J. Mol. Microbiol. Biotechnol. (1999) 1(1): 107-125.

The RND Permease Superfamily: An Ancient, Ubiquitous and Diverse Family that Includes Human Disease and Development Proteins

Tsai-Tien Tseng[1], Kevin S. Gratwick[1], Justin Kollman[1], Daniel Park[1], Dietrich H. Nies[2], André Goffeau[3], and Milton H. Saier, Jr.[1]*

[1] Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA
[2] Institut für Mikrobiologie, University of Halle, D-06099 Halle, Germany
[3] Unité de Biochimie Physiologique, Université Catholique de Louvain, Place Croix du Sud 2-20, B-1348 Louvain-La-Neuve, Belgium

Abstract

A previous report identified and classified a small family of Gram-negative bacterial drug and heavy metal efflux permeases, now commonly referred to as the RND family (TC no. 2.6). We here show that this family is actually a ubiquitous superfamily with representation in all major kingdoms. We report phylogenetic analyses that define seven families within the RND superfamily as follows: (1) the heavy metal efflux (HME) family (Gram negative bacteria), (2) the hydrophobe/amphiphile efflux-1 (HAE1) family (Gram negative bacteria), (3) the nodulation factor exporter (NFE) family (Gram negative bacteria), (4) the SecDF protein-secretion accessory protein (SecDF) family (Gram negative and Gram positive bacteria as well as archaea), (5) the hydrophobe/amphiphile efflux-2 (HAE2) family (Gram positive bacteria), (6) the eukaryotic sterol homeostasis (ESH) family, and (7) the hydrophobe/amphiphile efflux-3 (HAE3) family (archaea and spirochetes). Functionally uncharacterized proteins were identified that are members of the RND superfamily but fall outside of these seven families. Some of the eukaryotic homologues function as enzymes and receptors instead of (or in addition to) transporters. The sizes and topological patterns exhibited by members of all seven families are shown to be strikingly similar, and statistical analyses establish common descent. Multiple alignments of proteins within each family allow derivation of family-specific signature sequences. Structural, functional, mechanistic and evolutionary implication of the reported results are discussed.

Introduction

Niemann-Pick type C (NP-C) disease in humans is an inherited homozygous recessive lipid storage disorder that most strongly affects the viscera and central nervous system and leads to mental retardation and early death (Pentchev *et al*., 1995). Cells of NP-C patients release low density lipoprotein-derived cholesterol from lysosomes to the endoplasmic reticulum and plasma membrane at subnormal rates (Pentchev *et al*., 1987; Neufeld *et al*., 1996). A genetically defective protein, responsible for the NP-C disorder, NPC, has been identified, but its biochemical function is not known (Vanier *et al*., 1996; Erickson *et al*., 1997; Watari *et al*., 1999).

NPC homologues in animals include the "Patched" morphogen receptor of *Drosophila melanogaster* (Johnson *et al*., 1996), the sterol regulatory element binding protein (SREBP) cleavage-activating protein, SCAP of Chinese hamster ovary cells (Hua *et al*., 1996), and a yeast protein encoded by gene YPL006w in *Saccharomyces cerevisiae* (Carstea *et al*., 1997; Loftus *et al*., 1997). Although the animal proteins have been implicated in cholesterol/sterol homeostasis and bear the "sterol-sensing domain" also present in the cholesterol biosynthetic enzyme, 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMG-CoA reductase) (Loftus *et al*., 1997), the biochemical function of none of the former proteins is understood (Lange and Steck, 1998). Their putative topologies have recently been reported (Lange and Steck, 1998).

In this report we provide evidence that the above-mentioned disease- and development-related homologues are members of a large ubiquitous superfamily of permeases, all biochemically well characterized members of which are efflux pumps from Gram-negative bacteria. Different members of this superfamily, the RND superfamily (TC no. 2.6) (Saier *et al*., 1994; Saier, 1998), have been shown to individually transport hydrophobic drugs, fatty acids, bile salts, organic solvents, heavy metals, autoinducers and lipooligosaccharides in Gram-negative bacteria (see our Web site http://www-biology.ucsd.edu/~msaier/transport/titlepage.html). On the basis of observations herein reported and considerations of transport evolution that have resulted from the phylogenetic classification of transport proteins, we propose that at least some of the eukaryotic RND superfamily members functionally resemble the prokaryotic transporters. The disease- and differentiation-related animal proteins may play specific roles in cholesterol transport and steroid hormone reception, respectively. We additionally show that the SecDF protein complex of the bacterial type II general secretory apparatus (TC no. 3.5) (Saier *et al*., 1989), a family of Gram-positive bacterial proteins that includes the so-called ActII3 actinorhodin transport-associated protein (Férnandez-Moreno *et al*., 1991) and a novel cluster of archaeal and spirochete proteins are members of this superfamily.

The topological features of the proteins in each family are described, a phylogenetic tree for each family is presented, and family-specific signature sequences are derived from well-conserved regions of the complete multiple alignments of the proteins that comprise each family. All families include members that exhibit the

Received April 7, 1999; accepted April 7, 1999. *For correspondence. Email msaier@ucsd.edu; Tel. (619) 534-4084; Fax. (619) 534-7108.

Table 1. Representative Proteins of the RND Superfamily (TC no. 2.6)

Family	Abbreviation	Protein	Size (no. aas)	Putative TMSs	Source	Accession number
1.	CzcA Aeu	Co^{2+}-Zn^{2+}-Cd^{2+} exporter-A	1063	12	*Alkaligenes eutrophus*	gbM91650
2.	AcrF Eco	Acriflavin resistance protein-F	1034	12	*Escherichia coli*	gbX57948
2.	AcrB Eco	Acriflavin resistance protein-B	1049	13	*Escherichia coli*	spP31224
3.	NolGHI Rme	Lipooligosarccharide exporter	946 (277+215+435)	10	*Rhizobium meliloti*	gbX58632
4.	SecDF Eco	Type II protein secretion system constituents SecD and SecF	938 (615+323)	10	*Escherichia coli*	spP19673 and spP19674
4.	SecDF Rca	Type II protein secretion system constituents SecD and Sec	F887 (554+333)	10	*Rhodobacter capsulatus*	gbU69979
5.	ActII3 Sco	Actinorhodin transport-associated protein II-3	711	12	*Streptomyces coelicolor*	gbM64683
6.	YMP Sce	YPL006w membrane protein	1170	14	*Saccharomyces cerevisiae*	pirS52525
6.	NPC Hsa	Niemann-Pick C disease protein 1	1278	14	*Homo sapiens*	gbAF002020
6.	Ptc Dme	"Patched" morphogen (segmentation polarity) receptor	1286	12	*Drosophila melanogaster*	spP18502
6.	SCAP Cgr	SREBP cleavage-activating protein	1276	8	*Cricetulus griseus*	gbU67060
7.	ORF Afu	Gene AF1229 protein	750	11	*Archeoglobus fulgidus*	gbAE001019
7.	ORF Mja	Gene MJ1562 protein	388	6	*Methanococcus jannaschii*	pirA64495

prototypical RND superfamily topology. In the concluding section the structural, functional, mechanistic and evolutionary implications of the findings reported are elaborated.

Family Definition and Establishment of Homology

As demonstrated previously for a few representative Gram-negative bacterial RND family members (Saier *et al.*, 1994), proteins which are established members of the RND superfamily exhibit similar but unusual topological features, usually with 12 putative transmembrane α–helical spanners (TMSs) and large hydrophilic extracytoplasmic domains between TMSs 1 and 2 and TMSs 7 and 8 (Lange and Steck, 1998; Saier *et al.*, 1998). These proteins arose as a result of an internal gene duplication event (Saier *et al.*, 1994). These features will be shown below to be characteristic of eukaryotic, archaeal and Gram-positive bacterial proteins that respectively comprise distinct newly identified families in the ubiquitous RND superfamily.

The RND superfamily had previously been shown to include three Gram-negative bacterial-specific phylogenetic families of proton motive force (pmf)-driven efflux pumps, each with specificity for a different type of compound (Saier, 1994; Saier *et al.*, 1998). All of the members of these families probably function by substrate:proton antiport (Nies, 1995). Family 1 transporters export heavy metals (Rensing *et al.*, 1997), Family 2 permeases exhibit specificity for drugs and other hydrophobic compounds (Saier *et al.*, 1998), and Family 3 permeases probably secrete lipooligosaccharides that function as signaling molecules in rhizobial nodulation of leguminous plants (Baev *et al.*, 1991; Göttfert, 1993). Table 1 lists representative members of these three families as well as those of four novel families (Families 4-7) which we show are constituents of the RND superfamily. Family 4 includes numerous bacterial and archaeal SecD-SecF complexes that function together as auxiliary constituents of the universal (type II) secretory pathway (Saier *et al.*, 1989; Sugai and Wu, 1992; Bolhuis *et al.*, 1998). In some organisms, the SecD and SecF proteins are fused in a single polypeptide chain (Bolhuis *et al.*, 1998). Family 5 includes about twenty currently sequenced proteins, all from Gram-positive bacteria. One of these proteins, the so-called ActII3 actinorhodin transport-associated protein, has been implicated in the export of actinorhodin from *Streptomyces coelicolor* which synthesizes this antibiotic (Férnandez-Moreno *et al.*, 1991). Family 6 includes the human Niemann-Pick C disease protein, NPC, which appears to function in cholesterol homeostasis (Carstea *et al.*, 1997; Loftus *et al.*, 1997), an uncharacterized NPC homologue from the yeast *Saccharomyces cerevisiae*, YMP, that is about 30% identical to NPC throughout its length (Nelissen *et al.*, 1997), several homologues from higher animals including the "Patched" morphogen receptor which controls segmentation polarity in *Drosophila* (Hooper and Scott, 1989; Nakano *et al.*, 1989), and the sterol regulatory element binding protein (SREBP) cleavage-activating protein, SCAP, point mutations in which render Chinese hamster ovary (CHO) cells resistant to sterols (Hua *et al.*, 1996). Finally, Family 7 includes proteins from archaea and spirochetes, none of which is functionally characterized. The archaeal and spirochete proteins show greatest sequence similarity to proteins of Family 5, and

Table 2. Binary Sequence Comparisons which Establish that the Proteins Which Comprise the Seven RND Families Share a Common Ancestry[a]

Protein 1	RND Family	Protein 2	RND Family	No. Residues Compared	% Identity	Comparison Score (S.D.)
AcrF Eco	2	NolGHI Rme	3	141	37	18
CzcA Aeu	1	NolGHI Rme	3	368	26	33
SecDF Rca	4	NolGHI Rme	3	132	26	9.1
ActII3 Sco	5	SecDF Rca	4	130	22	9.4
YMP Sce	6	AcrB Eco	2	151	25	10
NPC Hsa	6	YMP Sce	6	1170	33	56
SCAP Cgr	6	NPC Hsa	6	189	29	18
Ptc Dme	6	NPC Hsa	6	1168	19	11
ORF Afu	7	ActII3 Sco	5	711	18	15

[a]The percent identity values and comparison scores were calculated using the GAP program with 500 random shuffles (Devereux *et al.*, 1984). A comparison score of 9 S.D. corresponds to a probability of 10^{-19} that the observed degree of similarity occurred by chance (Dayhoff *et al.*, 1983). This degree of similarity is deemed too great to have occurred by chance or by a convergent evolutionary process for compared regions of greater than 60 residues (Doolittle, 1986; Saier, 1994). Abbreviations are as described in Table 1.

