins. Figure 4 presents a GAP analysis of the 4 TMS Hma1 homologue with the 6 TMS Gka1 protein; it reveals the aforementioned alignment with a comparison score of 15.3 SD. Proteins with 4 TMSs are found predominantly in cluster 1, the three exceptions being Pae1 from
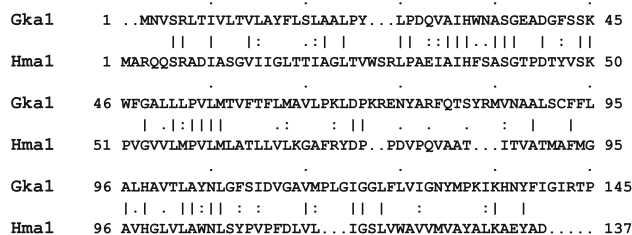

Fig. 4 GAP alignment demonstrating the regions of homology between the 6 and 4 TMS topological types within the SdpI family. Gka1 (residues 1 to 145 of 214), a 6 TMS representative, is compared with Hma1 (from residues 1 to 137), a 4 TMS representative. Quality: 106; gaps: 4; percentage similarity: 40.9; percentage identity: 29.5. The average comparison score was 15.3 SD

Prosthecochloris aestuarii, found in cluster 5, Sau1 from *Staphylococcus aureus,* located in cluster 2, and Hma1 from *Haloarcula marismortui,* located in cluster 6. Hma1 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 any of the standard 6 TMS proteins. The first peak of Cte1 aligns with the third peak of Afu2 (Fig. S5), which corresponds to peak 6 of the standard 6 TMS proteins, which suggests that this unique 5 TMS topology arose from the inverted 7 TMS protein through a 2 TMS deletion event at the N-terminus. Pae1 and Cte1 are found in cluster 5 along with Aba1. Aba1 is the longest 6 TMS protein with 303 aas. Only the first 210 aas code for the membrane-integrated portion of the protein. The remainder of the protein did not show homology with any region of any of the other proteins examined in this study. A BLAST search of this tail region yielded only hypothetical proteins, so no function of the region can be inferred.

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 (Fig. S6). The first peak of Cte1 aligns with peak 6 of the standard 6 TMS homologues. Bcl2, with a differing 5 TMS topology, has peaks 1–5 aligning with peaks 2–6 of the standard 6 TMS proteins (Fig. S7). 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 (Fig. S8). 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 standard 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 (Fig. S9). Their fifth peak corresponds best to the eighth 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 SD between proteins Rsa1 from *Renibacterium*

salmoninarum, a 5 TMS protein, and Lsp1 from *Lysinibacillus sphaericus,* an 8 TMS homologue (Fig. S10).

The 8 TMS homologues, although 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 (Fig. S11). The eighth peak of the 8 TMS homologues and the fifth peak of Rsa1 and Cgl2 are designated "A" and do not match any of the TMSs within other members of the family. The 8 TMS homologues align with the inverted 7 TMS proteins, with peaks 1–7 of the inverted 7 TMS proteins aligning with peaks 1–7 of the 8 TMS proteins (Fig. S12). Thus, the first three TMSs of the 8 TMS homologues correspond to peaks 4–6 of the standard 6 TMS proteins. 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 8 TMS proteins arose by deletion of three TMSs at the N-terminus of a 12 TMS Dge1-like protein and deletion of either one or two TMSs at its C-terminus. If two TMSs were deleted at the C-terminus of a Dge1-like protein, then a subsequent fusion event of one TMS must have occurred at the C-terminus of the 8 TMS homologues which corresponds to the eighth peak of the 8 TMS proteins. If, on the other hand, one TMS was deleted at the C-terminus of a 12 TMS protein, then the eleventh peak of the 12 TMS protein, corresponding to the eighth peak of the 8 TMS homologues, diverged in sequence so much that statistically significant similarity could not be found between the two. The 5 TMS topology most likely arose from an 8 TMS protein precursor by deletion of three TMSs at the N-terminus of the 8 TMS protein. It is possible that the 4 TMS topology may have arisen several times within the SdpI family, and that Sau1, as it clusters most closely with the 8 and 5 TMS homologues, may have arisen from a 5 TMS protein precursor by a deletion event at the C-terminus.

There are two variations of the 7 TMS topology. The first is an inverted topology as already discussed. The second is observed in Dlo1 with TMSs 1–6 aligning with TMSs 1–6 of the standard 6 TMS proteins (Fig. S13). The seventh peak of Dlo1 does not align with any other peak within the SdpI family and is designated "B." This protein is found in cluster 9 with 6 TMS proteins and Bcl2 of 5 TMSs. This clustering leads to the possibility that Dlo1 originated from a 6 TMS protein by addition of a C-terminal TMS, but it may equally well have arisen from a larger precursor derived from a 12 TMS protein by deletion of 5 TMSs at the C-terminus.

