Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochimica et Biophysica Acta 1618 (2003) 79–92
BIOCHIMICA ET BIOPHYSICA ACTA
BBA
www.bba-direct.com

The ion transporter superfamily

Shraddha Prakash, Garret Cooper, Soumya Singhi, Milton H. Saier Jr.*

Division of Biological Sciences, University of California at San Diego, La Jolla, CA 92093-0116, USA

Received 21 August 2003; received in revised form 15 October 2003; accepted 17 October 2003

Abstract

We define a novel superfamily of secondary carriers specific for cationic and anionic compounds, which we have termed the ion transporter (IT) superfamily. Twelve recognized and functionally defined families constitute this superfamily. We provide statistical sequence analyses demonstrating that these families were in fact derived from a common ancestor. Further, we characterize the 12 families in terms of (1) the known substrates transported, (2) the modes of transport and energy coupling mechanisms used, (3) the family sizes (in numbers of sequenced protein members in the current NCBI database), (4) the organismal distributions of the members of each family, (5) the size ranges of the constituent proteins, (6) the predicted topologies of these proteins, and (7) the occurrence of non-homologous auxiliary proteins that may either facilitate or be required for transport. No member of the superfamily is known to function in a capacity other than transport. Proteins in several of the constituent families are shown to have arisen by tandem intragenic duplication events, but topological variation has resulted from a variety of dissimilar genetic fusion, splicing and insertional events. The evolutionary relationships between the members of each family are defined, leading to predictions of functionally relevant orthologous relationships. Some but not all of the families include functionally dissimilar paralogues that arose by early extragenic duplication events.

Keywords: Transporter; Carrier; Ion; Phylogeny; Classification; Superfamily

1. Introduction

For the past decade, our laboratory has been concerned with the characterization and classification of transmembrane transport systems [1]. For this purpose, we developed a systematic, functional/phylogenetic classification scheme that is theoretically applicable to every type of transporter found in living organisms on Earth [2,3]. Nearly 400 families of transporters are currently included within the Transporter Classification (TC) system [4]. This system of classification was adopted by the International Union of Biochemistry and Molecular Biology (IUBMB) as the internationally acclaimed system for classifying transport proteins.

A major task we have undertaken is to establish distant phylogenetic relationships between families, thus creating superfamilies. Among the largest superfamilies of secondary carriers we have characterized are (1) the Major Facilitator (MF; TC #2.A.1) [5,6], (2) the Amino Acid-Polyamine-Organocation (APC; TC #2.A.3) [7], (3) the Resistance-Nodulation Division (RND; TC #2.A.6) [8], (4) the Drug-Metabolite Transporter (DMT; TC #2.A.7) [9], (5) the Multidrug-Oligosaccharidyl-lipid-Polysaccharide (MOP; TC #2.A.66) [10] and (6) the L-Lysine Exporter (LysE, TC #2.A.75–77) [11] superfamilies.

In an earlier paper, we provided preliminary motif evidence for the existence of an additional superfamily, which we called the ion transporter (IT) superfamily [12]. The recognized substrates of all of the functionally characterized members of these families proved to be ionic, either cationic or anionic species, none being neutral, hence the name IT superfamily. Eleven previously recognized families were found to exhibit common matrices of motifs as demonstrated using the MEME and MAST programs [12,13]. However, interfamilial statistical analyses of the protein sequences were not conducted. Consequently, we could not establish the relatedness of these proteins.

Recently, Lolkema and Slotboom [14] used hydropathy profile analysis to provide evidence for evolutionary relationships between several of the constituent families of the IT superfamily. This method correctly identified the IT

* Corresponding author. Tel.: +1-858-534-4084; fax: +1-858-534-7108.
E-mail address: msaier@ucsd.edu (M.H. Saier).

0005-2736/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbamem.2003.10.010

superfamily families we describe here. Additional functionally uncharacterized protein families of putative transporters not included in the TC system were also identified. However, the method may have given false positives, including families, which we do not believe belong to the IT superfamily (see Discussion). No statistical data substantiating their claim of homology were provided.

In the present paper, we carry out detailed analyses of the sequences of proteins comprising the families of the IT superfamily. We show that the 11 previously recognized putative IT superfamily constituents and a 12th recently identified family, the *p*-aminobenzoyl-glutamate transporter (AbgT) family [15], belong to this superfamily. We provide concrete statistical data using the IntraCompare (IC) and GAP programs [16,17] demonstrating that these 12 families are, in fact, related. In addition, we provide detailed information about the characteristics of each of the 12 constituent families. Tables listing family members, multiple alignments of their sequences, phylogenetic trees and average hydropathy, amphipathicity and similarity plots can be found on our supplementary materials website (www-biology.ucsd.edu/~msaier/supmat/).

2. Computer methods

Protein sequences that comprise the 12 families within the IT superfamily were obtained by recursive PSI-BLAST searches without iterations until all potential members had been retrieved from the NCBI database (*e*-value $\leq 10^{-4}$) [18]. Homologues were retrieved September–December, 2002. Redundant sequences were eliminated using an unpublished program (S. Singhi and M.H. Saier, Jr., unpublished), and protein fragments were extended using a frame translation method found on the BCM search launcher website (http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html). The CLUSTAL X program [19] and the TREE program [20] were used for multiple alignment of homologous sequences and construction of phylogenetic trees with the aid of the BLOSUM30 scoring matrix and the TREEVIEW drawing program [21].

Family assignments were based upon the phylogenetic results and statistical analyses obtained with the GAP program [16]. Our standard for establishing homology between two proteins is 9 SDs, when large regions are compared using the GAP program with 500 random shuffles, a gap opening penalty of 8, and a gap extension penalty of 2, as outlined and rationalized previously [1]. Interfamilial sequence comparisons presented in Table 2 were obtained by first running the IC program [17], comparing all members of each family with the members of all other families. The sequences were then individually compared using the GAP program [16]. The BLAST2 program [22] and the BASS program [23] were also used for preliminary binary comparisons to identify well aligning sequences.

The TMS SPLIT program [23], along with the IC program [17], was used to analyze proteins for internal duplications. The TMS-ALIGN program [23] was also valuable for locating the homologous TMSs and identifying internal duplications (see Figs. S3A, 3B, 3C and 3D on our supmat website). The programs developed in our laboratory are available on our ''biotools'' website (http://www-biology, ucsd.edu/~yzhai/biotools.html).

The topologies of individual integral membrane proteins were estimated using the TMHMM [24], HMMTOP [25] and WHAT [26] programs. The last of these programs also provides predictions of amphipathicity and secondary structure. The AveHAS program [27] was used for plotting

Table 1
The 12 established families in the IT superfamily

Family TC number	Family name	Family abbreviation	Number of members	Source organisms[a]	Size range (number of residues)[b]	Number of putative TMSs[c]	Well-characterized example
2.A.68	*p*-Aminobenzoyl-glutamate Transporter	AbgT	20	B	496–561	12	AbgT Eco
2.A.45	Arsenite–Antimonite Efflux	ArsB	90	B, A, E	370–846	12 – 13 (12)	ArsB Eco
2.A.11	Citrate-Mg^{2+}:H^+ (CitM)-citrate:H^+ (CitH) Symporter	CitMHS	21	B	426–489	12	CitM Bsu
2.A.47	Divalent Anion:Na^+ Symporter	DASS	154	B, A, E	432–923	11–14 (11)	SoT1 Sol
2.A.13	C_4-Dicarboxylate Uptake	Dcu	26	B	432–474	12 (10)	DcuA Eco
2.A.61	C_4-Dicarboxylate Uptake C	DcuC	12	B	431–495	12	DcuC Eco
2.A.8	Gluconate:H^+ Symporter	GntP	75	B, A	422–498	12	GntP Bsu
2.A.14	Lactate Permease	LctP	52	B, A	466–623	14 – 15	LctP Eco
2.A.34	NhaB Na^+:H^+ Antiporter	NhaB	14	B	499–533	10–12 (9)	NhaB Val
2.A.35	NhaC Na^+:H^+ Antiporter	NhaC	65	B, A	419–607	11–12; 14 (14)	NhaC Bfi
2.A.62	NhaD Na^+:H^+ Antiporter	NhaD	13	B, E	420–576	11–12 (13)	NhaD Vpa
2.A.56	Tripartite ATP-independent Periplasmic Transporter	Trap-T	127	B, A	418–904	12 – 24 (12)	DctM Rca

[a] B, bacteria; A, archaea; E, eukaryotes.
[b] Size range based on full-length proteins.
[c] Predicted topologies, based on average hydropathy plots for all members from each of the families are underlined. The range of predicted topologies is also provided. When experimental data supporting a particular topology for any one member of the family are available, the number of TMSs is presented in parentheses.

average hydropathy, similarity and amphipathicity as a function of alignment position for each family after aligning the sequencing with the CLUSTAL X program [19]. This last method provides more reliable estimates of topology for families of proteins with uniform lengths and structures. It was used for the topological assignments presented in Table 1.

3. Results

Table 1 lists the 12 previously identified functionally characterized families that we have been able to establish as constituents of the IT superfamily. All of these families fall into TC category 2.A and therefore consist exclusively of secondary carriers. Eight of the 12 families consist of cation symporters and function in nutrient uptake. Three systems (NhaB, C and D) are Na^+:H^+antiporters, and one, the ArsB family, includes arsenite/antimonite efflux pumps which possibly function by uniport. These families exhibit tremendous size variation, with the smallest family (DcuC) having only 12 members and the largest family (DASS) having 154 members.