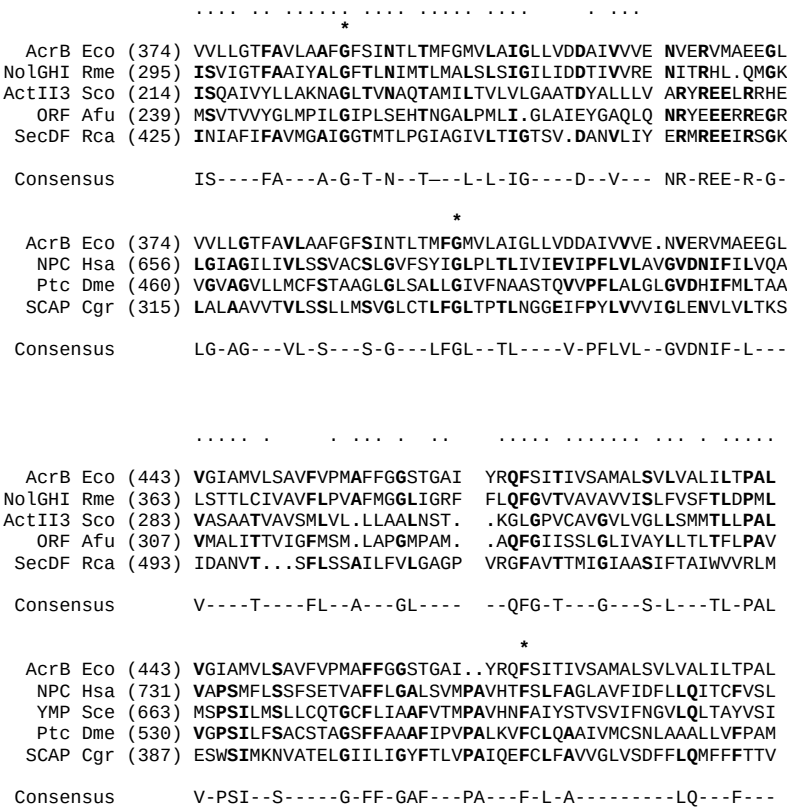

Figure 1. Multiple Alignments of RND Superfamily Members
Partial multiple alignments of four distant prokaryotic RND superfamily members with the AcrB protein of *E. coli* (top) and of three eukaryotic RND superfamily members with the same protein (bottom). Abbreviations of the proteins are presented in Table 1. Residues conserved in at least 3 of 5 proteins in each alignment are presented in bold print and appear in the consensus sequence. Fully conserved residues in each alignment are indicated with asterisks above the aligned sequences. Dots above the top alignment indicate that conservative substitutions in the majority of the residues in both alignments occur at a particular position. Residues considered to represent conservative substitutions are grouped for this purpose as follows: hydrophobic: L, I, V, M, F, Y, W; semipolar: A, S, T, C, G, P; basic: R, K, H; acidic and amide: D, E, N, Q. Sixty-six out of the 128 residue positions (52%) exhibit residue type conservation. Note that the two alignments are superimposable.

consequently they may function in drug efflux. Homologous archaeal proteins are either internally duplicated and of the expected size (*e.g.*, the gene AF1229 product of *Archeoglobus fulgidus*), or unduplicated and of about half the expected size (*e.g.*, the gene MJ1562 product of *Methanococcus jannaschii*). Similarly, the animal proteins are either internally duplicated (*e.g.*, YMP, NPC and Patched) or not duplicated (*e.g.*, SCAP and HMG CoA reductase) (Lange and Steck, 1998, see below). The nonduplicated proteins may function as homodimers as has been demonstrated for HMG-CoA reductase. It should be noted that with the exceptions of a few RND superfamily members such as the *M. jannaschii* protein of Family 7, and HMG-CoA reductase and the hamster SCAP protein, both of Family 6, all putative RND superfamily members are internally duplicated, exhibiting between 10 and 14 putative TMSs (Table 1).

Table 2 summarizes the statistical analyses of binary sequence comparisons which allow us to conclude that these proteins are all members of a single diverse superfamily. Families 1-3 had been shown previously to be evolutionarily related (Saier *et al.*, 1994), and the comparison scores summarized in Table 2 confirm this conclusion. The SecDF Rca protein complex gave a comparison score of 9.1 S.D. with the NolGHI protein complex (Table 2), a value which is sufficient to establish homology (see footnote to Table 2). As expected for two polypeptide chains that correspond to the two repeat elements within a single RND transporter, SecD and SecF are homologous to each other (Gardel *et al.*, 1990; Johnson *et al.*, 1992).

The *Streptomyces* ActII3 protein is demonstrably homologous to the *Rhodobacter* SecDF complex, showing that the Gram-positive bacterial family 5 is part of the RND superfamily. This family includes 10 paralogues from *Mycobacterium tuberculosis* (see below) and may therefore be of great importance to the drug resistance phenotype of this organism. The extraordinary degree of drug resistance exhibited by pathogenic Mycobacteria renders antibiotic treatment of this pathogen unusually problematic (Brennan and Nikaido, 1995).

The large and diverse eukaryotic-specific family 6 includes the yeast ORF, YMP Sce, which is clearly homologous to the *E. coli* AcrB drug efflux pump of RND family 1. This yeast protein exhibits a high degree of sequence identity with the Niemann-Pick C disease protein of humans as noted above and documented in Table 2. The involvement of this latter protein in cholesterol homeostasis (Johnson *et al.*, 1996; Loftus *et al.*, 1997) leads us to suggest that both NPC and YMP are sterol transporters. The hamster SCAP protein, mutations in which renders Chinese hamster ovary cells resistant to sterols, is clearly related to NPC Hsa as is the “Patched” segment polarity protein of *Drosophila melanogaster*. We suggest that these animal proteins generally function in sterol transport or reception (Lange and Steck, 1998). The proposed transport mechanism operative for YMP and NPC may be similar to those used by the well-characterized Gram-negative bacterial RND permeases (*i.e.*, pmf-driven solute efflux via proton antiport; Nies, 1995). This postulate should provide a guide to researchers concerned with the biochemical functions of the sequence similar yeast YMP and human NPC proteins.

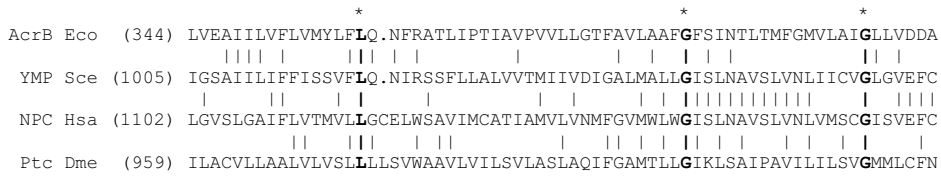

Figure 2. Alignment of Proteins of RND Family 6
Partial multiple alignment of three proteins of RND Family 6 (YMP Sce, NPC Hsa and Ptc Dme; see Table 1) with the *E. coli* AcrB drug efflux pump, an established member of RND Family 2. The comparison scores reported in Table 2 for the corresponding proteins were based on larger regions of compared sequences that included the aligned sequences shown. Vertical lines between sequences indicate identities while residues presented in bold print beneath asterisks are fully conserved in all four sequences. Residue numbers are provided in parentheses following the abbreviations of the proteins.

The partial multiple alignment shown in the top panel of Figure 1 reveals limited sequence similarity between representative members of all seven RND superfamily families. The proteins shown are ArcB of *E. coli* and NolGHI of *Rhizobium meliloti*, both established members of the RND family (Saier *et al.*, 1994), together with a representative of each one of the three new prokaryotic RND superfamily. Fully conserved residues are indicated with asterisks. The partial multiple alignment shown in the bottom panel of Figure 1 reveals the sequence similarity observed between different putative eukaryotic protein members of the RND superfamily as well as with AcrB of *E. coli*. While statistical means such as those reported in Table 2 are required to establish homology, significant sequence similarity including some fully conserved residues for each group are apparent. Figure 2 depicts a different partial alignment of the C-terminal domain of the eukaryotic proteins with the homologous portion of the N-terminal domain of AcrB. A striking degree of sequence similarity is apparent.