An Internal Duplication Within Dge1

Dge1, a 12 TMS protein, 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

A

```
Dge1    10 FLLGVVLTLALLGLAWGRVPAQEWLALLPLSVSSLLLGGLLAWLGRLEVE 59
               |||  .: || .  | |  :|  :     |   . |  :| .  |:    |
Dha1    60 FLLPLI.TLGVYILFW.IIPRID.....PQKANYLKMGRVF.WI....VS 97

Dge1    60 QRPVADAAAQALVLQAAVAAAAFAFAWSLPRALCVGTGLALVVMGNATSR 109
               ||  .    .:    | |  :  .|||   | |:  :.:||    :
Dha1    98 TTLVAFLS....LMYWGTIAVALGYLETLPRWYFSGIGIIFILLGNYFGK 143

Dge1   110 ARPGLWFGFRTRWALLSERAWYATQRQAAPALVSTGAVFTVFAALTPAPV 159
               :    || || | | .| |  | |||| |  :  |:    | . ||
Dha1   144 IKFNYTFGIRTPWTLANEEVWAKTHRFAGPIWI.VGGILMALAGVLPA.. 190

Dge1   160 LIPWVLPV.GL.LVLLAPVGISLHRASYR 186
               |  |.  |:  :||:| |  ..     ||
Dha1   191 ..AWTEPLFGIVIVLIAVVPMAYSYLVYR 217
```

B

```
Dha1     6 TSTHGRLLKVIAGILVIINIIIGIWAYPRLPEQVPSHWNLAGQVDGYSGA 55
               ||    ||  .. | | ::.:   :   | ||||||| |..|||. | |
Dge1   209 TSVERLLLALMLG.LPLLSLAACVVVLPWLPEQVPVHFDLAGRPDRYGSP 257

Dha1    56 LTGAFLLPLITLGVYILFWIIPRIDPQKANYLKMGRVFWIVSTTLVAFLS 105
               |    |||:||.   | .|            :   .   .. ||  |
Dge1   258 LE.LLALPLVGLGLAGFFAAMMRFGSATPAQRHLLLLTGALAGALTAPLP 306

Dha1   106 LMYWGTIAVALGYLETLPRWYFSGIGIIFILLGNYFGKIKFNYTFGIRTP 155
               |  | ...||    |      :  : :|    || :       |: |
Dge1   307 LGVSGDMSLPLGLGHVL....MLAVLALALLFPGPDGKRRPRLAAGLATL 352

Dha1   156 WTLANEEVWAKTHRFAGPI...WIVGGILMALAGVLPAAWTEPLFG.... 198
               |   .    . | |:   .:| | |: | .|     :| |
Dge1   353 AALLLPTLCLLPDQAAQPVGILFLVFGGLLFLVPMLLYGVPQPTAGRSKR 402
```

Fig. 5 a GAP comparison of Dge1 (*top, first half,* residues 10 to 186), a 12 TMS protein obtained with Dha1 (*bottom,* residues 60 to 217), a 6 TMS protein obtained with the GAP program. Quality: 114; gaps: 10; percentage similarity: 42.3; percentage identity: 31.4. The average comparison score was 15.4 SD. b GAP comparison of Dha1 (residues 6 to 198), a 6 TMS protein, with Dge1 (*bottom, second half,* residues 209 to 402), a 12 TMS protein, using the GAP program. Quality: 94; gaps: 4; percentage similarity: 34.8; percentage identity: 26.7. The average comparison score was 14.6 SD
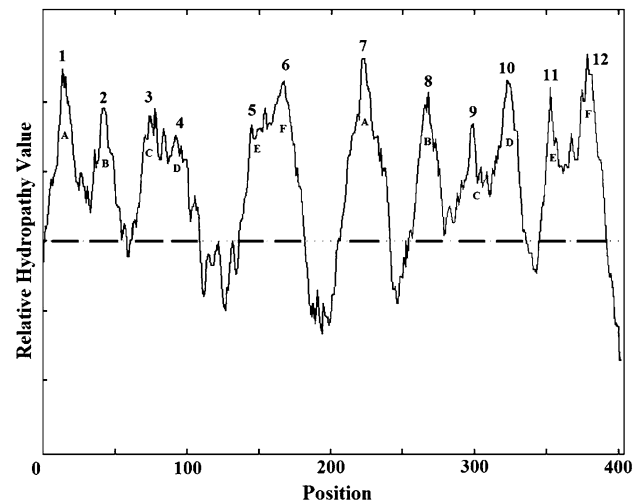

Fig. 6 Hydropathy plot of the SdpI homologue from *Deinococcus geothermalis* (Dge1) with numbered TMSs. Letters correspond to the homologous TMSs within the protein that probably arose by intragenic duplication. The plot was generated using the WHAT program

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 SD, generated by the comparison of the first half of Dge1 with Bcl1, with TMSs 4–6 of Dge1 corresponding to TMSs 4–6 of Bcl1 (data not shown). The 6 TMS protein, Dha1, aligned with both halves of Dge1. Alignment with the first half of Dge1 gave a comparison score of 14.6 SD (Fig. 5a), while alignment of the second half of Dge1 with Dha1 gave a comparison score of 15.4 SD (Fig. 5b).