All of the 12 families listed in Table 1 have representation in the bacterial domain, but two of the families (ArsB and DASS) are ubiquitous, with constituents in all three domains of life (Archaea, Bacteria and Eukaryotes). Including these two families, six of the twelve families include archaeal homologues. The principal families are therefore (1) bacterial-specific (five families), (2) prokaryotic-specific with representation in both the bacterial and archaeal domains (four families), (3) bacterial and eukaryotic-specific, not yet being represented in the archaeal domain (one family) and (4) ubiquitous, being found in all three domains of living organisms (two families).

The size ranges of the proteins are presented in the sixth column of Table 1. Seven of the twelve families include members that fall into a relatively narrow size range (420–576 amino acyl residues). Exceptions where there is greater size range include the two families (ArsB and DASS) with established eukaryotic members that are generally much larger than their prokaryotic homologues (up to 923 residues). These larger proteins have N- (or occasionally C-) terminal hydrophilic domains of up to 300–450 residues, which are not homologous to domains in any characterized protein in the databases. These observations are in agreement with previous interkingdom analyses of transporter homologue sizes [28]. These studies revealed that bacterial homologues are on the average larger than archaeal protein while eukaryotic homologues are generally much larger than their prokaryotic homologues [28]. The LctP and the NhaC families both show substantial size variation. For LctP, almost all members of the family exhibit 14 well-conserved TMSs showing that this topological feature is a characteristic of this family (see below). The larger sizes of some members of the LctP, NhaC and TRAP-T families will be discussed in the sections devoted to these families.

Table 2 presents binary comparison scores that establish that all of the 12 families listed in Table 1 are homologous. In the derivation of these scores, the entire protein sequences were used. The values obtained suggest that the ArsB family is most closely related to the NhaD and NhaB families (comparison scores of 20.4 and 17.2 SD, respectively). The CitMHS family is most closely related to the ArsB and GntP families (scores of 15 SD). The Trap-T family appears to be most closely related to the DcuC family (score of 17.8 SD). Since members of both of these two last-mentioned families transport dicarboxylates, this observation suggests that a single component DcuC-like protein may have served as the evolutionary precursor for tripartite Trap-T family systems. Below we provide descriptions of the 12 constituent families of the IT superfamily.

3.1. The p-aminobenzoyl-glutamate transporter (AbgT) family (TC #2.A.68)

The AbgT family consists of a single functionally characterized protein, the AbgT (YdaH) protein of *E. coli* [15].

Table 2
Interfamilial binary comparisons for the 12 established families of the IT superfamily[a]

Family	Member	GI number	Family	Member	GI number	Gap score SD[a]
CitMHS	Bsu	16079739	ArsB	Pho1	14591655	15.1
CitMHS	Pae	15600661	DcuC	Pmu3	15602095	13.5
CitMHS	CitM Bsu	1705890	NhaC	Bha1	15616508	10.2
CitMHS	Bsu	16079739	DASS	Vch3	15600796	10.3
CitMHS	CitH Bsu	16080957	GntP	Sen1	16761410	15.0
DASS	Ecl	4323058	ArsB	ArsB Sxy	231566	13.8
AbgT	Bha1	15614315	GntP	Pae2	15596248	9.5
AbgT	AbgT Eco	2506659	DcuC	Pmu3	15602095	9.9
ArsB	Bsu1	16080656	Dcu	Hpy	15645345	10.7
ArsB	Mtu2	15609822	NhaB	Sty	16765147	17.2
NhaD	NhaD Vpa	3123728	ArsB	Mtu2	15609822	20.4
LctP	Oih	23097823	GntP	Fnu2	19703889	9.6
Trap-T	Oih1	23097696	DcuC	Pmu1	15603164	17.8

[a] The GAP program was used to align the sequences with a gap opening penalty of 8, a gap extension penalty of 2 and 500 random shuffles.

This protein is apparently cryptic in wild-type cells, but when expressed on a high copy number plasmid, or when expressed at higher levels due to mutation, it allows utilization of *p*-aminobenzoyl-glutamate as a source of *p*-aminobenzoate for *p*-aminobenzoate auxotrophs. *p*-Aminobenzoate is a constituent of, and a precursor for the biosynthesis of folic acid.

The *E. coli abgT* gene is preceded by two genes, *abgA* and *abgB*, which code for homologous amino acyl amino hydrolases. Because of the structural similarity of *p*-aminobenzoyl-glutamate to peptides, and the enzymatic activities of the *abgA* and *abgB* gene product homologues, it is possible that AbgT is a peptide transporter. Although the mechanism of energy coupling has not been established, an H^+ symport mechanism is inferred.

Twenty bacterial homologues comprise the AbgT family (see Table SA1 on our supmat website). The multiple alignment for this family is shown in Fig. SA1 on our website while the average hydropathy, amphipathicity and similarity plots for these proteins are shown in Fig. SA2 on the same website. Twelve peaks of hydrophobicity, possibly corresponding to 12 transmembrane segments (TMSs) were observed, and between putative TMSs 7 and 8, and TMSs 10 and 11, well-conserved amphipathic α-helical regions were identified. Each peak of hydrophobicity corresponds to a peak of average similarity except for putative TMS 12, which was poorly conserved although present in all family members. We conclude that all members of the AbgT family probably have 12 TMSs.

Examination of the multiple alignment revealed 26 fully conserved residues scattered throughout the alignment. Of these, seven were glycines (G), eight were prolines (P) and four were phenylalanines (F). Only five conserved residues were hydrophilic or semipolar (W, E, D, R and S). The most conserved region was in the C-terminal half. Putative TMSs 10 and 11 included 8 of the 26 fully conserved residues (see Fig. SA1).

The AbgT family phylogenetic tree revealed eight clusters that branched from points near the center of the tree (Fig. SA3). Several organisms exhibit multiple paralogues, and usually these proved to be divergent in sequence, suggesting that they arose by early gene duplication events. Two pairs of paralogues, from *Oceanobacillus iheyensis* (Oih) and *Caulobacter crescentus* (Ccr), clustered loosely together suggesting that each of these pairs arose by a late gene duplication event. The functionally characterized AbgT protein of *E. coli* does not cluster with any other protein. Gram-positive and Gram-negative bacterial proteins are interspersed. The appearance of the tree does not allow functional assignment of any homologue, as orthologous relationships are not obvious.

3.2. The arsenite–antimonite (ArsB) efflux family (TC #2.A.45)

Arsenite resistance (Ars) efflux pumps of bacteria consist either of two proteins, ArsB, the integral membrane constituent with 12 TMSs, and ArsA, the ATP-hydrolyzing, transport energizing subunit, as for the chromosomally encoded *E. coli* system, or of one protein, the ArsB integral membrane protein of the plasmid-encoded *Staphylococcus* system [29–31]. ArsA proteins have two ATP binding domains and probably arose by a tandem intragenic duplication event. ArsB proteins possess 12 transmembrane spanners and may also have arisen by tandem intragenic duplication (see below). Structurally, the Ars pumps superficially resemble ABC-type efflux pumps, but there is no significant sequence similarity between the Ars and ABC pumps. When only ArsB is present, the system operates by a membrane potential-dependent mechanism, and consequently it belongs in TC subclass 2.A. When ArsA is also present, ATP hydrolysis drives efflux, and consequently the system belongs in TC subclass 3.A. ArsB therefore appears twice in the TC system, but ArsA appears only once. This versatility of energy coupling is an exceptionally unusual characteristic of a transporter [3]. ArsB pumps actively expel both arsenite and antimonite.

The large ubiquitous ArsB family includes 90 members (Table SB1). The average hydropathy and similarity plots for these proteins are shown in Fig. SB2. The Fig. reveals two hydrophobic regions, the first with seven peaks of hydrophobicity, and the second with six peaks. A poorly conserved hydrophilic region separates these two relatively well-conserved hydrophobic regions. A poorly conserved N-terminal hydrophilic domain (~ 370 alignment positions) is found in seven proteins derived from *Drosophila* and mammals. Peaks 1–5 are well conserved as is peak 6, but peak 5*, between peaks 5 and 6, is found only in two mammalian proteins. It seems unlikely that peak 5* traverses the membrane, as this would cause these two proteins to have a topology that differs from all other homologues. It may therefore dip into the membrane from the extracytoplasmic side. We therefore predict that essentially all ArsB family members have 12 TMSs.

The multiple alignment (Fig. SB1) of the ArsB homologues revealed that only the 12 putative TMSs are well conserved (see Fig. SB2). Hydrophilic domains and inter-TMS loops are poorly conserved. No residue was fully conserved in all sequences, and no position exhibited exclusively conservative substitutions.

The ArsB phylogenetic tree is shown in Fig. SB3. The overall configuration of the tree reveals many divergent sequences branching from the center of the tree. The archaeal proteins are found in five clusters, but none of these archaeal protein clusters closely with a bacterial or eukaryotic protein. Among the archaea represented, only the two *Pyrococcus* species have two paralogues, and the two proteins in each of these two organisms are very divergent in sequence. The eukaryotic proteins fall into two major clusters, one consisting only of plant proteins, the other of animal proteins. There are two mouse paralogues and five sequence divergent *Drosophila* paralogues.