Families of RND Superfamily Homologues in Gram-negative Bacteria (TC no. 2.6.1-2.6.3)

As noted above, we previously identified three distant groups of RND superfamily members (Saier *et al.*, 1994), and all proteins then identified were from Gram-negative bacteria. These proteins included (1) heavy metal efflux pumps such as CzcA of *Alkaligenes eutrophus* strain C34 (now *Ralstonia* spec. CH34), (2) the AcrF multidrug efflux pump of *E. coli*, and (3) the NolGHI putative lipooligosaccharide extrusion system of *R. meliloti*. These three proteins are now prototypical members of three phylogenetically and functionally defined families within the RND

Table 3. RND Superfamily Proteins that Cluster with Three Previously Identified Families (the Heavy Metal Efflux [HME] Family [TC no. 2.6.1], the Hydrophobe/Amphiphile Efflux-1 [HAE1] Family [TC no. 2.6.2], and the Nodulation Factor Exporter [NFE] Family [TC no. 2.6.3] (Mostly from Gram-negative Bacteria)

Abbreviation	Protein Designation in Databases	Size	Source	Accession no.
IfeB Atu	Transmembrane efflux pump protein	1046	*Agrobacterium tumefaciens*	gbAF039653
NccA Axy	NccA protein	1076	*Alcaligenes xylosoxydans*	pirI39580
Orf1 Aae	Cation efflux	1000	*Aquifex aeolicus*	gbAE000707
Orf2 Aae	Cation efflux system	1019	*Aquifex aeolicus*	gbAE000691
Orf3 Aae	Cation efflux system	1050	*Aquifex aeolicus*	gbAE000724
Orf4 Aae	Cation efflux system	1082	*Aquifex aeolicus*	gbAE000702
Orf Bsu	Putative acriflavin resistance protein	1065	*Bacillus subtilis*	gbZ99107
Orf Bbu	Putative acriflavin resistance protein	1036	*Borrelia burgdorferi*	gbAE001125
RagC Bja	Putative cation efflux or multidrug resistance protein	1060	*Bradyrhizobium japonicum*	gbAJ225023
CeoB Bce	CeoB	1027	*Burkholderia cepacia*	gbU97042
AmrB Bps	Putative transporter AmrB	1043	*Burkholderia pseudomallei*	gbAF072887
YegO Eco	Hypothetical 111.0 kd protein in *alkA-baeS* intergenic region	1025	*Escherichia coli*	spP76399
AcrF Eco	Acriflavin resistance protein F	1034	*Escherichia coli*	spP24181
YhiV Eco	Hypothetical 111.5 kd protein in *hdeD-gadA* intergenic region	1037	*Escherichia coli*	spP37637
AcrD Eco	Acriflavin resistance protein D	1037	*Escherichia coli*	spP24177
YegN Eco	Hypothetical 112.1 kd protein in *alkA-baeS* intergenic region	1040	*Escherichia coli*	spP76398
YbdE Eco	Hypothetical 114.7 kd protein in *nfrB-pheP* intergenic region	1047	*Escherichia coli*	spP38054
AcrB Eco	Acriflavin resistance protein B	1049	*Escherichia coli*	spP31224
Orf Hin	Hypothetical protein HI0895	1032	*Haemophilus influenzae*	spQ57124
Orf1 Hpy	Putative acriflavin resistance protein (AcrB)	1028	*Helicobacter pylori*	gbAE000574
Orf2 Hpy	Putative cation efflux system protein (CzcA)	1020	*Helicobacter pylori*	gbAE000605
Orf3 Hpy	Putative cation efflux system protein (CzcA)	1035	*Helicobacter pylori*	gbAE000634
HelA Lpn	HelA protein	1052	*Legionella pneumophila*	spQ48815
MtrD Ngo	Efflux transporter membrane protein	1067	*Neisseria gonorrhoeae*	gbU60099
MexD Pae	Cytoplasmic membrane-associated proton-motive-force-driven efflux transporter component of a multidrug resistance efflux pump	1043	*Pseudomonas aeruginosa*	gbU57969
MexB Pae	Multidrug resistance protein MexB	1046	*Pseudomonas aeruginosa*	spP52002
MexF Pae	Cytoplasmic membrane component of multidrug efflux system	1062	*Pseudomonas aeruginosa*	gbX99514
Orf Ppu	Inner membrane transporter protein	1049	*Pseudomonas putida*	gbAF029405
TtgB Ppu	Putative efflux pump TtgB	1050	*Pseudomonas putida*	gbAF031417
CzcA Reu	Cation efflux system protein CzcA	1063	*Ralstonia eutropha*	spP13511
CnrA Reu	Nickel and cobalt resistance protein CnrA	1075	*Ralstonia eutropha*	spP37972
NolGHI Rme	Lipooligosarccharide exporter	946	*Rhizobium meliloti*	gbX58632
Orf1 Ssp	Acriflavine resistance protein	1061	*Synechocystis sp*	gbD90911
Orf2 Ssp	Acriflavin resistance protein	1083	*Synechocystis sp.*	gbD63999
Orf3 Ssp	Putative acriflavin resistance protein	909	*Synechocystis sp*	gbD90917
Orf4 Ssp	Putative cation efflux system protein	1054	*Synechocystis sp*	gbD64005
Orf5 Ssp	Cation efflux system protein CzcA	1075	*Synechocystis sp*	gbD90915

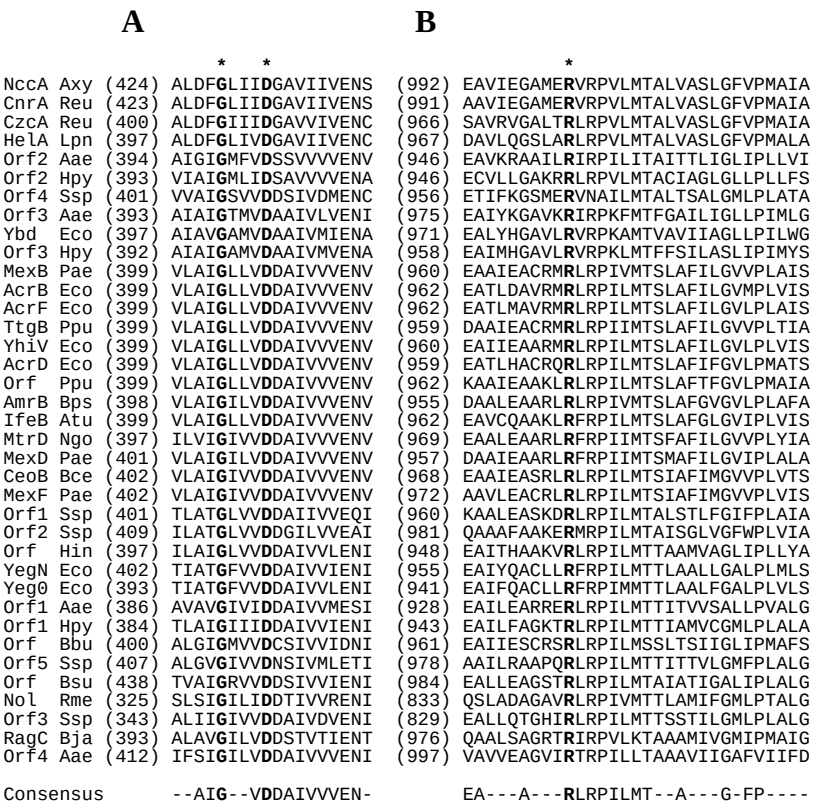

Figure 3. Alignment of Proteins from the HME, HAE1, and NFE Families
Two well-conserved, gap-free portions of the complete multiple alignment of thirty-seven sequenced proteins encompassing and closely related to families HME (TC no. 2.6.1), HAE1 (TC no. 2.6.2) and NFE (TC no. 2.6.3) of the RND superfamily. Abbreviations of the proteins are as indicated in Table 3, and the residue number of the first residue in each sequence shown is presented in parentheses. Fully conserved residues are presented in bold print with an asterisk above the alignment. The consensus sequence (Consensus) is provided below the alignment. A residue appears in the consensus sequence if that residue appears in a majority (19 of 37) of the proteins shown in the multiple alignment. The TREE program of Feng and Doolittle (1990) was used to generate the alignment (see Young *et al.*, 1999).

superfamily as follows: (1) the heavy metal efflux (HME) family (TC no. 2.6.1), (2) the (largely Gram-negative bacterial) hydrophobe/amphiphile efflux-1 (HAE1) family (TC no. 2.6.2), and (3) the putative nodulation factor exporter (NFE) family (TC no. 2.6.3). The HAE1 family in particular has expanded and now includes several functionally characterized proteins which catalyze export of a large variety of drugs, aliphatic and aromatic organic solvents, bile salts, fatty acids, isoflavenoids, and autoinducers (see our web page). All of these porters exhibit broad specificity for hydrophobic and amphiphilic compounds, and every functionally characterized member of the RND superfamily functions with outwardly directed polarity.

Table 3 lists all of the nearest RND superfamily homologues retrieved from the databases using the BLAST search tool, using any one of the three prototypical proteins cited above. Several points are worthy of note. First, all but one of the 37 proteins listed are from Gram-negative bacteria. The one exception, a protein from *Bacillus subtilis*, may have been obtained by *Bacillus* by horizontal transmission of genetic material. None of the proteins listed is from an archaeon or a eukaryote. Second, several bacteria exhibit multiple paralogues. For example, the *E. coli* genome encodes 7 paralogues, *Synechocystis* encodes 5 paralogues, *Aquifex aeolicus* encodes 4 paralogues, and *Helicobacter pylori* encodes 3. Third, a few fully sequenced Gram-negative bacterial genomes do not encode recognizable members of these families. For example, the genomes of *Treponema pallidium* and *Chlamydia trachomatis* fall into this category. Fourth, with only three exceptions, the proteins exhibit a very narrow size range varying from 1000 to 1083 amino acyl residues in length. The three exceptions are only somewhat smaller (909-964 residues). It should be noted that imprecise assignment of initiation codon may have given rise to incorrect estimation of protein length. Fifth, in spite of the large number of new RND superfamily homologues listed in Table 3, none of the functionally characterized proteins exhibits substrate specificities that differ markedly from those described in 1994 (Saier *et al.*, 1994).