The duplication event that led to the appearance of the 12 TMS Dge1 was evidently followed by extensive sequence divergence within both halves of this protein. 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 SD

(Fig. S14). The appearance of the hydropathy plot (WHAT program) for Dge1 also supports the conclusion of an internal duplication (Fig. 6). The evidence supports the proposal that the 6 TMS proteins represent the basic element for the SdpI family from which other family members evolved. These observations suggest that the standard 6 TMS protein may have duplicated internally to give 12 TMS proteins, that 12 TMS proteins may have led to the formation of the 8 TMS homologues, and that further deletions led to the inverted 7 and 6 TMS proteins.

Topological Comparisons

Figure 7 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 Fig. S15. Alignment of the proteins is shown according to their topologies (Fig. 7) as summarized in Fig. 8. Proteins of the 6 TMS topology, with the exception of the four inverted proteins, all align with TMSs 1–6 of all the others. The 4 TMS proteins align with each other as well as with TMSs 1–4 of the standard 6 TMS proteins. The 3 TMS proteins also align with each other and with TMSs 4–6 of the standard 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 standard 6 TMS proteins, and the first TMS of Cte1 corresponds to the sixth TMS of the standard 6 TMS proteins. In Bcl2, TMSs 1–5 align with TMSs 2–6 of the standard 6 TMS proteins. TMSs 1–5 of Sgo1 and Ssa2 align with each other and with TMSs 1–5 of
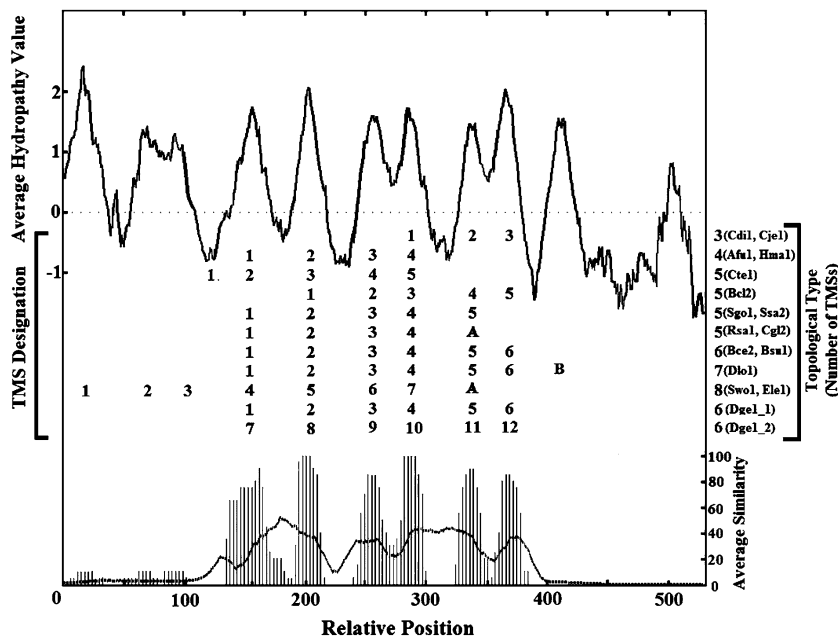
Fig. 7 Average hydropathy (*top*) and similarity (*bottom*) plots for the SdpI family excluding the four inverted proteins Afu2, Tsp3, Ton1, and Tko1, and with the 12 TMS protein, Dge1, spliced into two 6 TMS long halves. These plots were generated using the AveHAS program based on the ClustalX multiple alignment shown in Fig. S2 on our Web site. 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 or B) if not conserved among the groups of proteins. At *right*, the total numbers of putative TMSs of each topological type are presented together with representative examples. All TMSs in a single vertical column are homologous regardless of the number designations used except for TMSs indicated by letters. TMSs A are not demonstrably homologous to TMS B. *Note:* the first peak of Cte1 marks the region where the first peak of Cte1 aligned, and because it is the only representative within the SdpI family to have this region, it is poorly displayed in the AveHAS plot

Fig. 8 Topological types of proteins of the SdpI family analyzed in this work. The *left column* indicates the number of TMSs in each topological type of proteins analyzed as well as representative proteins. The *center 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 demonstrably homologous to each other unless they are designated by letter. The number of each TMS is assigned by its corresponding TMS of homology within the standard 6 TMS proteins. 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 (Fig. 1) in which proteins of the topological type of the same row are found. i denotes inside the cell; o denotes outside the cell

the standard 6 TMS proteins. Rsa1 and Cgl2 align with each other, and their TMSs 1–4 align with TMSs 1–4 of the standard 6 TMS proteins. TMSs 1–6 of Dlo1 (7 TMS topology) align with TMSs 1–6 of the standard 6 TMS proteins. TMSs 1–3 of the 8 TMS proteins align with TMSs 4–6 of the standard 6 TMS proteins, and TMSs 4–7 of the 8 TMS proteins align with TMSs 1–4 of the standard 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 and Elkan 1995). The best conserved motif, motif 1 (AL[YW]PXLP[ED]R[VI][PA][VI]H[WF][NG]ASGE[VP][DN][GR][YF][GM]SKF[EV][GL]) (alternative residues at a single position are in brackets; X = any residue), is found in most members of the family that include TMSs 1 and 2. On the basis of results obtained with the MEME and WHAT programs, motif 1 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 1. 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. Therefore, motif 1 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 1.