The bacterial proteins are derived from several bacterial kingdoms, and no prokaryotic organism has more than three paralogues. However, three are found in each of two Gram-positive bacteria, *Bacillus anthracis* and *Mycobacterium tuberculosis*. In both cases, two of the three paralogues are closely related while the third is distant. Several bacteria have two paralogues, and these are usually divergent in sequence.

The functionally characterized bacterial ArsB homologues, one from *E. coli* and one from *S. aureus*, are found in two tight clusters, one for Gram-negative bacteria, and one for Gram-positive bacteria, respectively. These proteins generally follow the phylogenies of the source organisms suggesting that they represent two clusters of orthologues. The Afe1 protein from *Acidithiobacillus ferrooxidans* is the only exception. A third related cluster consists of three archaeal proteins (Fac, Tac1 and Tvo1). These may also be orthologous ArsB homologues. Examination of the overall configuration of the tree reveals that proteins from distantly related organisms never cluster closely together, suggesting the absence of recent horizontal transfer of genes encoding these proteins between distantly related organisms.

3.3. The citrate-Mg^{2+}:H^+ (CitM)–citrate-Ca^{2+}:H^+(CitH) symporter (CitMHS) family (TC #2.A.11)

The two characterized members of the CitMHS family are both citrate uptake permeases from *Bacillus subtilis*. CitM is believed to transport a citrate^{2}-Mg^{2+} complex in symport with one H^+ per Mg^{2+}-citrate while CitH apparently transports a citrate^{2}-Ca^{2+} complex in symport with protons [32,33]. The cation selectivity of CitM is: Mg^{2+}, Mn^{2+}, Ba^{2+}, Ni^{2+}, Co^{2+}, Ca^{2+} and Zn^{2+} with an order of preference in this order [34]. CitM is highly specific for citrate and D-isocitrate and does not transport other di- and tri-carboxylates including succinate, *cis*-aconitate and tricarballylate [34,35]. For CitH, the cation specificity (in order of preference) is: Ca^{2+}, Ba^{2+} and Sr^{2+} [33]. The two proteins are 60% identical, contain about 400 amino acyl residues and possess 12 putative transmembrane spanners.

The CitMHS family consists of 21 bacterial proteins (Table SC1). Most represented organisms have just one CitMHS family member, but *Azotobacter vinelandii* and *Pseudomonas syringae* both have two very distant paralogues while *B. subtilis* has three fairly closely related paralogues, two of which are functionally characterized as noted above.

The multiple alignment (Fig. SC1) for the CitMHS family proteins revealed several well-conserved regions, the most impressive preceding and overlapping TMS 3. The average hydropathy and similarity plots are shown in Fig. SC2. The two halves show similar profiles with six TMSs each. Similar to the ArsB plot (Fig. SB2), peaks 1 and 2 and 7 and 8 are relatively close to each other. Amphipathic peaks were found between TMSs 2 and 3, 3 and 4, 8 and 9, and 9 and 10. These characteristics suggest that these two halves arose by an internal duplication event.

The CitMHS phylogenetic tree (Fig. SC3) revealed a reasonably tight cluster of γ-proteobacterial proteins (lower right hand side of the tree) which may consist exclusively of orthologues. The Gram-positive bacterial proteins are scattered and distantly related to each other, suggesting the occurrence of early gene duplication events during the evolution of this family.

3.4. The divalent anion:Na^+ symporter (DASS) family (TC #2.A.47)

Functionally characterized proteins of the DASS family transport (1) organic di- and tricarboxylates of the Krebs cycle as well as dicarboxylate amino acid and (2) inorganic sulfate, thiosulfate, selenate and phosphate. These proteins are found in Gram-negative and Gram-positive bacteria, cyanobacteria, archaea, plant chloroplasts, yeast and animals. Full-length members of this family vary in size from 432 amino acyl residues (*M. jannaschii*) to 923 residues (*Saccharomyces cerevisiae*). The three *S. cerevisiae* proteins are large (881–923 residues); the animal proteins are substantially smaller (539–616 residues), and the bacterial proteins are still smaller (461–612 residues). They exhibit 11–14 putative TMSs. An 11-TMS model for the animal NaDC-1 has been proposed [36]. This and the other NaDC isoforms cotransport three Na^+ with each dicarboxylate. Protonated tricarboxylates are also cotransported with three Na^+. Several organisms possess multiple paralogues of the DASS family (e.g., three for *E. coli*, three for *S. cerevisiae* and four for *C. elegans*).

The ubiquitous DASS family is the largest family in the IT superfamily with 154 members (Table SD1). The average hydropathy and similarity plots (Fig. SD2) reveal a poorly conserved N-terminal region, 450 residues long, found only in yeast and fungi with the exception of a single bacterial homologue from *Ralstonia metallodurans*. In view of the large size and diversity of organismal sources for this family, it is not surprising that no residue position proved to be fully conserved or even exhibited exclusively conservative substitutions (Fig. SD1). Only the TMSs proved to be fairly well conserved (Fig. SD2).

The DASS family phylogenetic tree (Fig. SD3) is large and reveals considerable sequence diversity. Only four archaeal proteins are found in the tree, and none of these proteins is closely related to any other protein. The animal proteins form a single large cluster plus an orphan protein (Dme3) that does not cluster with anything else. The yeast and fungal homologues form a single tight cluster, while the plant homologues fall into three fairly small clusters with one, two and five members, respectively.

Among eukaryotes, multiple paralogues are often observed. For example, *S. cerevisiae*, *Arabidopsis thaliana* and *C. elegans* each has three paralogues while *H. sapiens* and *D. melanogaster* have four and the mouse has six. No

archaeon has more than one, but bacteria may have as many as seven as found in *Desulfitobacterium hafniense* (where three clusters include four, two and one paralogues, respectively). *Novosphingobium aromaticivorans* and *Salmonella typhimurium* have four and five paralogues, respectively. With the exception of two close paralogues in *N. aromaticivorans*, all of these paralogues are distantly related to each other.

3.5. The C_4-dicarboxylate uptake (Dcu) family (TC #2.A.13)

Proteins of the Dcu family possess 12 putative transmembrane α-helical spanners, but DcuA has 10 experimentally determined TMSs with both the N- and C-termini localized to the periplasm [37]. For DcuA, the ''positive inside'' rule is obeyed, and two putative TMSs are localized to a cytoplasmic loop between TMSs 5 and 6 and in the C-terminal periplasmic region, respectively [37]. If this model is correct, the two halves of the protein have opposite orientation in the membrane. This could have arisen from a more typical 12 TMS permease with its N- and C-termini in the cytoplasm by topological inversion of the N-terminal five-TMS domain. A similar inversion has been documented for the lactose permease of *E. coli* induced by depletion of phosphatidyl ethanolamine in the membranes [38,39].

The two *E. coli* proteins, DcuA and DcuB, transport aspartate, malate, fumarate and succinate and function as antiporters with any two of these substrates. They exhibit 36% identity and 63% similarity, and both transport fumarate in exchange for succinate with about the same affinity (30 μM) [40]. Since DcuA is encoded in an operon with the gene for aspartase, and DcuB is encoded in an operon with the gene for fumarase, their physiological functions may be to catalyze aspartate:fumarate and fumarate:malate exchange during the anaerobic utilization of aspartate and fumarate, respectively. The electroneutral antiport of fumarate for succinate during anaerobic fumarate respiration has been demonstrated. Both permeases are induced under anaerobic conditions and are subject to catabolite repression. The two transporters can apparently substitute for each other under certain physiological conditions [41–43].

The Dcu family consists of 26 bacterial proteins (Table SE1). The average hydropathy plot reveals 12 putative TMSs with the first half being about equally well conserved as the second half (Fig. SE2). Many residues in these homologues are fully conserved as revealed in the multiple alignment (see Fig. SE1). Only one organism, *S. typhimurium*, has three paralogues, and as revealed by the phylogenetic tree (Fig. SE3), they are all sequence divergent. Several bacteria have two paralogues, and in no case are the two paralogues closely related. Two clusters of distantly related γ-proteobacterial orthologues (from *Haemophilus influenzae*, *H. somnus* and *Pasteurella multocida*) can be seen at the bottom of the tree (Fig. SE2).

3.6. The C_4-dicarboxylate uptake C (DcuC) family (TC #2.A.61)

A single functionally characterized protein is found in the DcuC family. This is an anaerobic C_4-dicarboxylate transporter (DcuC) of *E. coli*. The DcuC protein is induced only under anaerobic conditions and is not repressed by glucose. It may therefore function as a succinate efflux system during anaerobic glucose fermentation. However, when overexpressed, it can replace either DcuA or DcuB in catalyzing fumarate–succinate exchange and fumarate uptake.

The DcuC family consists of only 12 proteins (Table SF1), all from bacteria. As expected for a small family, many residues are fully conserved, and these are distributed fairly uniformly through the alignment (Fig. SF1). The average hydropathy plot (Fig. SF2) can be interpreted in terms of 12 TMSs with TMSs 1 and 2 and TMSs 7 and 8 close to each other. TMSs 5 and 11 show substantial hydrophilic character. Substantial peaks of amphipathicity are seen between TMSs 2 and 3, 3 and 4, 8 and 9, and 9 and 10. These are all characteristics of the CitMHS family. They provide evidence that these proteins arose by an intragenic duplication event.