Figure 3 presents two gap free regions of the complete multiple alignment that exhibit high degrees of sequence similarity. In addition to the few fully conserved residue positions, several positions are conserved in all but one or two of the aligned sequences. From each of these two alignments, signature sequences were derived. These sequences are:

1. (L I V T A S) (L I V F) X (L I V F T) G X (L I V F M) (L I V) D (D S A G) (S A G T) (L I V T) (L I V) X (L I V M R) (E D) (N Q A S T)

2. (S A T C) (L I V A T) X_2 (G A S T) X_3 R (L I V F M T) (R N) (P A) X (L I V F M A) X (T S) (S A T V F) X_5 (G A) (L I V F M A) (L I V F M W) (P V) (L I V M T)

(X = any residue; amino acids in parentheses represent alternative residues at any one position)

Figure 4A presents the average hydropathy plot obtained from the multiple alignment of the 38 sequences corresponding to the proteins listed in Table 3. The plot is diagnostic of the proteins of the RND superfamily in general (see below). Thus, a single N-terminal hydrophobic peak corresponds to putative TMS 1, and this is followed by a

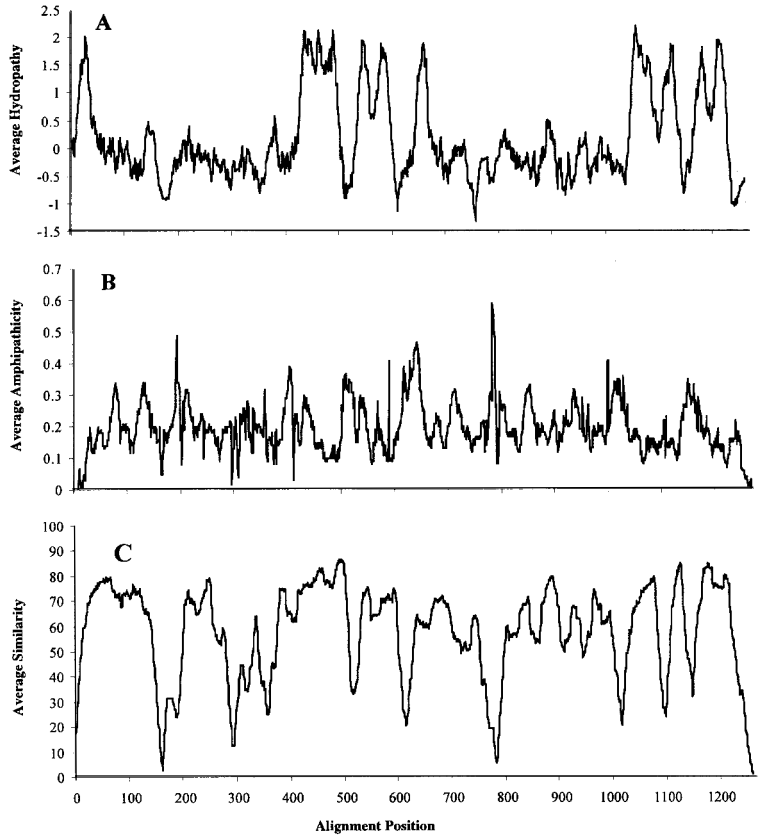

Figure 4. Hydropathy, Similarity, and Amphipathicity Plots
Average hydropathy (A), average amphipathicity (B) and average similarity (C) plots for the 37 proteins included in Figure 3. In all plots, a sliding window of 21 residues was used. The hydropathy plot was based on the amino acid hydropathy values of Kyte and Doolittle (1982). For the average amphipathicity plot, an angle of 100° was used as required for an (helix in the TREE moment program (Le *et al.*, 1999). The partial multiple alignments shown in Figures 3A and B were taken from the complete multiple alignment upon which the plots shown were derived.

long hydrophilic, periplasmically localized region (alignment positions 50-430). Residue positions 430-600 exhibit five peaks of hydrophobicity in a 3 + 2 arrangement corresponding to TMSs 2-4 and 5 + 6, respectively. This first 6 TMS segment (alignment positions 1-600) corresponds to the first tandem repeat unit which is characteristic of all RND superfamily permeases. The second repeat unit begins at residue position 601. The hydrophobic peak at position 650 (putative TMS 7), the hydrophilic, periplasmically-localized domain at positions 620-1040, and the C-terminal five hydrophobic peaks in a 3 + 2 arrangement, corresponding to putative TMSs 8-10 and 11 + 12, respectively, correspond to the second repeat unit in these permeases. Because the hydropathy plots of so many protein sequences could be averaged, the topological predictions can be considered to be highly reliable. The topology of a few RND superfamily member has been verified and corresponds to prediction (T. Pribyl and D.H. Nies, unpublished results).

The average amphipathicity plot (angle set at 100° for an α-helix, Figure 4B) reveals that the largest peaks of amphipathicity occur immediately before putative TMSs 2 and 8 as well as between TMSs 4 and 5, and TMSs 10 and 11, in equivalent positions in the two repeat elements. However, an additional peak is observed preceding putative TMS 7 although no corresponding peak precedes putative TMS 1. As will be noted below, this topological arrangement is also observed for other families within the RND superfamily. The putative amphipathic α-helices that precede hydrophobic segments in the RND superfamily proteins may have biogenic significance (Saier *et al.*, 1989; Yamada *et al.*, 1991; Le *et al.*, 1999).

An average similarity plot (Figure 4C) revealed that the Gram-negative bacterial proteins exhibit comparable degrees of sequence similarity throughout most of their lengths although troughs, corresponding primarily to gaps in the multiple alignment, occur in hydrophilic regions of the proteins. All twelve putative TMSs are well-conserved as are the putative amphipathic helices that precede the hydrophobic segments of the proteins. This fact emphasizes the potential functional, structural or biogenic significance of these structures.

The phylogenetic tree for 37 proteins listed in Table 3 is shown in Figure 5. As noted in our previous publication where only a few proteins were represented, the known heavy metal efflux pumps (CzcA Reu, CnrA Reu and NccA Axy), the multidrug pumps (*i.e.*, the Acr proteins of *E. coli*, the Mex proteins of *P. aeruginosa* and others) and the putative NolGHI nodulation factor expulsion system can each be found clustered at the end of a distinct branch. However, in contrast to our previous results, many new proteins are found to branch from points near the center of the unrooted phylogenetic tree. Because of their considerable sequence divergence and the fact that no functionally characterized protein clusters with them, it is not possible to predict the substrate specificities of these putative transporters. Nevertheless, since all functionally characterized members of the RND superfamily appear to function as efflux pumps using a proton antiport mechanism, we predict that all transporters of the RND superfamily will prove to exhibit outwardly directed polarity.

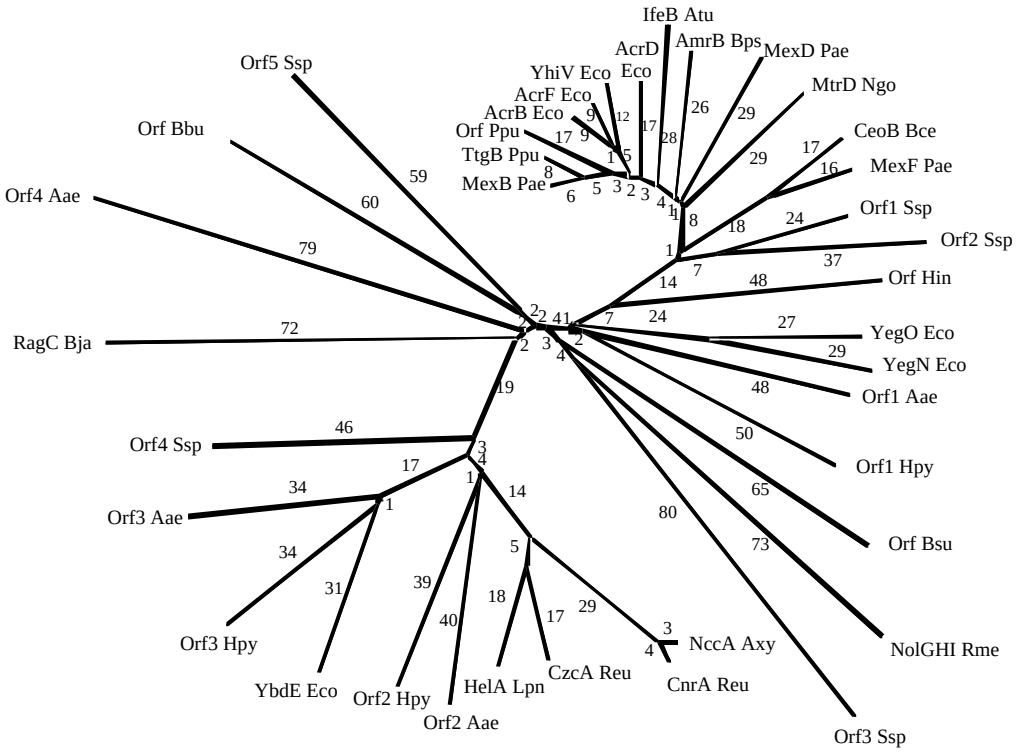

Figure 5. Phylogenetic Tree
Phylogenetic tree for the proteins included in Figure 3. Protein abbreviations are given in Table 3. Branch lengths, in arbitrary units, are provided adjacent to the branches and are approximately proportional to phylogenetic distance. The TREE program of Feng and Doolittle (1990) was used to generate the tree.

The SecDF Family (TC no. 2.6.4)

Proteins of the SecDF family proved to be homologous to previously defined, well-characterized members of the RND superfamily (Gardel et al., 1987, 1990; Pogliano and Beckwith, 1994; Bolhuis *et al.*, 1998; see Table 2 above). While the precise biochemical functions of these proteins are not known, they are found associated with the general secretory (Sec; type II) protein complexes of bacteria and archaea in substoichiometric amounts where they function as essential constituents of the Sec system at low temperatures (Pogliano and Beckwith, 1994; for a current review see Economou, 1998). Defects in either of these proteins thus give a cold-sensitive phenotype (Gardel *et al.*, 1987, 1990; Schatz *et al.*, 1991; Sugai and Wu, 1992; Bolhuis *et al.*, 1998).