The second best conserved motif, motif 2 ([IV]G[LI]L[FL]I[VG][LI]GNY[LM][PG]KX[KR]PN[YW]F[VI]GIRT-PWTLS[SN][ED]EVW[RN]KT[HN]R[LF][GA]G[KR][LV][FW]V[IAV][GA]G), 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 2 is found between the first and second TMSs as expected because these proteins align with TMSs 4–6 of the standard 6 TMS proteins. Figures 8 and S16 depict the locations of the recognized motif 2 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 2. 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 majority of the proteins with the standard 6 TMS topology have one of three combinations of these two motifs. The 6 TMS proteins from clusters 8 and 10 contain both motifs. The four "inverted" proteins of 6 and 7 TMSs were also found to contain the same combination of motifs albeit in an inverted manner.

Of the 6 TMS proteins of clusters 5, 6, and 7, MEME recognized only motif 1 with the exception of Lpl1 of cluster 7, which displays both motifs. Finally, cluster 9 contains 6 TMS proteins in which only motif 2 was recognized by MEME except for the aforementioned 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 2 (A and B) as are the sequence similarities between the consensus motifs 1 and 2 (Table 2C).

Operon Analyses

The genomes encoding 17 SdpI homologues were examined, those whose genomes were found within the NMPDR Genome Browser database. Analysis of the genomic environments of *sdpI* genes was limited by the data currently available in the genomes of the species examined. Of the 17 genomes encoding SdpI homologues, only 6 contained *sdpI* genes encoded within multicistronic operons. The remaining 11 *sdpI* genes could not be conclusively linked to other genes. All six operons included a transcriptional regulator upstream of the SdpI homologue gene. Both Bce2 from *Bacillus cereus* ATCC 10987 and Bsu1 from *Bacillus subtilis* subsp. *subtilis* str. 168 encoded ArsR family transcriptional regulators, 3 bp upstream of the *sdpI* genes. Lpl1 from *Lactobacillus plantarum* WCSF1 had an unidentified transcriptional regulator encoded immediately upstream of it. The gene for Afu2 in *Archaeoglobus fulgidus* DSM 4304 overlapped with a PadR family transcriptional regulator gene by 8 bp; Swo1 from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen overlapped by 4 bp with the gene for a GntR family transcriptional regulator, and Tko1 from *Thermococcus kodakarensis* KOD1 had a ParR family transcriptional regulator encoded immediately upstream of it. Uniquely, Tko1 additionally had an unidentified 357-bp gene coding for a hypothetical membrane protein 3 bp downstream of the *sdpI* gene, one 723-bp gene coding for a hypothetical protein 7 bp upstream of the transcriptional regulator, and a 342-bp gene coding for a hypothetical protein 3 bp upstream of the 723-bp gene. The hypothetical proteins could not be identified using BLAST searches and were not homologous to Tko1 or to any of the transcriptional regulators. PadR and ParR were found to be

Table 2 Summary of the similarities and differences within and between the sequences of motifs 1 (A) and 2 (B) among clusters; (C) shows an alignment of the consensus motif 1 (M1) with the consensus motif 2 (M2)

A   Motif 1

```
Cluster
1        I  I [YW] G  X  L  P E  E  V  P A H  Y  N  A X G  E  V  D  R W  G  S  K W E L
            :     .        :    ^      :          ^        ^      .   :    :
2     [VL|H] Y  P  X [ML] P D  X  V  P V H W  N [GF] S G  E [AV] D  X  W  X  X  K A D K
                               :                        :
4        L  L  W  D  K  L  P E  X  V  A X H  F [ND] A  S G  E  P  D  X  Y  X  S  K F A G
            :     .        ^    :  ^    :      .        .    .      .
5        Y  H  Y  X  X  L  P X  X  I  A X H  F  G  F  S G  X  P  D  A W  G  P  K G P K
         ^.               ^        ^  ^ ^            ^.            ^
6        C  L  Y  P  Y  L  P N  R  I  A V H W [ND] F  N G  X  P  N  E  F  M  S  K Q V V
         :   :              ^        :                ^                :
7        W  V  Y  X  X  L  P X  N [LM|A I  H  F  D  L  S G  K [GP] N [AG] F  L  P  K L P K
         ^             ^    ^        .  :  .      ^.        .          .
8        X  L  W  X [KQ] L  P E  E [VI] A I  H  F  D  X  S G  X  P  D [GN] Y [AS] [PS]K V X A
         :               ^    .        ^  ^  :        ^            :          :
10       A  A  Y  P  H  L  P D  T  I  P T H W  N  S  S G  E  V  D  G  Y  G  S  S K F E
                               .  :                  ^                  :
```