The DcuC phylogenetic tree (Fig. SF3) shows that the three paralogues from *P. multocida* are all distantly related. Two of these have possible orthologues in *E. coli* while one has a probable orthologue in *H. influenzae*.

3.7. The gluconate:H^+ symporter (GntP) family (TC #2.A.8)

Protein members of the GntP family include known gluconate permeases of *E. coli* and *Bacillus* species as well as several functionally uncharacterized Orfs [44,45]. Four of the seven *E. coli* paralogues have been found to possess active gluconate uptake activity, and one of them can accommodate both L-idonate and D-gluconate although L-idonate is the physiological substrate [46]. These proteins are of about 450 residues and possess 12 putative transmembrane α-helical spanners.

The GntP family includes 75 members derived from both bacteria and archaea (Table SG1). Many organisms have multiple paralogues with *Salmonella enterica* (subspecies *typhi* and *typhimurium*) having five and both *E. coli* and *P. syringae* having four. The multiple alignment (Fig. SG1) revealed three fully conserved glycines at alignment positions 122, 126, and 133 (overlapping and following TMS 3) as well as a fully conserved proline at position 220. The average hydropathy and similarity plots could be best interpreted in terms of 12 TMSs (see Fig. SG2).

The GntP phylogenetic tree is shown in Fig. SG3. All five *S. enterica* paralogues are distantly related to each other. Most surprisingly, the seven *E. coli* proteins (four from *E. coli* K12 and three from *E. coli* 0157) are all fairly distantly related, suggesting that these two *E. coli* strains lack even one set of orthologues.

3.8. The lactate permease (LctP) family (TC #2.A.14)

Proteins of the LctP family have been found in Gram-negative and Gram-positive bacteria as well as archaea. One member of the family, from *E. coli*, is the lactate:H^+ symporter [47]. Two closely related paralogues have been found encoded within the *E. coli* genome, and one of these has been shown to be the glycolate uptake permease [48]. Both permeases transport the same acids (L- and D-lactate as well as glycolate), but their physiological substrates differ. The lactate permease transports both L- and D-lactate while the glycolate permease transports glycolate under physiological conditions. Regulatory effects determine the substrates transported in vivo.

Fifty-two proteins comprise the LctP family. Table SH1 lists these homologues which derive from bacteria and a few archaea. One bacterium, *D. hafniense*, has four paralogues while two *Bacillus* species, *B. anthracis* and *Bacillus halodurans*, have three each. All other organisms represented have one or two LctP family members.

The multiple alignment (Fig. SH1) reveals only three fully conserved residues near the end of the alignment, a glycine (position 602), a serine (position 611) and a phenylalanine (position 619). This alignment as well as the average hydropathy and similarity plots (Fig. SH2) suggest the presence of about 14 TMSs. Topological predictions for the individual proteins (Table SH1) suggest that most proteins have 14 TMSs, but three [two from *D. hafniense* (Dha3 and Dha4) as well as one from *Bacillus halodurans* (Bha2)] have 15, due to a C-terminal extension that includes an additional hydrophobic region. Based on motif analysis [12], the LctP family is most closely related to the GntP family. We noted that the average size for GntP homologues, predicted to have 12 TMSs, is about 450 residues, while that for the LctP homologues, predicted to have 14 TMSs, is about 535 residues. Alignment of representative GntP and LctP homologues revealed that the extra 85 residues in LctP family members occur in the N-terminal halves of these proteins relative to the GntP homologues. Both sets of proteins end together in a multiple alignment including proteins from both families (not shown). The LctP family may therefore have arisen by addition of two hydrophobic domains to the N terminus of an ancestral 12 TMS protein.

The phylogenetic tree for the LctP family (Fig. SH3) reveals minimally six clusters of proteins. Two of the four *D. halfniense* paralogues are found on a single branch, but the other two are distantly related to each other, one of them clustering loosely with the functionally characterized lactate and glyoxalate permeases of *E. coli*. It seems likely that the additional members of this cluster are orthologues that transport lactate and glyoxalate (top center of Fig. SH3). Considering the *Bacillus* homologues, two of the *B. subtilis* and *B. anthracis* paralogues and one of the *B. halodurans* paralogues are found in the large diverse cluster at the top of the tree that includes the two *E. coli* paralogues. However, none of these five proteins appears to be orthologous to any of the others. The remaining two *B. halodurans* paralogues cluster loosely together with the two closely related *D. hafniense* proteins, but the remaining *B. anthracis* protein is not part of this cluster. We therefore have no evidence for orthology among the different *Bacillus* proteins. The pairs of LctP paralogues in some organisms (i.e., *H. pylori*, *S. aureus* and *A. vinlandii*) are tightly clustered although the pairs of paralogues in other organisms (*M. magnetotacticum* and *D. desulfuricans*) are distantly related.

3.9. The NhaB Na^+:H^+ antiporter (NhaB) family (TC #2.A.34)

The *E. coli* NhaB is 58% identical to the orthologous *Vibrio alginolyticus* Na^+/H^+ antiporter [49]. Although the latter protein is predicted to exhibit 10–12 TMSs, construction of *nhaB–phoA* fusions led to evidence for a nine-TMS model with the N terminus in the cytoplasm and the C terminus in the periplasm [50]. A centrally located aspartyl residue in the third TMS of the *V. alginolyticus* homologue, conserved in all members of the family, has been shown to be essential for activity [51].

Only 14 proteins, all from γ-proteobacteria, comprise the NhaB family (Table SI1). These proteins are of fairly uniform size (499–533 residues). In the multiple alignment, about 50% of the residue positions exhibit identity in all proteins (Fig. SI1). In the C-terminal region, a 22-residue stretch shows absolute identity for all 14 proteins.

Between 8 and 12 TMSs were predicted for the individual proteins, and 12 peaks of hydrophobicity may be present in the average hydropathy plot (Fig. SI2). No organism has more than one NhaB family member. The phylogenetic tree (Fig. SI3) shows clustering of these proteins according to organismal (16S RNA) phylogeny. We therefore predict that they are all orthologues with the same function.

3.10. The NhaC Na^+:H^+ antiporter (NhaC) family (TC #2.A.35)

Two members of the NhaC family have been functionally characterized. One is believed to be a Na^+:H^+ antiporter [52]; the other is a malate·H^+:lactate·Na^+ antiporter [53]. Several paralogues are found in *Vibrio cholerae*, and two paralogues each are found encoded in the completely sequenced genomes of *H. influenzae* and *Bacillus subtilis*. *E. coli* lacks such a homologue. *Pyrococcus* species also have at least one homologue each. Thus, members of the NhaC family are found both in Gram-negative bacteria and Gram-positive bacteria as well as archaea. NhaC of *B. firmus* is 462 amino acyl residues long and possesses 11 or 12 putative transmembrane α-helical segments. MleN of *B. subtilis* (468 aas) exhibits 12 putative TMSs.

Sixty-five NhaC homologues were identified (Table SJ1). These proteins are derived from bacteria and archaea. Several organisms have multiple NhaC paralogues. For

example, *Fusobacterium nucleatum* has 10, *Halobacterium* sp. has five, *V. cholerae* has four, and *B. anthracis* and *D. desulfuricans* each have three. All other organisms represented have only one or two. Most of these proteins are predicted to have 11 or 12 TMSs (Table SJ1).

The multiple alignment revealed the presence of a single, fully conserved residue, a proline preceding the C-terminal TMS (Fig. SJ1). The hydropathy plot (Fig. SJ2) revealed 12 conserved peaks of hydropathy which were preceded by two well-defined but poorly conserved hydrophobic peaks present in three homologues, Dde2, Cje and Cpe1. The first of these is a 607-residue protein (Dde2) from *D. desulfuricans*, and it is predicted to have 14 TMSs. The other two are a 577-residue protein (Cje) from *Campylobacter jejuni* and a 571-residue protein (Cpe1) from *Clostridium perfringens*. Minimal sequence similarity between the common N-terminal TMSs in these three proteins suggests (but does not prove) a common origin. The Hsp4 protein from *Halobacterium* sp. has a C-terminal hydrophilic extension of about 100 residues.

The phylogenetic tree (Fig. SJ3) shows that the three large proteins with N-terminal extensions form a single tight cluster in spite of the considerable phylogenetic distance between their source organisms. It seems that an N-terminal hydrophobic extension that is lacking in the other family members was added to these three proteins during their evolutionary histories, possibly in a single common event.

The phylogenetic tree shown in Fig. SJ3 reveals about seven major branches stemming from the center of the tree. Six of the archaeal homologues cluster together at the bottom of the tree and are probably orthologous. Two of the remaining three halobacterial homologues cluster together while the third clusters loosely with a *D. desulfuricans* protein. The 10 fusobacterial paralogues all proved to be distantly related (Fig. SJ3). Two of the four *V. cholerae* paralogues cluster together with several other β- and γ-proteobacterial proteins while the other two are distantly related. Several orthologous relationships can be deduced from the tree.

3.11. The NhaD Na^+:H^+ antiporter (NhaD) family (TC #2.A.62)

A single member of the NhaD family has been characterized [54]. This protein is the NhaD protein of *Vibrio parahaemolyticus*. The protein has 443 amino acyl residues with 11 putative TMSs. It has been shown to catalyze Na^+/H^+ antiport, but Li^+ is also a substrate. The transporter exhibits activity only at basic pH (8–9) with virtually no activity at pH 7.0–7.5 [54].