The proteins of the SecDF family form a coherent group. For all members of the family from organisms with completely sequenced genomes, either both SecD and SecF are present (Table 4) or neither protein is present. Thus, for example, both SecD and SecF homologues could be found in the completely sequenced genomes of 12 of the 15 fully sequenced bacterial genomes, but neither protein was found encoded within the *Mycoplasma genitalium* or *M. pneumoniae* genome. These latter two organisms do exhibit an abbreviated Sec system including the SecY and SecA proteins (Fraser *et al.*, 1995; Himmelreich *et al.*, 1996). Similarly, although both SecD and SecF were found encoded within the genomes of three archaea, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii* and *Pyrococcus horikoschii*, neither protein was identified in the fully sequenced genome of *Archaeoglobus fulgidus* (Klenk *et al.*, 1997). Like the mycoplasmas, *A. fulgidus* does have a type II secretory system that includes the SecY and SecE proteins (Klenk *et al.*, 1997). Interestingly, the *A. fulgidus* Sec system seems to resemble eukaryotic Sec systems (Klenk *et al.*, 1997), and *Saccharomyces cerevisiae* appears to lack genes encoding SecD and SecF homologues (Table 4). These observations suggest that SecD and SecF may not be required for secretion via the Sec system when growth rates are slow or growth occurs at very high temperatures. They further suggest that these proteins are not constituents of the secretory apparatuses of some archaea and eukaryotes.

As indicated in Table 4, SecD and SecF occur as two distinct polypeptide chains encoded by two distinct genes in all bacteria represented except *Bacillus subtilis* and *Chlamydia trachomatis*. In these two organisms, SecD and SecF are fused in a single gene-encoded polypeptide chain. Surprisingly, the chlamydial protein is nearly twice as large as the *B. subtilis* protein (1400 amino acyl residues as compared with 737 residues). The former protein includes a substantial amount of nonhomologous material. The sizes of the SecD proteins vary from 396 residues (*M. jannaschii*) to 701 residues (*Mycobacterium leprae*) while the SecF proteins vary from 251 residues (*M. thermoautotrophicum*) to 442 residues (*M. tuberculosis*). Thus, size variation is a characteristic of the SecDF family.

Partial multiple alignments of the SecD and SecF proteins from completely sequenced genomes are shown in Figures 6A and 6B, respectively. Within the regions shown, both alignments exhibit fully conserved residues. Signature sequences were derived from these alignments as follows:

Table 4. Proteins of the SecDF Type II Secretory System Auxiliary Constituents (SecDF) Family (TC no. 2.6.4) (from Bacteria and Archaea)

Abbreviation	Description	Organism	Size (no. of residues)	Database and accession no.
SecDF Bsu	SecDF protein	*Bacillus subtilis*	737	gbAF024506
SecD Eco	SecD protein	*Escherichia coli*	615	spP19673
SecF Eco	SecF protein	*Escherichia coli*	323	spP19674
SecD Rca	SecD protein	*Rhodobacter capsulatus*	554	gbU69979
SecF Rca	SecF protein	*Rhodobacter capsulatus*	333	gbU69979
SecD Hpy	SecD protein	*Helicobacter pylori*	503	gbAE000652
SecF Hpy	SecF protein	*Helicobacter pylori*	323	gbAE000652
SecD Aae	SecD protein	*Aquifex aeolicus*	501	gbAE000716
SecF Aae	SecF protein	*Aquifex aeolicus*	288	gbAE000747
SecD Hin	SecD protein	*Haemophilus influenzae*	616	spP44591
SecF Hin	SecF protein	*Haemophilus influenzae*	325	spP44590
SecD Tpa	SecD protein	*Treponema pallidum*	583	gbAE001219
SecF Tpa	SecF protein	*Treponema pallidum*	420	gbAE001219
SecD Ssp	SecD protein	*Synechocystis sp.*	472	gbD64000
SecF Ssp	SecF protein	*Synechocystis sp.*	315	gbD64000
SecD Bbu	SecD protein	*Borrelia burgdorferi*	586	gbAE001166
SecF Bbu	SecF protein	*Borrelia burgdorferi*	299	gbAE001166
SecD Mtu	SecD protein	*Mycobacterium tuberculosis*	573	spQ50634
SecF Mtu	SecF protein	*Mycobacterium tuberculosis*	442	spQ50635
SecD Mle	SecD protein	*Mycobacterium leprae*	701	spP38387
SecF Mle	SecF protein	*Mycobacterium leprae*	394	spP38386
SecD Sco	SecD protein	*Streptomyces coelicolor*	570	spQ53955
SecF Sco	SecF protein	*Streptomyces coelicolor*	373	spQ53956
SecD Mth	SecD protein	*Methanobacterium thermoautotrophicum*	403	gbAE000861
SecF Mth	SecF protein	*Methanobacterium thermoautotrophicum*	257	gbAE000861
SecD Mja	SecD protein	*Methanococcus jannaschii*	396	spQ57575
SecF Mja	SecF protein	*Methanococcus jannaschii*	282	pirD64456
SecD Pho	SecD protein	*Pyrococcus horikoshii*	507	gbAP000007
SecF Pho	SecF protein	*Pyrococcus horikoshii*	293	gbAP000007
SecDF Ctr	SecDF protein	*Chlamydia trachomatis*	1400	gbAE001318
Orf3 Taq	Hypo. Protein 3	*Thermus aquaticus*	275	pirS52278
SecF Rsp	SecF protein	*Rhodobacter sphaeroides*	324	gbU83136
SecF Pde	SecF protein	*Paracoccus denitrificans*	183	gbZ71971

A. SecD

```
                    *                                *         *                *
SecD Bbu   (377)  EAQDLALVFKTAAFPVDIKIDDLRIIGPTLGARTIDLGIKASALALCLVFLFICVYYG
SecD Aae   (306)  EARDLALILRTGSLPSPLKFLQEKIVGPSLGKDAIEQGIKAGILAIILLAVVLIARYK
SecD Eco   (416)  EARQLSLLLRAGALIAPIQIVEERTIGPTLGMQNIEQGLEACLAGLLVSILFMIIFYK
SecD Hin   (417)  EAHNLSTLLKSGALIAPIQIVEERTIGPSLGAQNVEQGINASLWGLVAVIAFMLFYYK
SecD Hpy   (296)  QASDLAIALRSGAMSAPIQVLEKRIIGPSLGKDSVKTSIIALVGGFILVMGFMVLYYS
SecD Bsu   (221)  EAKDLASILNAGALPVKLTEKYSTSVGAQFGQQALHDTVFAGIVGIAIIFLFMLFYYR
SecD Mle   (260)  TARQLANVLKYGSLPLSFEPSEAQTVSATLGLTSLRAGLIAGAIGLSLVLLYSLLYYR
SecD Mja   (211)  EAMAIYSALKSGALPVKLDIEYISTISPEFGKEFLKGTAIALLLAFIAVGIIVSIRYK
SecD Mtu   (349)  TARQLANVLKYGSLPLSFEPSEAQTVSATLGLSSLRAGMIAGAIGLLLVLVYSLLYYR
SecD Pho   (310)  DAQVVAVVLRSGSLPIKLSIERIDYISPKLGENFKKQVLIAGIAALLVVGGIVYLHYR
SecD Rca   (356)  EATDLALLLRAGALPAGMTFLEERTIGPELGADSVKAGMVASVIGFVAVVAYMIASYG
SecD Sco   (334)  EAQSLANMLSYGALPLTFKEDSVTTVTAALGGEQLKAGLIAGAIGLALVVLYLLFYYR
SecD Ssp   (260)  TANDLAVQLRGGSLPFPVEVVENRTVGATLGQESIRRSLVAGFVGLVLVLVFMAVYYR
SecD Tpa   (378)  EAQNLKTALRSAWLNVALEIENQQVVGASMGEESIRQGTRALVWGLCAVLLFMLVWYQ
SecD Mth   (205)  QAKEIETLLKSGSLPVKVKIVGVSSVSPELGKQFAEGAVIAGLLAVLAIAVILIVRYR

Consensus         EA--LA--L--GALP---------TVGP-LG------G--AG--GL--V------YY-
```

B. SecF

```
                      *                          * *
SecF Bbu   (141)  SIFHDIFFIVAFLGVF    RIEINSYIIVAILTIIGYSLNDTIIIFDRIRDNVKRLT
SecF Aae   (153)  ALAHDVITVLGAYSIT    QREVNLEVVSAILVVAGYSVADTVVIFDRIRENLRKKK
SecF Eco   (176)  ALAHDVIITLGILSLF    HIEIDLTIVASLMSVIGYSLNDSIVVSDRIRENFRKIR
SecF Hin   (182)  SLAHDVIITLGVFSAL    QIEIDLTFVAAILSVVGYSINDSIVVFDRVRENFRKIR
SecF Hpy   (167)  ALVHDVILVASSVIVF    KIDMNLEVIAALLTLIGYSINDTIIIFDRIREEMLSQK
SecF Bsu   (600)  SLLYDAFFIVTFFSIT    RLEVDVTFIAAILTIIGYSINDTIVTFDRVREHMKKRK
SecF Mle   (168)  TMCFDLTVTAGVYSLV    GFEVTPATVIGLLTILGFSLYDTVIVFDKVEENTHGFQ
SecF Mja   (152)  SALSDIIMALGAMSLL    GIELSSATIAALLMVIGYSVDSDILLTTRVLKRLTKSF
SecF Mtu   (221)  AMLFDLTVTAGVYSLV    GFEVTPATVIGLLTILGFSLYDTVIVFDKVEENTHGFQ
SecF Pho   (169)  SAFSDMVIAVALMDIF    GIELSQATIAALLMLIGYSVDSNILLTTRLLRRKEFSV
SecF Rca   (185)  ALVHDVLLTVGLFAVL    QLKFDLTTVAALLTITGYSINDTVVVFDRLRENLIKYK
SecF Sco   (171)  ALIHDITITVGIYALV    GFEVTPGTVIGLLTILGYSLYDTVVVFDSLKEQTRDIT
SecF Ssp   (171)  ALLYDALITMGAFAIFGLVGGVEVDSLFLVALLTIIGFSVNDTVVIYDRVRETLERHS
SecF Tpa   (285)  ALVHDACIMVSFMVWF    GLEFNSASIAAILTIIGYSINDTVVVFDRVRQTILLDP
SecF Mth   (131)  AAASDIIIAVGGMSLF    GIPLSLASVGAILMLIGYSVDTDILLTTRVLKRRKGTI

Consensus         AL-HD-----G--S------G-E-----VAALLTIIGYS-NDT-V-FDRVRE------
```

Figure 6. Two Well-Conserved Portions of the Complete Multiple Alignment of the 15 Complete SecD-SecF Protein Systems Presented in Table 4
A, a portion of the SecD constituents. B, a portion of the SecF constituents. The incomplete *T. aquaticus*, *R. sphaeroides* and *P. denitrificans* sequences (see bottom of Table 4) were not included in the alignment. The SecDF proteins of *B. subtilis* and *C. trachomatis* are single polypeptide chains, but in all other pairs of SecD-SecF proteins, the SecD proteins were artificially fused N-terminally to the SecF proteins before the complete multiple alignment was generated. The format of presentation and computer program used were as described in the legend to Figure 3.