B   Motif 2

```
Cluster
3     [VL|L  L L  V  X G X L [GA]W X  A X L  P  G N P V V G I R V  P E V R  K  S  Q E L W D M A H R V A G P L W V L S V
         :     ..     ^  ^    .      :.  ^ ^   :.  :   ^ ^ ^ ^    ^   .    :    ^.  ^
4     L [TL|L L  A  A G L L  T  X  A  H X N  R  R N X X I G F R X  G Y T Y  X  S  D R A W R K T  N T F A G I  F X V A X S
                          ^         .        ^:  ^   ^              .    .    :.          :        .
7     V  A  L V  F [IM|I I G  N  Y  P  K X V  Q  X N R X X M N R K  L A X L  F  V  G I G  F L X  F I V A I  F Y L
         :  ^.       :: :   :^  ^ ^ ^               :.:            .  :: : :      :: :     :      :.  ^ :
8     L  G I [ML|[FLI|IV I G  N  Y [LM] P K X [KR] Q N Y T V G I R L  P W T L [ND][ND]E X N W N K T H R L A G K L W V L G G
         : .^       . . : .      ^ ^       ^      .  ^ :^  :  .  .    : ^  ::   :.    :.  ^^
9     L  G I M  X  I V I G  N  I  X  P K X  K  Q N R X I G I R T  P W T L  S  S  E E V W K K T N X L X G K L L V I V G
         : :         ^ : .        .  ^       .  :   .:  .   :  :                 ::             :   :.:
10    [IV|G L L  F  I V L G  N  Y  L  P K C  K  P N Y F V G I R T  P W T L  S  S  E E V W R K T H R F G G K V  F V I L G
                  .                 ^                :  .   .      :  :    :            ::         ::
```

C

```
M2   1  VLLILIGLLFIVLGNY..LPKAKPNYFVGIRTPWTLSSE.EVWRKTHRLGGKLFVIGGILML  59
        .| :|:|  |  | : :   |  ||.  |  :       |  |  |  : :    |  | |    |:  :|
MI   4  LLTLLLGPLILTIALYPALPERVPVH.......WNASGEVDGYGSKFEGGSKL...EGVFLL  55
```

(A,B) * Identity; :, close similarity; ., more distant similarity within the cluster as defined by the ClustalX program. *Note:* Motif 1 was only found in one protein (Lpl1) in cluster 7 by the MEME program, while the ClustalX program showed conservation of motif 1 in all members of cluster 7. The consensus sequence for motif 1 in cluster 7 is based on all members of the cluster. Three fully or nearly fully conserved residues were found in each motif. These are boxed with a circled asterisk at the top of the alignment. Other less well-conserved residues are also boxed but not indicatd by a circled asterisk. (C) |, identity; :, close similarity; ., more distant similarity as defined using the GAP program. The residues indicated in motifs 1 and 2 are the dominant residues at the various aligned positions

homologous to one another, although PadR, ParR, and GntR did not show obvious homology with SdpR. ArsR was found to be homologous to SdpR. We tentatively suggest that these putative transcriptional regulators bind on the cytoplasmic sides of the membrane to the SdpI homologues to effect autoregulation.

Promoter analyses (see Methods) predicted promoters ending 18 bp upstream of the gene encoding the Bce2 transcriptional regulator with a score of 0.98, 49 bp upstream of the Bsu1 regulatory gene with a score of 1.00, 7 bp upstream of the Lpl1 regulatory gene with a score of 0.84, 32 bp upstream of Afu2 s with a score of 0.94, 38 bp upstream of Swo1 s with a score of 0.99, and 33 bp upstream of Tko1's transcriptional regulator with a score of 0.94; no promoters were found within 120 bp of the first hypothetical protein in the Tko1 operon.

Discussion

### Proposed Pathway for the Evolution of Varying Topologies

Figure 9 diagrams the proposed pathways for the evolution of proteins of the SdpI family and shows their differing topologies. 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 and 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 postulate that in the SdpI family, the primary 6 TMS proteins may have arisen through intragenic duplication of a primordial 3 TMS-encoding DNA segment. Deletions within this basic element over evolutionary time led to the formation of the 4 TMS, the 3 TMS, and two of the 5 TMS variant proteins. A fusion event may have led to the appearance of the noninverted 7 TMS protein (Dlo1). The 12 TMS protein undoubtedly arose by intragenic duplication of the basic 6 TMS element followed by extensive sequence divergence of both halves, particularly in the first and last 3 TMS segments. Deletion events of a primordial 12 TMS protein led to the formation of the 8 TMS proteins. Deletions in the 8 TMS homologues led to the formation of the inverted 7 TMS proteins, Tko1, Tsp3, and Ton1, and one of the 5 TMS topological types, represented by Rsa1 and Cgl2. Deletions in the inverted 7 TMS proteins led to the formation of the inverted 6 TMS protein, Afu2, and the 5 TMS variant, Cte1. The 4 TMS proteins may have also arisen by deletion of one TMS from the N-termini of the 5 TMS proteins, Rsa1 and Cgl2. It is probable that the 4 TMS topology arose at least twice as some of the 4 TMS proteins cluster closer to the standard 6 TMS proteins, while other 4 TMS proteins cluster closer to 5 and 8 TMS proteins.