The NhaD family consists of 13 proteins, all from Gram-negative bacteria except for a single plant protein, from *A. thaliana* (Table SK1). *A. thaliana* and one bacterium, *Microbulbifer degradans*, have two paralogues, but the other organisms represented have only one. Except for one large plant homologue (576 residues), all fall into the size range 420–477 residues.

The multiple alignment (Fig. SK1) showed 20 identities, eight of which are found together in putative TMS 5. The average hydropathy and similarity plots (Fig. SK2) show 12 well-conserved hydrophobic peaks preceded by one and possibly two poorly conserved hydrophobic peaks. The multiple alignment revealed that the large plant protein has an N-terminal hydrophilic extension of 100 residues not found in the other proteins. However, five of the homologues, four proteobacterial proteins as well as the large *A. thaliana* protein, have a poorly conserved 20-residue hydrophobic region that could span the membrane. These proteins may therefore have an extra N-terminal TMS lacking in the other NhaD family members.

The phylogenetic tree (Fig. SK3) shows that the four proteobacterial proteins with the extra putative N-terminal TMS cluster tightly together. These proteins may be orthologues. The three chlamydial proteins also cluster together as expected for orthologues, and surprisingly, the plant protein clusters with them. The two remaining members of the NhaD family, Mde2 of *M. degradans* and Rpa of *Rhodopseudomonas palustris*, are clearly not orthologous to the others. Therefore, a single unified function for the members of this family cannot be proposed.

3.12. The tripartite ATP-independent periplasmic transporter (TRAP-T) family (TC #2.A.56)

TRAP-T family permeases generally consist of three components, and these systems have so far been found in Gram-negative bacteria, Gram-positive bacteria and archaea. Several members of the family have been both sequenced and functionally characterized. The first system to be characterized was the DctPQM system of *Rhodobacter capsulatus* [55], and it is the prototype for the TRAP-T family [12,56].

DctP is a periplasmic dicarboxylate (malate, fumarate, succinate) binding receptor that is biochemically well characterized. DctQ is an integral cytoplasmic membrane protein (25 kDa) with four putative TMSs. DctM is a second integral cytoplasmic membrane protein (50 kDa) with 12 putative TMSs. These three proteins have been shown to be both necessary and sufficient for the proton motive force-dependent uptake of dicarboxylates into *R. capsulatus*.

In some TRAP-T systems, fused Q-M-type proteins instead of two separate Q- and M-type proteins are found, while in others, Q-P-type fusion proteins are found. The operon encoding the *Synechocystis* system includes a protein homologous to the glutamine binding protein of *E. coli*, and biochemical evidence has suggested that a glutamate transporter from *Rhodobacter sphaeroides* is a periplasmic binding protein-dependent, pmf-dependent secondary carrier [57]. An *E. coli* homologue can catalyze uptake of L-xylulose [58,59], while a homologous system in *Halomonas elongata* takes up

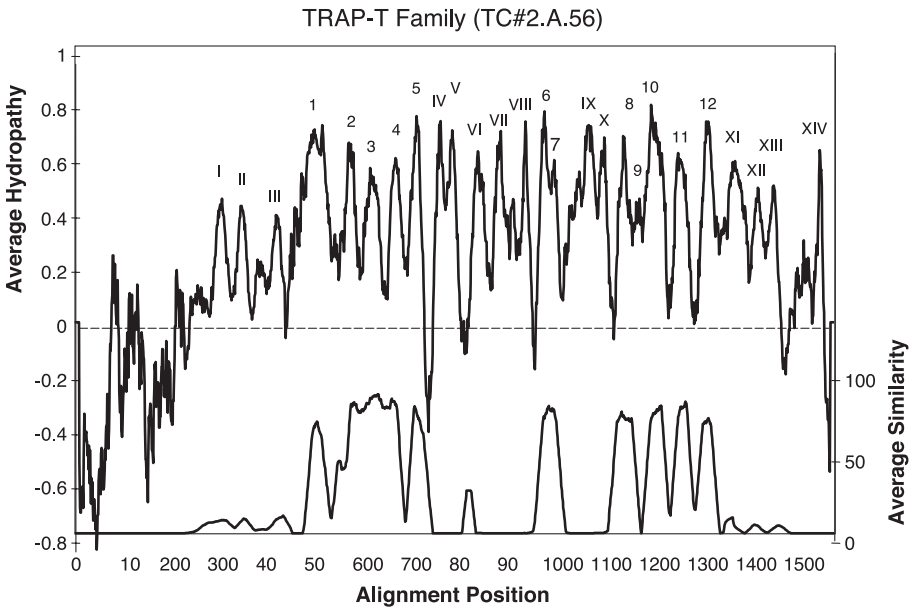

Fig. 1. Average hydropathy (top) and average similarity (bottom) for the members of the TRAP-T family (DctM homologues). The proteins listed in Table SL1 on our website were included in the analysis. The plots were generated with the AveHAS program, based on a multiple alignment generated with the CLUSTAL X program (see Computer methods).

ectoine and hydroxyectoine [60]. Surprisingly, the DctP dicarboxylate receptor is demonstrably homologous to both the YiaO L-xylulose receptor and the TeaA ectoine receptor. Thus, the TRAP-T family of permeases may be involved in the uptake of widely divergent types of compounds [56].

As noted above, the TRAP-T family consists of bacterial and archaeal tripartite systems with a 12-TMS protein, a

Table 3
Occurrence of extra (poorly conserved) TMSs in protein members of the TRAP-T family (TC #2.A.56)

TMS #[a]	Total # of proteins[b]	Cluster #[c]	Protein abbreviations[d]	Organismal types
I	44	1	Ape1, Afu3, Ape2, Afu2, Vch5, Sme7, Pmu4, Dra, Hsp, Son1, Pae3, Mma2, Rpa7, Afu1, Bme2, Pmu5, Hso4, Oih3, Bha4, Oih4, Bha3, Bha5, Dha5, Rpa8, Vch4	Proteobacteria, G+, Archaea
		2	Tpa	Spirochete
II		3	Atu5	Alpha
		5	Fnu2	unknown
		7c	Dde2, Pae2	Delta, Gamma
		9	Atu3, Fnu1, Hin1, Pmu2, Sme1, Vch3	Alpha, Gamma
III		10d	Dha1, Dde1	Delta, G+
		11	Atu4, Sme4, Bfu2, Rpa5	Alpha
IV	3	1	Vch5, Sme7, Pmu4	Gamma, Alpha
V	3	1	Vch5, Sme7, Pmu4	Gamma, Alpha
		9	Atu3	Alpha
VI	6	12	Sme5, Rsp5, Mlo, Sme6, Rpa6	Alpha
VII	2	12	Rpa6, Sme5	Alpha
VIII	5	12	Sme5, Rsp5, Mlo, Sme6, Rpa6	Alpha
IX	1	1	Hsp	Archaea
X	8	1	Ape2, Afu2, Hsp, Son1, Pae3, Rsp3, Mma3, Rpa7	Gamma, Alpha, Archaea
XI				
XII	27	1	Ape1, Afu3, Ape2, Afu2, Vch5, Sme7, Pmu4, Dra, Hsp, Son1, Pae3, Mma2 Rpa7, Afu1, Bme2, Pmu5, Hso4, Oih3, Bha4, Oih4, Bha3, Bha5, Dha5,	Proteobacteria, G+ Archaea
XIII			Rpa8, Vch4, Cgl	
XIV	3	1	Vch5, Sme7, Pmu4	Gamma, Alpha

[a] Number of the putative poorly conserved TMSs (hydrophobic peaks labeled with Roman numerals in Fig. 1).
[b] Total number of proteins out of 127 exhibiting this (or these) putative TMSs.
[c] Phylogenetic clusters including proteins that possess this TMS (or these TMSs) (see Fig. 2).
[d] The abbreviations are as indicated in Table SL1. They indicate the organism and number of the paralogues (e.g., Ape2 = *A. pernix*, paralogue #2).

four-TMS protein and an extracytoplasmic receptor [12,56]. One hundred twenty-seven homologues of the large 12-TMS protein (DctA homologues) were identified (Table SL1), and of the three TRAP-T family constituents, only these proteins proved to be members of the IT superfamily. The organisms with the most paralogues were all α-proteobacteria (*R. sphaeroides*, 10; *R. palustris* and *Sinorhizobium meliloti*, eight, and *Agrobacterium tumefaciens*, seven). Other organisms have one to six TRAP-T family members. Of the archaea, *Archaeoglobus fulgidus* has three while *Aeropyrum pernix* has two.

The multiple alignment (Fig. SL1) and the average hydropathy/similarity plots (Fig. 1) revealed numerous peaks of hydrophobicity, some well conserved and others poorly conserved. The best-conserved region (residue positions 561–676) encompasses TMSs 2, 3 and 4. There is not a single gap in the alignment of this entire 115-residue segment. It can therefore be concluded that this region is functionally and structurally important.