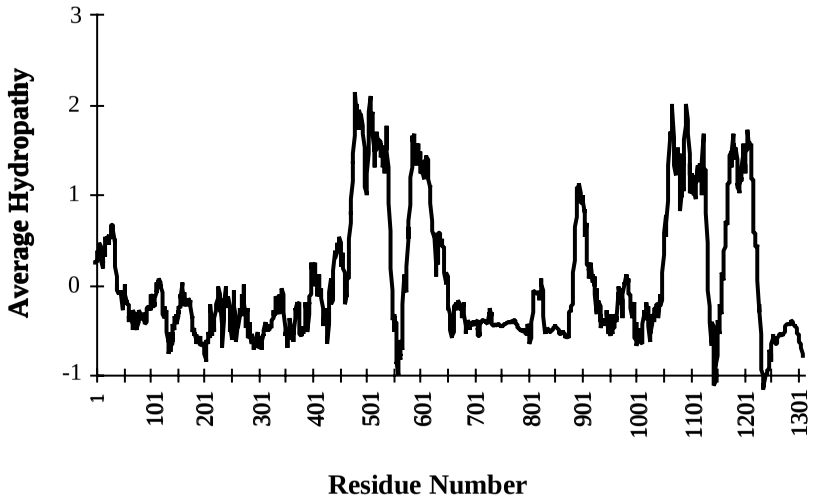

Figure 7. Average Hydropathy Plot for the (Fused) SecDF Proteins
The format of presentation and computer program used were as described in the legend to Figure 4.

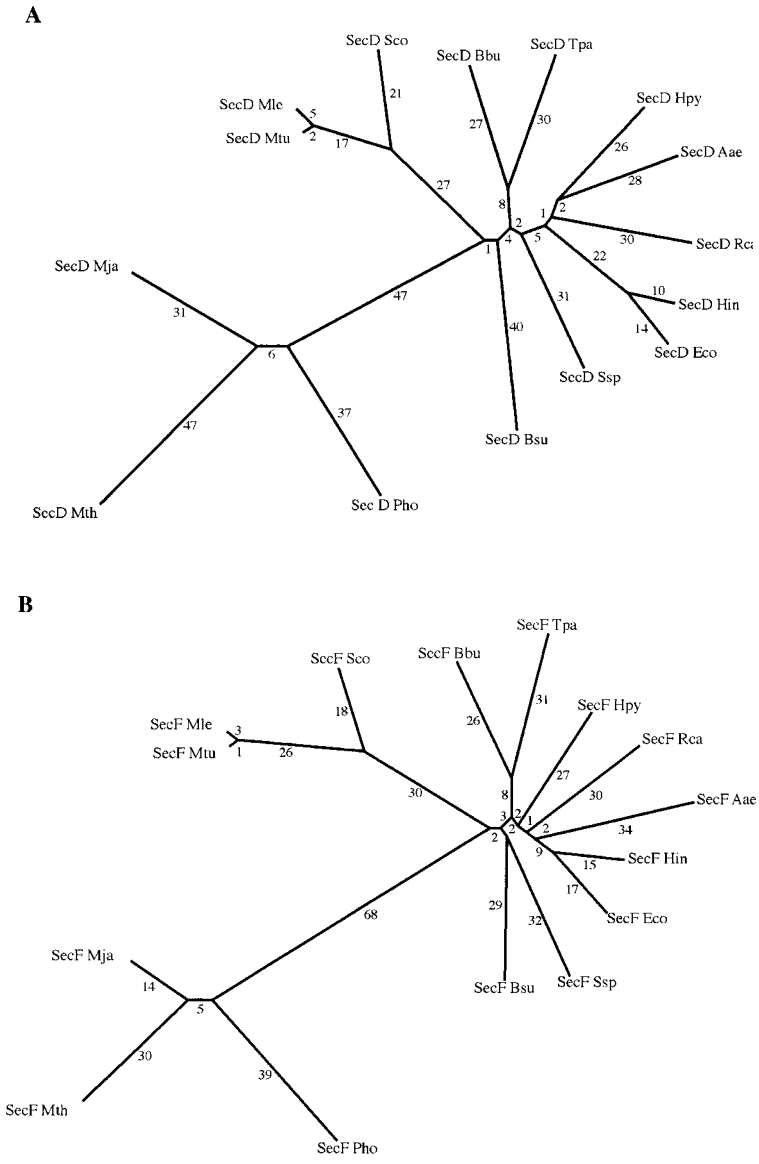

Figure 8. Phylogenetic Trees for the SecD Proteins (A) and the SecF Proteins (B) of the SecDF Family (TC no. 2.6.4)
The format of presentation and computer program used were as described in the legend to Figure 5.

SecD: (L I V) (G S T) (P A) X_2 G X_9 A X (L I V A F) (L I V A G W) (A G) (L I V F) (L I V C A S) (L I V A) (L I V S) X (L I V G A) (L I V F Y) (L I V M S) (L I V C S A Y) (L I V A F) X Y

SecF: (L I V) (L I V G A S) (G A S) (L I V) (L M) X (L I V) X G (Y F) S (L I V) X (D S T) (D N S T) $(L I V)_2$ (L I V T) (F T Y) (D T) (R K S) (L I V)

(X = any residue; residues in parentheses represent alternative residues at a particular position)

Based on the complete multiple alignments for the naturally fused SecDF proteins and the artificially fused SecD and SecF proteins, the average hydropathy plot shown in Figure 7 was obtained. The plot is typical of RND permeases with a single hydrophobic peak at the N-terminus probably corresponding to the first transmembrane helical spanner (TMS 1) followed by a large periplasmic, hydrophilic region (residue positions 50-470). This hydrophilic region is followed by a strongly hydrophobic region exhibiting 5 TMSs in a probable 3 + 2 arrangement (positions 470-650). SecD is known to have 6 TMSs in full accord with the topological prediction based on the average hydropathy analyses (Pogliano and Beckwith, 1994). A single hydrophobic peak appears at position 900 within a large hydrophilic region (residue positions 650-1050). This peak presumably corresponds to TMS 7 in the SecDF complex or to TMS 1 in SecF. Finally, the C-terminal region of the alignment again shows a high degree of hydrophobicity with 5 putative TMSs in a 3 + 2 arrangement (positions 1050-1230). These results are consistent with the 6 TMS topology suggested for the *E. coli* SecF (Johnson *et al.*, 1992; Pogliano and Beckwith, 1994) as well as the 12 TMS topology suggested for several Gram-negative bacterial RND family members (Saier *et al.*, 1994). The patterns of the left and right sides of the hydropathy plot reflect the homology of SecD with SecF in agreement with the proposed internal duplication observed for many members of the RND superfamily (Pogliano and Beckwith, 1994; Saier *et al.*, 1994; Bolhuis *et al.*, 1998).

An average similarity plot (not shown) revealed that as is often typical of RND superfamily members, the transmembrane regions are best conserved. However, two regions just N-terminal to both large hydrophobic domains are also well conserved. These hydrophilic regions of strong conservation proved to be amphiphilic and may provide an important function in the periplasm. An average amphipathicity plot (not shown) revealed that all regions of strong hydrophobicity exhibit lower amphipathicity values than the hydrophilic regions when the angle is set at 100° per residue as is appropriate for an α-helix.

Figures 8A and B show phylogenetic trees for the SecD and SecF proteins, respectively. For this analysis, the *B. subtilis* and *C. trachomatis* proteins were artificially spliced at the junction corresponding to the natural splice sites for other SecD and SecF proteins. The two trees are strikingly similar. In fact, the only minor difference in branching position is inversion of the positions of the *R. capsulatus* and *A. aeolicus* proteins. In general, clustering patterns are as expected on the basis of the phylogenies of the organisms. Thus, clustering is seen for (1) the three archaea (Mja, Mth and Pho), (2) the three high G + C Gram-positive bacteria (Mtu, Mle and Sco), (3) the two spirochetes (Bbu and Tpa), and (4) *E. coli* and *H. influenzae*. The lack of clustering of the *B. subtilis* protein with those of the high G + C Gram-positive bacteria was unexpected. Nevertheless, the observations presented in Figure 8 generally suggest that all of the SecD/SecF proteins are orthologues with a single pair of these proteins existing in each of the organisms analyzed.

The Gram-positive Bacterial Putative Hydrophobe/ Amphiphile Efflux-2 (HAE2) Family (TC no. 2.6.5)

As summarized above and in Table 2, the HAE2 family, with members exclusively from Gram-positive bacteria, represents a novel family within the RND superfamily. The sequenced proteins that are included in this family are listed in Table 5. Almost all members of the family are from high G + C Gram-positive bacteria (*Mycobacterium leprae* and *M. tuberculosis* as well as *Streptomyces coelicolor*) or from *B. subtilis*, an organism with about 43.5% G + C content. Thus, none of the several low G + C Gram-positive bacteria, for which completely sequenced genomes are available, are represented. Surprisingly, ten of the 18 proteins listed are from a single organism, *M. tuberculosis*. If these putative efflux pumps prove to be specific for antibiotics and other drugs, they may provide a partial explanation for the relative insensitivity of this organism to drug treatment (Brennan and Nikaido, 1995).