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 (Fig. 8) in such a way that motif 1, between TMSs 1 and 2 in the standard 6 TMS proteins, is always found to be externally localized, while motif 2, between TMSs 4 and 5 in these same proteins, is always located on the inside, facing the cytoplasm. 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 internally 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 localize to 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 Fig. 8) were predicted to be localized to the inside of the cell, while the C-termini were predicted to be on the outside. On the basis of all of these predicted orientations, which were in surprising agreement with each other, motif 1 is always on the external surface to the membrane, while motif 2 always faces the cytoplasm. Because we postulate that motif 1 is responsible for neutralizing the extracellular SdpC toxin by forming an SdpI–SdpC complex in the membrane, motif 1 should be localized to the outer surface of the cellular membrane. By contrast, because motif 2 is predicted to be responsible for promoting expression of the *sdpRI* operon by sequestering the cytoplasmic autorepressor, SdpR, it would follow that this process occurs on the inside of the membrane. The predicted topologies therefore fully support the functional predictions.

Fig. 9 Proposed pathway for the evolution of the proteins of differing topologies within the SdpI family. *Black arrows* indicate probable direction of evolution; *striped arrows* indicate possible evolutionary pathways when two different pathways are equally probable
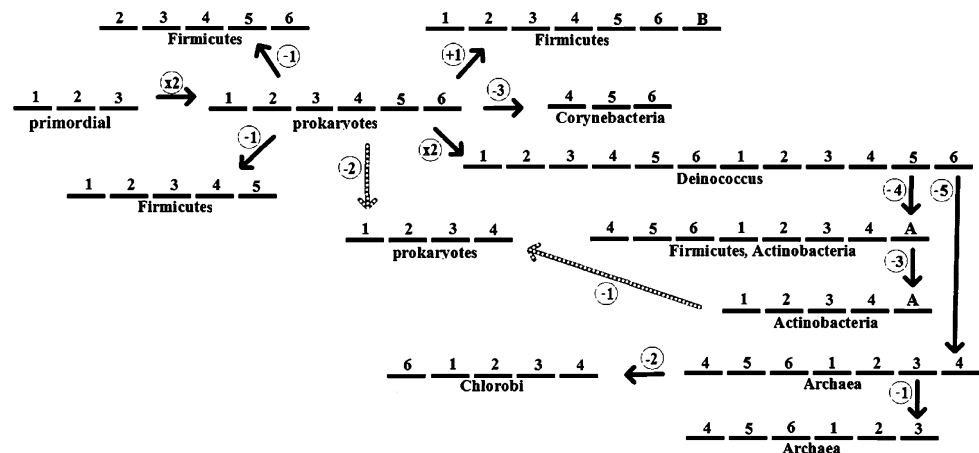
Conserved Motifs Confirm Homology of SdpI Family Members

Analyses of the motifs present in the proteins of the SdpI family confirmed homology of most family members despite variations in their topologies. Figure 8 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 1 and 2 of the standard 6 TMS proteins, while motif 2, when present, is always found between TMSs 4 and 5 in these same homologues. 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 the AveHAS program (Fig. 7).

Motif analyses of the four inverted proteins confirmed the proposed inversion. Motif 2, 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 1, found in the region between TMSs 4 and 5 in the inverted proteins, 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 inverted topology of the former proteins with respect to the standard 6 TMS proteins 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 on the basis of our motif analyses. Cluster 6 contains 6 TMS proteins which only exhibit motif 1, and Hma1 also contains only motif 1. 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 1, and Pae1 is very closely related to Cte1, leading to the possibility that the 4 TMS protein, Pae1, originated from a 5 TMS protein like Cte1 by deletion of one TMS at the N-terminus.

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. Alternatively, they may have 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 1 in the other 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 no. 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 1–2 or motif 2 between TMSs 4–5 may alone be adequate for one of the two 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 may form heterodimers to ensure a complex possessing both of the conserved motifs.

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 topological variation within members of the SdpI family. The first 3 TMSs of the *B. subtilis* 6 TMS SdpI protein are likely to be 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 likely to be responsible for one of the functions. Proteins with 3, 4, and 5 TMSs may be unifunctional 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. Because 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. In the case of the 8 TMS proteins, only Swo1 displays both conserved motifs; the remainder display only motif 1. Thus, the 8 TMS proteins may have started out with both functions, but they have since diverged to become unifunctional. Alternatively, motif 2 may have diverged to provide for a distinct but related function (e.g., binding of a protein dissimilar to SdpR).