In the plot depicted in Fig. 1, a total of 26 hydrophobic peaks are observed, 12 well conserved (indicated by Arabic numbers) and 14 poorly conserved (indicated by Roman numerals). Some of the poorly conserved peaks preceded (three peaks at the N-termini) and followed (four peaks at the C-termini) the 12 well-conserved TMSs. However, others were present between well-conserved TMSs 5 and 6 (five poorly conserved peaks) and 7 and 8 (two poorly conserved peaks). Thus, referring to Fig. 1 and Table 3, the N-terminal poorly conserved peaks (I–III) are found in about 44 proteins from all types of prokaryotes. These include proteobacteria, *Flavobacterium*, spirochetes, Gram-positive bacteria, and archaea (phylogenetic clusters 1, 2, 3, 5, 7, 9, 10 and 11; see Table 3). The central poorly conserved peaks (IV–X) are found in between 1 and 8 proteins, depending on the specific peak. These proteins with these extra TMSs are from α- and γ-proteobacteria as well as three archaea that have only peaks IX and X (Table 3). Finally, peaks XI, XII and XIII are found in 27 proteins derived from both bacteria and archaea. These proteins are all found in phylogenetic cluster #1 (see Table 3 and below). Just three proteins have peak XIV, and these are from α- and γ-proteobacteria. If we assume that the conserved peaks have the same orientation in

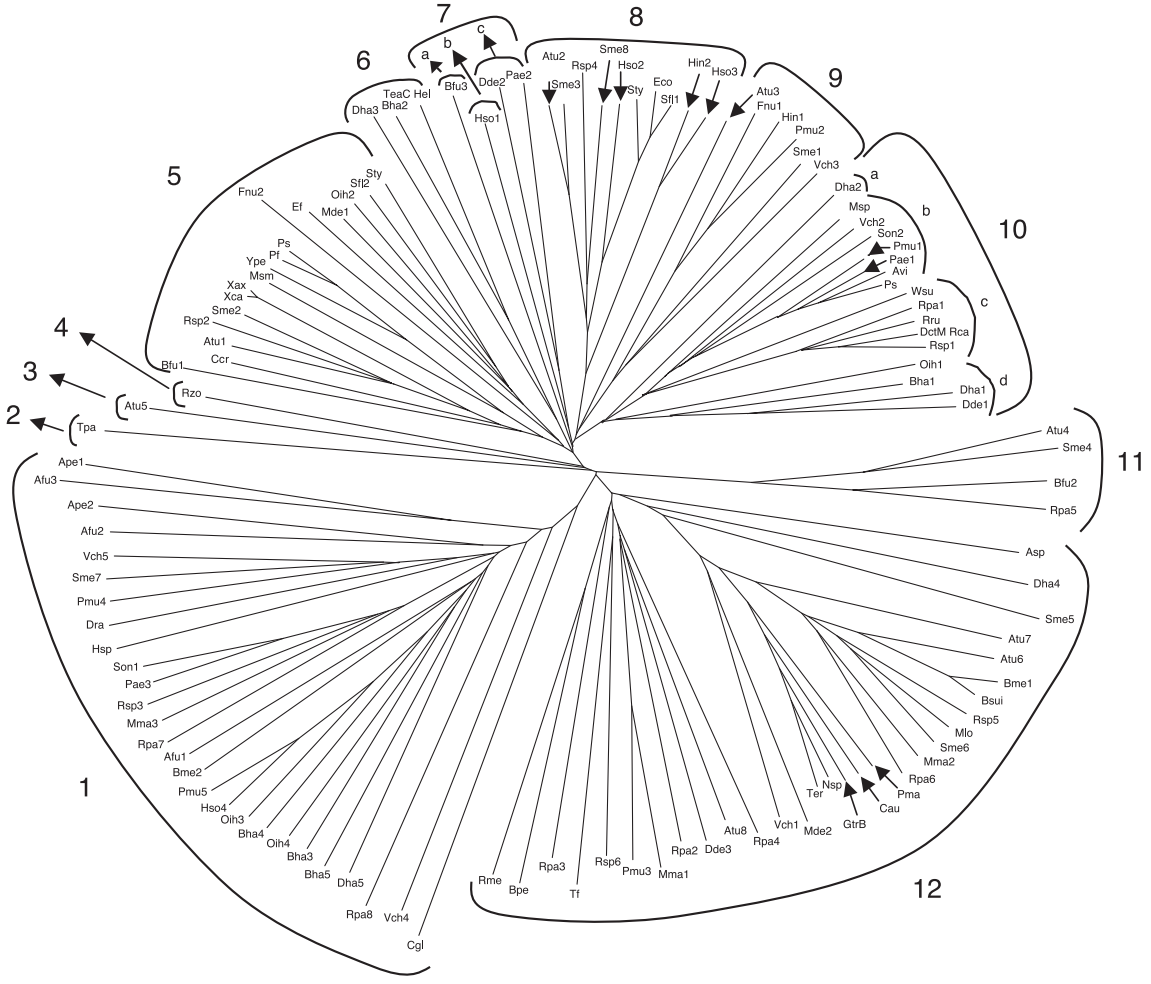

Fig. 2. Complete phylogenetic tree for the DctM homologues of the TRAP-T family. Abbreviations of the proteins are provided in Table SL1 on our supmat website. Phylogenetic clusters are labeled by number and discussed in the text. The CLUSTAL X program provided the multiple alignment upon which the tree was based (see Computer methods).

the membrane, then the two poorly conserved central hydrophobic regions (peaks IV–VIII and IX–X) may each span the membrane an even number of times (possibly twice). These results suggest that in contrast to other families in the IT superfamily, the DctA homologues of the TRAP-T family are extremely diverse topologically with additions and/or insertions at both ends and at two central locations.

The phylogenetic tree for the DctA homologues is shown in Fig. 2. All of the 27 proteins represented in cluster 1 (lower left hand side of Fig. 2) have the large N- and C-terminal hydrophobic extensions discussed above except for Cgl that lacks the N-terminal extension. Additional proteins with only the N-terminal extensions include the cluster 2, 3, 9 and 11 proteins plus a few proteins that fall into other clusters. Homologues with the first central hydrophobic insertion (putative TMSs IV–VIII) are all in clusters 1 and 12, except for one protein (Atu3) found in cluster 9. Homologues with the second central hydrophobic insertion (putative TMS IX and X) are found in subclusters of cluster 1 (the archaea, Afu2, Ape2 and Hsp, as well as the proteobacteria, Son1, Pae3, Mma3, Rsp3 and Rpa7).

Potential groups of orthologous proteins can be identified. For example, in clusters 10b and 10c, α-, γ- and ε-proteobacterial proteins cluster approximately according to organismal phylogeny. Only one protein per organism is found in this region. Since this cluster includes the functionally characterized DctM dicarboxylate transport protein of the TRAP-T family, it can be assumed that all of these proteins are dicarboxylate transporters. Similarly, clusters 5 and 8 probably include orthologues suggesting that each of these two sets of proteins serves a single unified function.

3.13. Internal duplications giving rise to proteins with two homologous halves, each of six TMSs

We attempted to demonstrate homology of the first and second halves of the 12 TMS proteins in all 12 families of the IT superfamily. We were successful in establishing homology for four of these families, the ArsB, DASS, GntP and NhaD families. Comparison scores for the alignments shown in Figs. S3A, 3B, 3C and 3D (see our supmat website) are 19 SD (ArsB), 23 SD (DASS), 12.2 SD (GntP) and 9.4 SD (NhaD). These values are sufficient to establish homology, showing that these proteins arose by a tandem internal gene duplication event. The same may be true for other families of the IT superfamily, and in several of them, sequence similarity between the two halves could be observed. However, comparison scores did not exceed 9 SD, so by this criterion, we could not establish homology.

4. Discussion

The TC system includes each of the recognized families of the IT superfamily under a distinct TC number (see Table 1). In order to integrate new information concerning superfamily status without disrupting the current TC numerology, we have constructed a TCDB hyperstructure that identifies established relationships between families [61].

The TCDB user interface automatically notifies the user when a TC family belonging to the IT superfamily, or any other superfamily, is viewed. The user can then traverse the superfamily hyperlink and view the familial members of the superfamily along with descriptions and references. A list containing all superfamilies thus constructed can be retrieved allowing direct access to members of all such superfamilies [61].

The work reported here extends a previous effort from our laboratory [12] to use established statistical methods to define a large superfamily of ITs that include 12 previously recognized families listed in the current TC system (see Table 1). Rabus et al. [12] showed that proteins within many of these families exhibit limited motif similarities, but homology was not established. More recently, Lolkema and Slotboom [14] used hydropathy profile analysis to provide independent evidence for a common evolutionary origin of many of these proteins, but again, homology was not established. Moreover, by our criteria, at least three of the families included in the "single structural class" of Lolkema and Slotboom [14] (e.g., AtoE (TC #2.A.1.37), CCS (TC #2.A.24) and ESS (TC #2.A.27)) do not appear to belong in the IT superfamily (unpublished results). The AtoE family members clearly belong to the major facilitator superfamily while we could not detect sequence or motif similarity of the CCS and ESS proteins with any of the IT superfamily proteins. This apparent discrepancy may reflect the probability that proteins arising independently can exhibit similar hydrophobic profiles. Alternatively, the approach of Lolkema and Slotboom [14] may have detected distant relationships not recognized by the methods we used.

Almost all of the functionally characterized transporters of the IT superfamily transport ionic species. However, these include both organic and inorganic species, as well as both cationic and anionic species. The only exceptions so far are members of the periplasmic receptor-dependent TRAP-T family, some of which may transport neutral substrates (e.g., ribulose and ectoine).