Only a single member of the HAE2 family has been partially characterized. This is the ActII3 protein of *S. coelicolor* which was suggested to be an “Actinorhodin transport-associated protein” (Férnandez-Moreno *et al.*, 1991). We here propose that it does not merely facilitate transport via a nonhomologous system as suggested by Férnandez-Moreno *et al.*, but that it instead is an independent efflux pump that functions as do the RND drug exporters from Gram-negative bacteria. It should be noted, however, that SecDF complexes somehow function as protein auxiliary constituents to the type II protein secretion system, and consequently a secondary transport role for both HAE2 and SecDF family members should not be ruled out without more substantial evidence. The proteins of the HAE2 family vary in size between 711 and 1146 amino acyl residues (Table 5). Even more size variation was noted for the SecDF family (see above).

Figure 9 shows a partial multiple alignment of a large, well conserved region of these proteins. The alignment reveals that only four residues are fully conserved within the segment of the multiple alignment shown, and examination of the consensus sequence suggests that the regions surrounding these two pairs of fully conserved residues are the best conserved. Surprisingly, there is only one 1-amino acid gap in the entire alignment shown. From this alignment, two signature sequences specific for the HAE2 family were derived. These sequences are:

1. (G A) (L I V A) (G A S) (T L I V) D Y (G A S C I) (L I V) (F L I V) (L I V M) (L I V T F) (S T A G M) (R K) (Y F H) (R H Q) (D E Q)

2. (L I V F A M) (L I V F A) (L I V S A T G) (A S L M) (L I V M A F) T (L I V F) (L I V G A T M) P (A L) (L I V F C A) (L I V M) X (L I V T) (L I V F T A G)

Table 5. Sequenced Proteins of the (Gram-positive Bacterial) Putative Hydrophobe/Amphiphile Efflux-2 (HAE2) family (TC no. 2.6.5)

Abbreviation	Description	Organism	Size (no. residues)	Database and accession no.
ydgH Bsu	Probable transport protein	*Bacillus subtilis*	885	gbAB001488
ydfJ Bsu	Possible transport protein	*Bacillus subtilis*	724	gbAB001488
Orf1 Mle	Transport protein	*Mycobacterium leprae*	1008	pirS72698
Orf2 Mle	Integral membrane protein	*Mycobacterium leprae*	959	spP54881
Orf3 Mle	Possible membrane transport protein	*Mycobacterium leprae*	955	gbZ95398
Orf4 Mle	Possible membrane transport protein	*Mycobacterium leprae*	1014	gbZ95398
Orf1 Mtu	Probable transmembrane protein	*Mycobacterium tuberculosis*	1002	gbAL010186
Orf2 Mtu	Possible membrane protein	*Mycobacterium tuberculosis*	1089	gbZ97188
Orf3 Mtu	Similar to transport protein	*Mycobacterium tuberculosis*	1146	gbZ77826
Orf4 Mtu	Unknown transmembrane protein	*Mycobacterium tuberculosis*	967	gbAL021932
Orf5 Mtu	Probable membrane protein	*Mycobacterium tuberculosis*	962	gbZ83860
Orf6 Mtu	Integral membrane protein	*Mycobacterium tuberculosis*	968	spQ11171
Orf7 Mtu	Hypothetical membrane protein	*Mycobacterium tuberculosis*	958	gbZ84725
Orf8 Mtu	Probable membrane protein	*Mycobacterium tuberculosis*	964	gbAL021943
Orf9 Mtu	Putative membrane transport protein	*Mycobacterium tuberculosis*	966	gbAL021928
Orf10 Mtu	Unknown membrane protein	*Mycobacterium tuberculosis*	762	gbZ92669
ActII-3 Sco	Antibiotic transport-associated protein	*Streptomyces coelicolor*	711	pirC40046
Orf1 Sco	Probable export protein	*Streptomyces coelicolor*	847	gbAL021529

Figure 9. Alignment of the HAE2 Family
An extended portion of the multiple alignment of the 18 sequenced proteins that comprise the HAE2 family of the RND superfamily (see Table 5 for protein abbreviations). The format of presentation and computer program used were as described in the legend to Figure 3.

The average hydropathy plot for the HAE2 family is shown in Figure 10. This plot is based on the complete multiple alignment, part of which is shown in Figure 9. The plot is typical of most other families within the RND superfamily, exhibiting a single N-terminal hydrophobic peak (TMS 1) centered at alignment position 50, followed by an unusually short hydrophilic domain (positions 70-230), and then by a hydrophobic region including three subregions as follows: (1) residue positions 230-305, putative TMSs 2-4; (2) positions 335-400, putative TMSs 5 and 6; and (3) a single sharp peak at positions 450-475, putative TMS 7. A strongly hydrophilic region, alignment positions 480-920, represents the second, large, putative periplasmic domain. Then another strongly hydrophobic region is observed, exhibiting five putative TMSs as follows: (1) positions 925-990, putative TMSs 8 and 9; (2) positions 990-1015, TMS 10, and positions 1030-1100, TMSs 11 and 12. The C-terminal region, positions 1100-1400, is strongly hydrophilic and presumably represents the cytoplasmic tails of these proteins.

An average similarity plot (not shown) revealed that: (1) the hydrophobic regions are best conserved; (2) the homologous hydrophilic regions following TMSs 1 and 7 are also quite well conserved; (3) much of the relatively short first periplasmic region between TMSs 1 and 2 is fairly well conserved; (4) the hydrophilic regions of about 60-80 residue positions preceding putative TMSs 2 and 8 are well conserved; (5) finally, the second, large, putative periplasmic region (positions 530-870) and the C-terminal cytoplasmic region (positions 1130-1400) are poorly conserved because they are of variable lengths in the different proteins. An average amphipathicity plot (angle = 100° for α-helix) showed that the transmembrane regions exhibited low degrees of amphipathicity although several striking peaks of amphipathicity were observed in the hydrophilic regions (data not shown).

A phylogenetic tree for the HAE2 family is shown in Figure 11. Several points are worthy of note. (1) The 18 proteins of the family fall into seven clusters, which are localized to the tips of the seven deep rooted branches. (2) Two of these branches bear a single protein, both from *B. subtilis*. These two proteins are among the most dissimilar members of the HAE2 family, suggesting that these paralogues arose very early in the evolutionary process. (3) The two *Streptomyces coelicolor* proteins, on the other hand, cluster loosely together suggesting that these paralogues arose at an intermediate time in evolutionary history. Finally, all of the proteins localized to the remaining four branches are derived exclusively from Mycobacteria (*M. tuberculosis* and *M. leprae*). One of each of the four *M. leprae* paralogues is localized to each of these four branches. Thus, Orf1 Mle is most closely related to Orf1 Mtu, Orf2 Mle is most closely related to Orf4 Mtu, Orf3 Mle is closely related to Orf10 Mtu, and Orf4 Mle is closely related to Orf9 Mtu. Each of these pairs of proteins may be orthologues.

Finally, three of the *M. tuberculosis* paralogues are about equidistant from each other on one branch (upper left in Figure 11) while five of the *M. tuberculosis* paralogues are about equidistant from each other on a second branch (lower left in Figure 11). Based on branch lengths, it seems likely that all of these paralogues arose by gene duplication events that occurred at about the same time in evolutionary history, some time before *M. tuberculosis* and *M. leprae*

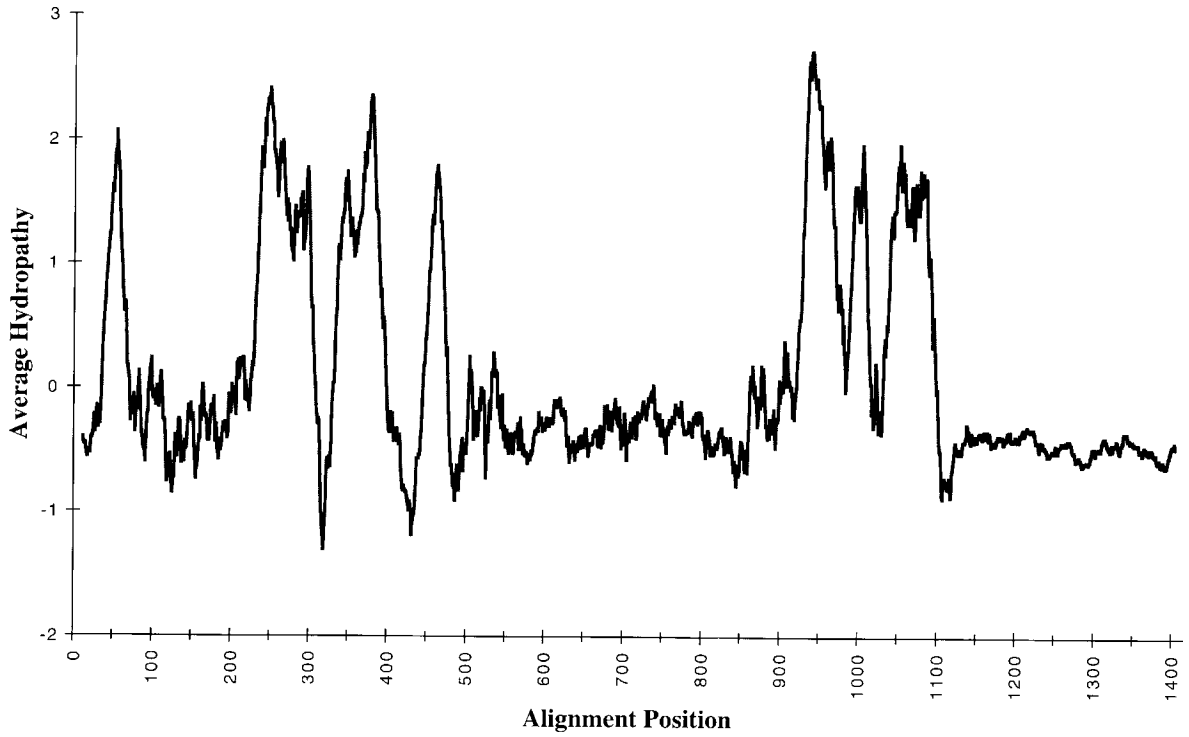

Figure 10. Average Hydropathy Plot for the 18 Proteins of the HAE2 Family
The format of presentation and computer program used were as described in the legend to Figure 4.