Two strains of *Corynebacterium glutamicum* and one of *C. efficiens* both have two SdpI homologues, a 3 TMS protein (e.g., Cgl1 in Table 1) and a 5 TMS protein (e.g., Cgl2 in Table 1; unpublished observations). The 3 TMS

proteins are 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 proteins are 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. By having two truncated proteins with complementary functions, possibly in complex with each other, regulated SdpC immunity could exist, involving two related but dissimilar proteins.

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 proteins. (2) Assuming that these conserved motifs bind SdpC (the toxin) and SdpR (the regulator), respectively, then SdpC would bind to the external surface of the membrane while SdpR would bind to the cytoplasmic side, as expected, on the basis of the mutational analyses (Ellermeier et al. 2006; 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, respectively, as suggested by Ellermeier et al. (2006), could be explained if the two sequence divergent halves of a 6 TMS SdpI protein arose from a 3 TMS protein binding precursor polypeptide. (5) The fact that several SdpI homologues exhibit an inverted topology relative to the standard 6 TMS proteins makes functional sense because the order of two 3 TMS halves in the polypeptide chain should be of no functional significance. (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, are 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 the second and third 3 TMS elements of the four 3 TMS elements have retained function (Fig. 8). Loss of the nonfunctional regions (TMSs 1–3 and 10–12) would yield the inverted homologues.

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, intragenic duplication of a 3 TMS–encoding genetic element. Further work, including the generation of high-resolution three-dimensional structural data, is likely to provide confirmation or refutation of this proposal.

Operon Analyses

Operon analyses of a few SdpI homologues suggested that a significant fraction of SdpI homologues are encoded in operons downstream of transcriptional regulators. The transcriptional regulator, ArsR from *Bacillus cereus* E33L, is homologous to SdpR, SdpI's autorepressor. The transcriptional regulators, ParR from *Thermococcus kodakarensis* KOD1, PadR from *Archaeoglobus fulgidus* DSM 4304, and GntR from *Syntrophomonas wolfei* subsp. *wolfei* str. Goettingen, could not be shown to be homologous to SdpR. Unique among the operons examined, Tko1 is found downstream of its transcriptional regulator, upstream of one hypothetical protein and flanked upstream of its transcriptional regulator by two hypothetical proteins in a single operon. It seems reasonable to suggest that these transcriptional regulators function in conjunction with the SdpI homologues to effect autoregulation in response to extracellular signals. The fact that several of these regulators are nonhomologous provides evolutionary rationalization for the sequence divergence of the companion SdpI homologues. Future identification of the hypothetical proteins will undoubtedly provide valuable clues as to the functions of Tko1 and the other SdpI homologues.

Alternative Potential Functions of SdpI Homologues

The basis for the presence of multiple SdpI paralogues encoded within single genomes is poorly understood from a functional standpoint. However, bacteria possess complex gene networks and process protein information to influence many cellular behavioral traits (Schultz et al. 2009). These gene circuits and functional molecules allow integration of complex signals impinging on a network of modules. The availability of multiple processed input signals, transmitted to a central integrator, allows fine-tuning of the decision-making process. This would be particularly advantageous when numerous developmental alternatives exist (Chagneau and Saier 2004; Schultz et al. 2009; Stephenson and Hoch 2002).

The need for functional integration of multiple input signals, sensed by cell surface sensors, is emphasized by the developmental complexity of *Bacillus subtilis*. In addition to vegetative growth, this organism can (1) sporulate and form fruiting bodies, (2) develop competence for DNA uptake, (3) become supermotile (swarming competent), and (4) form complex organized sessile communal lifestyles (biofilms) with different functions delegated to

different cell populations (Branda et al. 2001; Chagneau and Saier 2004; Verhamme et al. 2009). Indeed, nonsporulating bacteria often retain several of these other possibilities.

The developmental fate of any one cell within a population is dependent on environmental conditions, internal signals, collective and individual sensing, and a nongenetic, stochastic (nondeterministic) process responding to quantitative chance events in a qualitative way (Ben-Jacob 2009; Losick and Desplan 2008). Stochastic decision making may be random, but it is likely to be influenced by dozens of input signals, each sensed by a different set of cytoplasmic and transmembrane sensors. This may explain the recurrence of multiple homologous receptors in a single organism possessing multiple programs. They provide immunity to or allow signal transmission in response to many other extracellular signals in addition to SdpC toxin. Among the bacteria possessing SdpI homologues are several present in the human intestinal tract (e.g., Lactobacilli, Clostridia, Bifidobacteria, etc.) The actions of these homologues may be to sense the presence of other bacterial cohabitants, including members of both the same and different species. Research on these proteins may be applicable to an understanding of their interactions and to their contributions to human health and disease. Targeting SdpI homologues could provide a new Achilles' heel against multidrug-resistant superbugs. One example is the common hospital bacterium, *Clostridium difficile*, which possesses an SdpI homologue. Targeting the agents mediating cannibalism could allow the development of novel antibiotics that under stress conditions would reduce specific bacterial populations *in vivo*. The need for the development of novel antibiotics with unique targets is emphasized by the emergence of multidrug resistance in several disease-causing microbes.