The substrate transport stoichiometries of IT superfamily members vary tremendously. For example, NaDC1 in the DASS family takes up dicarboxylates in symport with 3 Na^+ ions [36,62] while ArsB proteins export arsenite, probably by a uniport mechanism. Several members of the IT superfamily catalyze Na^+/H^+ antiport (members of the NhaB, C, and D families), but one member of the NhaC family has surprisingly been reported to catalyze malate•H^+/lactate•Na^+ antiport [53]. Other substrates of the IT superfamily include a variety of di- and tricarboxylates, including dicarboxylate amino acids. In the CitMHS family, the substrate tricarboxylates must be complexed with a divalent cation. Inorganic anions such as arsenite, antimonite, sulfate, thiosulfate, phosphate and selenate are also substrates

of specific DASS family transporters. One family (AbgT) transports *p*-aminobenzoyl glutamate and possibly other peptide-like substances. It is not clear why a superfamily of transporters capable of evolving specificity for such a broad range of ionic compounds would be so restricted with respect to the ability to evolve the capability to transport neutral or zwiterionic molecules of similar structure. However, many families of transporters exhibit specificity for a particular class of compounds [3,4].

Two novel features of the IT superfamily appear to be highly unusual. One is the ability of some of the members to acquire auxiliary subunits as is observed for members of the TRAP-T family which incorporate and depend upon an extracytoplasmic receptor as well as a non-homologous integral membrane protein [12,55,56] and the ArsB family, members of which can function either with or without a coupling ATPase [29–31]. The other is the tremendous topological variation that is observed for members of certain families within the IT superfamily. Since some members of the IT superfamily arose by duplication of a six-TMS-encoding genetic element, it is reasonable to suggest that the same is true for all families within the IT superfamily. However, LctP family members differ from GntP family members by the apparent presence of two extra hydrophobic TMSs, and three LctP family members have a C-terminal extension with still one more putative TMS. In the NhaC and NhaD families, some members have large hydrophilic N-terminal domains that include either one or two hydrophobic segments that could span the membrane. Additionally, for two of the families, Dcu and NhaB, 12 TMS topologies can be predicted based on average hydropathy profiles, but experimental data support a 10-TMS topology with N- and C-termini in the periplasm for the Dcu family member, DcuA of *E. coli* [37], and a nine-TMS topology with the N terminus in the cytoplasm and the C terminus in the periplasm for an NhaB family member from *V. alginolyticus* [50]. Most striking, however, is the tremendous topological variation observed for DctM homologues in the TRAP-T family (see Fig. 1 and Table 3). Thus, extra putative TMSs were found both N- and C-terminal to the common 12 TMSs as well as in two internal sites. Moreover, since several members of this family exhibit each of these characteristics, they cannot be attributed to sequencing errors or some other type of artifact. It may be that members of the IT superfamily will prove to exhibit striking topological flexibility for specific functional or structural reasons that are not yet understood.

Using the phylogenetic approach, we could define orthologous and paralogous relationships of many of the proteins that comprise a family within the IT superfamily. Thus, for example, we provided evidence that the NhaB family consists entirely of orthologues. These proteins may all prove to exhibit a single unified function, both biochemically and physiologically. Within specific phylogenetic clusters of the TRAP-T, ArsB, NhaC and CitMHS families, the phylogenetic relationships of the proteins to each other suggested orthology, and hence common function. The virtual lack of closely related proteins identified in phylogenetically distant organisms clearly points to a rarity of horizontal transfer of genes encoding IT superfamily members both between the different domains of life, and also between different kingdoms within a single domain. This observation is in general agreement with a similar conclusion resulting from the analyses of many other families of transporters [62,63, but also, see Ref. [10]].

The tremendous variation in the numbers of paralogues present in different bacteria proved surprising. Thus, *F. nucleatum* encodes within its genome 10 sequence dissimilar paralogues of the NhaC family of (putative) Na^+/H^+ antiporters although *E. coli* has none at all. Again, *R. spheroides* and several other α-proteobacteria have between 7 and 10 paralogues of DctM within the TRAP-T family although no other prokaryote has so many members of this family. These remarkable observations must reflect the specific physiological needs of each of these organisms. However, what these needs are and how large numbers of paralogues satisfy these needs remain topics for future study.

Recognition of the 12 families as constituents of a single superfamily allows us to extrapolate structural, functional and mechanistic data from one or a few members of the superfamily to all others to an extent dictated by the phylogenetic distances between any two such superfamily members. This fact provides what may prove to be the greatest value of phylogenetic analyses such as those reported here. We hope that this work will provide a guide for future molecular genetic, biophysical and biochemical analyses of function and mechanism for the diverse group of related transporters found within the IT superfamily.

Acknowledgements

We wish to thank Nung R. Lin for screening AtoE, CCS and ESS family members for similarity to IT superfamily members, Torsten von Rozycki for assistance with statistical analyses of internal repeat sequences, Abraham B. Chang for help with the preparation of the figures, and Mary Beth Hiller for assistance in the preparation of this manuscript. The work in our laboratory was supported by NIH grants GM 64368 and GM 55434.

References

[1] M.H. Saier Jr., Computer aided analysis of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution, Microbiol. Rev. 58 (1994) 71–93.
[2] M.H. Saier Jr., Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya, in: R.K. Poole (Ed.), Advances in Microbial Physiology, Academic Press, San Diego, 1998, pp. 81–136.
[3] M.H. Saier Jr., A functional-phylogenetic classification system for transmembrane solute transporters, Microbiol. Mol. Biol. Rev. 64 (2000) 354–411.