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 SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bailey TL, 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 AJ (2005) Where asymmetry in gene expression originates. Mol Microbiol 57:611–620

Ben-Jacob E (2009) Learning from bacteria about natural information processing. Ann N Y Acad Sci 1178:78–90

Branda SS, Gonzalez-Pastor JE, Ben-Yehuda S, Losick R, Kolter R (2001) Fruiting body formation by *Bacillus subtilis*. Proc Natl Acad Sci USA 98:11621–11626

Britton RA, Eichenberger P, Gonzalez-Pastor JE, Fawcett P, Monson R, Losick R, Grossman AD (2002) Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of *Bacillus subtilis*. J Bacteriol 184:4881–4890

Butcher BG, Helmann JD (2006) Identification of *Bacillus subtilis* sigma-dependent genes that provide intrinsic resistance to antimicrobial compounds produced by bacilli. Mol Microbiol 60:765–782

Chagneau C, Saier MH Jr (2004) Biofilm-defective mutants of *Bacillus subtilis*. J Mol Microbiol Biotechnol 8:177–188

Chung JD, Stephanopoulos G, Ireton K, Grossman AD (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 JP, Havarstein LS (2007) Cannibalism and fratricide: mechanisms and raisons d'etre. Nat Rev Microbiol 5:219–229

Dayhoff MO, Barker WC, Hunt LT (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 RF (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 ST, Conlon EM, Rudner DZ, Wang ST, Ferguson C, Haga K, Sato T, Liu JS, Losick R (2004) The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*. PLoS Biol 2:e328

Ellermeier CD, Hobbs EC, Gonzalez-Pastor JE, 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 JE, Hobbs EC, Losick R (2003) Cannibalism by sporulating bacteria. Science 301:510–513

Grossman AD (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 RA, Martin L, Saier MH 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 EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

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

Lee JH, Harvat EM, Stevens JM, Ferguson SJ, Saier MH 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

Losick R, Desplan C (2008) Stochasticity and cell fate. Science 320:65–68

McNeil LK, Reich C, Aziz RK, Bartels D, Cohoon M, Disz T, Edwards RA, Gerdes S, Hwang K, Kubal M, Margaryan GR, Meyer F, Mihalo W, Olsen GJ, Olson R, Osterman A, Paarmann D, Paczian T, Parrello B, Pusch GD, Rodionov DA, Shi X, Vassieva O, Vonstein V, Zagnitko O, Xia F, Zinner J, Overbeek R, Stevens R (2007) The National Microbial Pathogen Database Resource (NMPDR): a genomics platform based on subsystem annotation. Nucleic Acids Res 35:D347–D353

Molle V, Fujita M, Jensen ST, Eichenberger P, Gonzalez-Pastor JE, Liu JS, Losick R (2003) The Spo0A regulon of *Bacillus subtilis*. Mol Microbiol 50:1683–1701

Pao GM, Wu LF, Johnson KD, Hofte H, Chrispeels MJ, Sweet G, Sandal NN, Saier MH Jr (1991) Evolution of the MIP family of integral membrane transport proteins. Mol Microbiol 5:33–37

Parker GF, Daniel RA, Errington J (1996) Timing and genetic regulation of commitment to sporulation in *Bacillus subtilis*. Microbiology 142(pt 12):3445–3452

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

Saier MH Jr (2003) Tracing pathways of transport protein evolution. Mol Microbiol 48:1145–1156

Saier MH Jr, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. Nucleic Acids Res 37:D274–D278

Schultz D, Wolynes PG, Ben Jacob E, Onuchic JN (2009) Deciding fate in adverse times: sporulation and competence in *Bacillus subtilis*. Proc Natl Acad Sci USA 106:21027–21034

Smith RF, Wiese BA, Wojzynski MK, Davison DB, Worley KC (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

Stephenson K, Hoch JA (2002) Evolution of signalling in the sporulation phosphorelay. Mol Microbiol 46:297–304

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Tusnady GE, Simon I (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850

Verhamme DT, Murray EJ, Stanley-Wall NR (2009) DegU and Spo0A jointly control transcription of two loci required for complex colony development by *Bacillus subtilis*. J Bacteriol 191:100–108

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

Vreeland RH, Rosenzweig WD, Powers DW (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. Nature 407:897–900

Wang B, Dukarevich M, Sun EI, Yen MR, Saier MH Jr (2009) Membrane porters of ATP-binding cassette transport systems are polyphyletic. J Membr Biol 231:1–10

Yen MR, Choi J, Saier MH Jr (2009) Bioinformatic analyses of transmembrane transport: novel software for deducing protein phylogeny, topology, and evolution. J Mol Microbiol Biotechnol 17:163–176

Zhai Y, Saier MH Jr (2001a) 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 MH Jr (2001b) 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 MH 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 MH Jr (2002) A Web-based Tree View (TV) program for the visualization of phylogenetic trees. J Mol Microbiol Biotechnol 4:69–70