[4] W. Busch , M.H. Saier Jr., The Transporter Classification (TC) system, 2002, CRC Crit. Rev. Biochem. Mol. Biol. 37 (2002) 287–337.
[5] S.S. Pao, I.T. Paulsen, M.H. Saier Jr., The major facilitator superfamily, Microbiol. Mol. Biol. Rev. 62 (1998) 1–32.
[6] M.H. Saier Jr., J.T. Beatty, A. Goffeau, K.T. Harley, W.H.M. Heijne, S.-C. Huang, D.L. Jack, P.S. Jahn, K. Lew, J. Liu, S.S. Pao, I.T. Paulsen, T.-T. Tseng, P.S. Virk, The major facilitator superfamily, J. Mol. Microbiol. Biotechnol. 1 (1999) 257–279.
[7] D.L. Jack, I.T. Paulsen, M.H. Saier Jr., The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations, Microbiology 146 (2000) 1797–1814.
[8] T.-T. Tseng, K.S. Gratwick, J. Kollman, D. Park, D.H. Nies, A. Goffeau, M.H. Saier Jr., The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins, J. Mol. Microbiol. Biotechnol. 1 (1999) 107–125.
[9] D.L. Jack, N.M. Yang, M.H. Saier Jr., The drug/metabolite transporter superfamily, Eur. J. Biochem. 268 (2001) 3620–3639.
[10] R.N. Hvorup, B. Winnen, A. Chang, Y. Jiang, X. Zhou, M.H. Saier Jr., The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily, Eur. J. Biochem. 270 (2003) 799–813.
[11] M. Vrljic, J. Garg, A. Bellmann, S. Wachi, R. Freudl, M.J. Malecki, H. Sahm, V.J. Kozina, L. Eggeling , M.H. Saier Jr., The LysE superfamily: topology of the lysine exporter LysE of *Corynebacterium glutamicum*, a paradyme for a novel superfamily of transmembrane solute translocators, J. Mol. Microbiol. Biotechnol. 1 (1999) 327–336.
[12] R. Rabus, D.L. Jack, D.J. Kelly, M.H. Saier Jr., TRAP transporters: an ancient family of extracytoplasmic solute-receptor-dependent secondary active transporters, Microbiology 145 (1999) 3431–3445.
[13] T.L. Bailey, M. Gribskov, Combining evidence using *p*-values: application to sequence homology searches, Bioinformatics 14 (1998) 48–54.
[14] J.S. Lolkema, D.-J. Slotboom, Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis, J. Mol. Biol. 327 (2003) 901–909.
[15] M.J. Hussein, J.M. Green, B.P. Nichols, Characterization of mutations that allow *p*-aminobenzoyl-glutamate utilization by *Escherichia coli*, J. Bacteriol. 180 (1998) 6260–6268.
[16] J. Devereux, P. Haeberli, O. Smithies, A comprehensive set of sequence analysis programs for the VAX, Nucleic Acids Res. 12 (1984) 387–395.
[17] Y. Zhai, M.H. Saier Jr., A simple sensitive program for detecting internal repeats in sets of multiply aligned homologous proteins, J. Mol. Microbiol. Biotechnol. 4 (2002) 29–31.
[18] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[19] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.
[20] D.-F. Feng, R.F. Doolittle, Progressive alignment and phylogenetic tree construction of protein sequences, Methods Enzymol. 183 (1990) 375–387.
[21] R.D. Page, TreeView: an application to display phylogenetic trees on personal computers, Comput. Appl. Biosci. 12 (1996) 357–358.
[22] T.A. Tatusova, T.L. Madden, BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences, FEMS Microbiol. Lett. 174 (1999) 247–250.
[23] X. Zhou, N.M. Yang, C.V. Tran, R.N. Hvorup , M.H. Saier Jr., Web-based programs for the display and analysis of transmembrane α-helices in aligned protein sequences, J. Mol. Microbiol. Biotechnol. 5 (2003) 1–6.
[24] A. Krogh, B. Larsson, G. von Heijne, E. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model. Application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[25] G.E. Tusnady, I. Simon, Principles governing amino acid composition of integral membrane proteins: application to topology prediction, J. Mol. Biol. 283 (1998) 489–506.
[26] Y. Zhai, M.H. Saier Jr., A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence, J. Mol. Microbiol. Biotechnol. 3 (2001) 501–502.
[27] Y. Zhai, M.H. Saier Jr., 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 (2001) 285–286.
[28] Y.J. Chung, C. Krueger, D. Metzgar, M.H. Saier Jr., Size comparisons among integral membrane transport protein homologues in bacteria, archaea, and eucarya, J. Bacteriol. 183 (2001) 1012–1021.
[29] C. Rensing, M. Ghosh, B.P. Rosen, Families of soft-metal-ion transporting ATPase, J. Bacteriol. 181 (1999) 5891–5897.
[30] B.R. Rosen, Bacterial resistance to heavy metals and metalloids, JBIC 1 (1996) 273–277.
[31] C. Xu, T. Zhou, M. Kuroda, B.P. Rosen, Metalloid resistance mechanisms in prokaryotes, J. Biochem. 123 (1998) 16–23.
[32] A. Boorsma, M.E. van der Rest, J.S. Lolkema, W.N. Konings, Secondary transporters for citrate and the Mg^{2+}-citrate complex in *Bacillus subtilis* are homologous proteins, J. Bacteriol. 178 (1996) 6216–6222.
[33] B.P. Krom, J.B. Warner, W.N. Konings, J.S. Lolkema, Complementary metal ion specificity of the metal-citrate transporters CitM and CitH of *Bacillus subtilis*, J. Bacteriol. 182 (2000) 6374–6381.
[34] H. Li, A.M. Pajor, Functional characterization of CitM, the Mg^{2+}-citrate transporter, J. Membr. Biol. 185 (2002) 9–16.
[35] J.B. Warner, J.S. Lolkema, Growth of *Bacillus subtilis* on citrate and isocitrate is supported by the Mg^{2+}-citrate transporter CitM, Microbiology 148 (2002) 3405–3412.
[36] A.M. Pajor, Sodium-coupled transporters for Krebs Cycle intermediates, Annu. Rev. Physiol. 61 (1999) 663–682.
[37] P. Golby, D.J. Kelly, J.R. Guest, S.C. Andrews, Topological analysis of DcuA, an anaerobic C_4-dicarboxylate transporter of *Escherichia coli*, J. Bacteriol. 180 (1998) 4821–4827.
[38] M. Bogdanov, W. Dowhan, Phospholipid-assisted protein folding: phosphatidyl-ethanolamine is required at a late step of the conformational maturation of the polytopic membrane protein lactose permease, EMBO J. 17 (1998) 5255–5264.
[39] M. Bogdanov, P.N. Heacock, W. Dowhan, A polytopic membrane protein displays a reversible topology dependent on membrane lipid composition, EMBO J. 21 (2002) 2107–2116.
[40] I.G. Janausch, E. Zientz, Q.H. Tran, A. Kroger, G. Unden, C4-dicarboxylate carriers and sensors in bacteria, Biochim. Biophys. Acta 1553 (2002) 39–56.
[41] P. Engel, R. Krämer, G. Unden, Transport of C_4-dicarboxylates by anaerobically grown *Escherichia coli*: energetics and mechanism of exchange, uptake and efflux, Eur. J. Biochem. 222 (1994) 605–614.
[42] S. Six, S.C. Andrews, G. Unden, J.R. Guest, *Escherichia coli* possesses two homologous anaerobic C_4-dicarboxylate membrane transporters (DcuA and DcuB) distinct from the aerobic dicarboxylate transport system (Dct), J. Bacteriol. 176 (1994) 6470–6478.
[43] G. Unden, J. Bongaerts, Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors, Biochim. Biophys. Acta 1320 (1997) 217–234.
[44] N. Peekhaus, S. Tong, J. Reizer, M.H. Saier Jr., E. Murray, T. Conway, Characterization of a novel transporter family that includes multiple *Escherichia coli* gluconate transporters and their homologues, FEMS Microbiol. Lett. 147 (1997) 233–238.
[45] A. Reizer, J. Deutscher, M.H. Saier Jr., J. Reizer, Analysis of the gluconate (gnt) operon of *Bacillus subtilis*, Mol. Microbiol. 5 (1991) 1081–1089.
[46] C. Bausch, N. Peekhaus, C. Utz, T. Blais, E. Murray, T. Lowary, T. Conway, Sequence analysis of the GntII (subsidiary) system for

gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*, J. Bacteriol. 180 (1998) 3704–3710.
[47] J.M. Dong, J.S. Taylor, D.J. Latour, S. Iuchi, E.C.C. Lin, Three overlapping *lct* genes involved in L-lactate utilization by *Escherichia coli*, J. Bacteriol. 175 (1993) 6671–6678.
[48] M.F. Nunez, M.T. Pellicer, J. Badia, J. Aguilar, L. Baldoma, The gene *yghK* linked to the *glc* operon of *Escherichia coli* encodes a permease for glycolate that is structurally and functionally similar to L-lactate permease, Microbiology 147 (2001) 1069–1077.
[49] E. Pinner, E. Padan, S. Schuldiner, Cloning, sequencing and expression of the NhaB gene, encoding a Na^+:H^+ antiporter in *Escherichia coli*, J. Biol. Chem. 267 (1992) 11064–11068.
[50] H. Enomoto, T. Unemoto, M. Nishibuchi, E. Padan, T. Nakamura, Topological study of the *Vibrio alginolyticus* Na^+/H^+ antiporter using gene fusions in *Escherichia coli* cells, Biochim. Biophys. Acta 1370 (1998) 77–86.
[51] T. Nakamura, Y. Fujisaki, H. Enomoto, Y. Nakayama, T. Takabe, N. Yamaguchi, N. Uozumi, Residue aspartate-147 from the third transmembrane region of Na^+/H^+ antiporter NhaB of *Vibrio alginolyticus* plays a role in its activity, J. Bacteriol. 183 (2001) 5762–5767.
[52] M. Ito, A.A. Guffanti, J. Zemsky, D.M. Ivey, T.A. Krulwich, Role of the nhaC-encoded Na^+/H^+ antiporter of alkaliphilic *Bacillus firmus* OF4, J. Bacteriol. 179 (1997) 3851–3857.
[53] Y. Wei, A.A. Guffanti, M. Ito, T.A. Krulwich, *Bacillus subtilis* YqkI is a novel malic/Na^+-lactate antiporter that enhances growth on malate at low protonmotive force, J. Biol. Chem. 275 (2000) 30287–30292.
[54] K. Nozaki, T. Kuroda, T. Mizushima, T. Tsuchiya, A new Na^+/H^+ antiporter, NhaD, of *Vibrio parahaemolyticus*, Biochim. Biophys. Acta 1369 (1998) 213–220.
[55] J. Forward, M.C. Behrendt, N.R. Wyborn, R. Cross, D.J. Kelly, TRAP Transporters: a new family of periplasmic solute transport systems encoded by the dctPQM genes of *Rhodobacter capsulatus* and by homologs in diverse Gram-negative bacteria, J. Bacteriol. 179 (1997) 5482–5493.
[56] D.J. Kelly, G.H. Thomas, The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea, FEMS Microbiol. Rev. 25 (2001) 405–424.
[57] M.H.J. Jacobs, T. van der Heide, A.J.M. Driessen, W.N. Konings, Glutamate transport in *Rhodobacter sphaeroides* is mediated by a novel binding-protein dependent secondary transport system, Proc. Natl. Acad. Sci. U. S. A. 93 (1996) 12786–12790.
[58] J. Reizer, A. Charbit, A. Reizer, M.H. Saier Jr., Novel phosphotransferase system genes revealed by bacterial genome analysis: operons encoding homologues of sugar-specific permease domains of the phosphotransferase system and pentose catabolic enzymes, Genome Sci. Technol. 1 (1996) 53–75.
[59] J.C. Sanchez, R. Gimenez, A. Schneider, W.D. Fessner, L. Baldoma, J. Aguilar, J. Badia, Activation of a cryptic gene encoding a kinase for L-xylulose opens a new pathway for the utilization of L-lyxose by *Escherichia coli*, J. Biol. Chem. 269 (1994) 29665–29669.
[60] K. Grammann, A. Volke, H.J. Kunte, New type of osmoregulated solute transporter identified in halophilic members of the *Bacteria* domain: TRAP transporter TeaABC mediates uptake of ectoine and hydroxyectoine in Halomonas elongata DSM 2581^T, J. Bacteriol. 184 (2002) 3078–3085.
[61] C.V. Tran, N.M. Yang, M.H. Saier Jr., TC-DB: an architecture for membrane transport protein analysis, Proc. 2nd Intl. IEEE Computer Society Computational Systems Bioinformatic Conference (CSB 2003), 2003, p. 658.
[62] A.M. Pajor, N. Sun, H.G. Valmonte, Mutational analysis of histidine residues in the rabbit Na^+/dicarboxylate co-transporter NaDC-1, Biochem. J. 331 (1998) 257–264.
[63] M.H. Saier Jr., Answering fundamental questions in biology with bioinformatics, ASM News 69 (2003) 175–181.