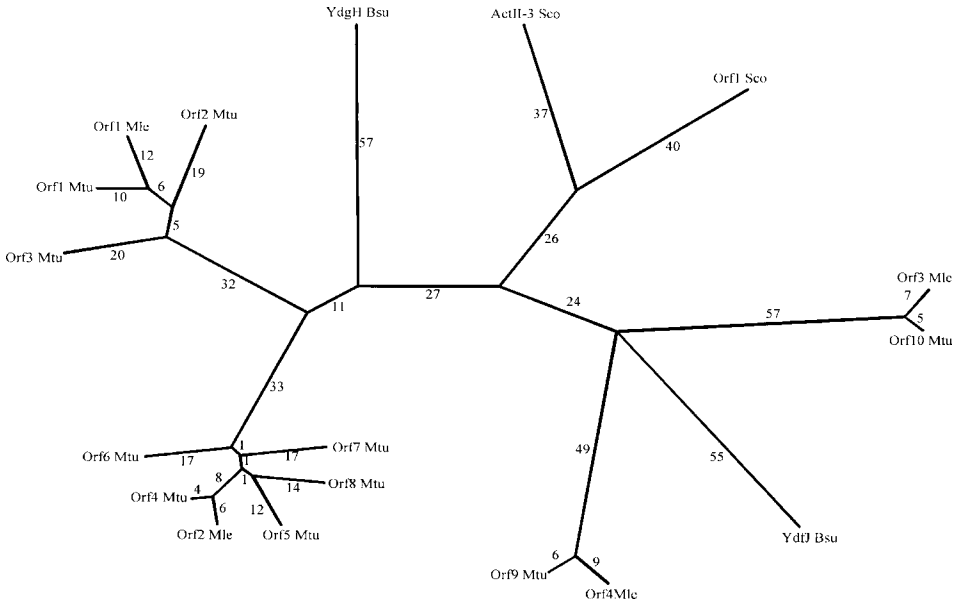

Figure 11. Phylogenetic Tree for the Sequenced Proteins of the HAE2 Proteins of the RND Superfamily The format of presentation and computer program used were as described in the legend to Figure 5.

diverged from each other. These gene duplication events presumably occurred at a time when there was strong pressure to generate many HAE2 family paralogues of similar function. If these proteins are drug efflux pumps, they may have evolved as a defense mechanism against microbial biological warfare. These proteins may provide a partial explanation for the resistance of pathogenic Mycobacterial species, particularly *M. tuberculosis*, to drug therapy (Brennan and Nikaido, 1995).

The Eukaryotic (Putative) Sterol Homeostasis (ESH) Family (TC no. 2.6.6)

The ESH family includes a group of functionally diverse proteins which serve as receptors, enzymes, and possibly transporters (Table 6). The putative transporters (*i.e.*, the Niemann-Pick C disease proteins, NPC, and the yeast YPL606w (YMP) protein) and the "patched" receptors exhibit an apparent internal repeat element as is true of most RND superfamily members (see below), but the SCAP proteins as well as the 3-hydroxy-3-methyl glutaryl-CoA reductases (HMG-CoA reductases) exhibit only one such repeat unit. These latter proteins possess minimally the 6 putative TMSs of a single repeat unit which in all of these proteins encompasses the (putative) cholesterol-sensing domain (Lange and Steck, 1998). The smallest of the homologues included in our study is 840 amino acyl residues long while the largest is 1456 residues long (Table 6). As discussed by Lange and Steck (1998) these proteins probably have differing topologies.

Figure 12 shows a partial alignment of the twenty proteins included in Table 6. Although no one residue position exhibits full conservation in the gap free alignment presented, the degree of sequence similarity between these functionally diverse proteins is appreciable as also demonstrated by the comparison scores presented in Table 2.

Table 6. Representative Members of the Eukaryotic Sterol Homeostasis (ESH) Family (TC no. 2.6.6)

Abbreviation	Description	Organism	Length	Accession Number
F31 *Cel*	F31F6.5	*Caenorhabditis elegans*	955	emblZ69884
HMG *Bge*	HMG-CoA reductase	*Blatella germanica* (German cockroach)	856	pirS30338
HMG *Spu*	HMG-CoA reductase	*Strongylocentrus purpuratus* (purple sea urchin)	932	spP16393
HMG *Xla*	HMG-CoA reductase	*Xenopus laevis*	883	spP20715
KIA *Hsa*	KIA0199 gene product	*Homo sapiens*	1277	ddbjD83782
NPC Hsa	Neimann-Picks C disease protein	*Homo sapiens*	1278	gbAF002020
NPC Mmu	Neimann-Picks C disease protein	*Mus musculus*	1278	gbAF003348
ORF Mmu	putative 12-transmembrane protein	*Mus musculus*	1182	ddbjAB010833
ORF1 Cel	coded for by *C. elegans* cDNA yk39e8.5	*Caenorhabditis elegans*	1456	gbU53340
ORF2 Cel	weakly similar to *C. elegans* proteins F54G8.5 and F44F4.4	*Caenorhabditis elegans*	1015	gbU26733
ORF3 Cel	contains similarity to transmembrane domains found in HMGR and PTC *Dme*	*Caenorhabditis elegans*	840	gbAF067945
ORF4 Cel	ORF predicted using Genefinder	*Caenorhabditis elegans*	983	emblZ82089
ORF5 Cel	similar to *Drosophila* and mouse patched proteins	*Caenorhabditis elegans*	889	gbU80447
ORF6 Cel	similar to *Drosophila* patched protein	*Caenorhabditis elegans*	933	gbU88308
PTC Dme	patched protein	*Drosophila melanogaster*	1286	spP18502
PTC Dre	patched protein	*Danio rerio*	1220	emblX98883
PTC Hsa	patched protein	*Homo sapiens*	1447	gbU59464
SCP Cgr	SREBP cleavage activating protein	*Cricetulus griseus* (Chinese hamster)	1276	gbU67060
Y39 Cel	Y39A1B.2	*Caenorhabditis elegans*	1003	embAL021482
YMP Sce	membrane protein ypl006w	*Saccharomyces cerevisiae*	1170	pirS52525

```
PTC  Dre  517 RTGDCLRRTGTSVALTSVNNMIAFFMAALVPIPALRAFSLQAAVVVVFNFAMALLIFPAILSLDL
ORF  Mmu  489 RMGECLRSTGTSVALTSVNNMVAFFMAALVPIPALRAFSLQAAIVVGCNFAAVMLVFPAILSLDL
PTC  Has  535 RTGECLKRTGASVALTSISNVTAFFMAALIPIPALRAFSLQAAVVVVFNFAMVLLIFPAILSMDL
PTC  Dme  521 QTKLILKKVGPSILFSACSTAGSFFAAAFIPVPALKVFCLQAAIVMCSNLAAALLVFPAMISLDL
NPC  Mmu  723 QLGRILGEVAPTMFLSSFSETSAFFFGALSSMPAVHTFSLFAGMAVLIDFLLQITCFVSLLGLDI
NPC  Hsa  723 QLGRVLGEVAPSMFLSSFSETVAFFLGALSVMPAVHTFSLFAGLAVFIDFLLQITCFVSLLGLDI
YMP  Sce  654 KIISAIGRMSPSILMSLLCQTGCFLIAAFVTMPAVHNFAIYSTVSVIFNGVLQLTAYVSILSLYE
ORF1 Cel  738 IVGMVMAGTMPAMFSSSLGCAFSFFIGGFTDLPAIRTFCLYAGLAVLIDVVLHCTIFLALFVWDT
ORF5 Cel  339 RIAECMADAAVSILITALTDALSFGVGTITTIPAVQIFCIYTMCALLLTFAYQLTFFCAILVYYT
ORF6 Cel  366 RMGECLADAAVSILITSSTDVLSFGVGAITTIPAVQIFCVYTGVAIFFAFIYQITFFAACLALAM
F31  Cel  382 RMSKTLSHAGVAVTITNVTDVMSFAIGCITDLPGIQFFCIYACVSVAFSYFYQLTFFSGAMAIMG
ORF2 Cel  444 RMIEAMSESAVAIFITSFTDVLSFGAGTITDIIAVQGFCAMTAACMFFTFLYQITFFAALMVISA
ORF3 Cel  357 RMGLALEEAGSAITVTSLTSVLSFGIGTYSTTPAIAIFCKFIALAIMFDWFYQLTFFAAVMAMGA
Y39  Cel  388 RMKETFADAAVSITVTSLTDLISFGVGCATPFFSVQMFCAYAVAAVIFTYIYQLTFFAAVMVYTN
ORF4 Cel  420 RMSEVMAEVGPAILISCLTNMFADAVGSFTSSPEITLLCTGNMLSMWFAFIYQMTFYAGLMSIVG
KIA  Hsa  377 RIAQGLSSESWSIMKNMATELGIILIGYFTLVPAIQEFCLFAVVGLVSDFFLQMLFFTTVLSIDI
SCP  Cgr  379 RIAQGLSSESWSIMKNVATELGIILIGYFTLVPAIQEFCLFAVVGLVSDFFLQMFFFTTVLSIDI
HMG  Spu  156 NIARGMAILGPTITLDTVVTTLVISIGTMSSIRKMEVFCCFGILSLIANYFVFMTFFPACLSLVL
HMG  Xla  155 NIARGMAILGPTFTLEALVECLVIGVGTMSGVRQLEIMCCFGCMSVLANYFAFMTFFPACVSLVL
HMG  Bge  154 NIARGIAMLGPTITLDTVVETLVIGVGMLSGVRRLEVLCCFACMSVIVNYVVFMTFYPACLSLIL
```

Figure 12. Partial Multiple Alignment of the Sequenced Proteins of the ESH Family
The format of presentation is as discussed in the legend to Figure 3 except that well-conserved residues are shaded, and residue number is provided at the beginning of each sequence (see Table 6 for protein abbreviations).

The phylogenetic tree for the 20 ESH family members is shown in Figure 13. The proteins fall into five clusters (Clusters I-V in Figure 13). Cluster I includes the two orthologous mammalian Niemann-Pick C disease proteins (human and mouse) as well as homologues from C. elegans and *S. cerevisiae*. Cluster II includes the "patched" receptor of *D. melanogaster* and its putative orthologues of mammals. Cluster III includes the SREBP cleavage activating protein of *C. griseus* as well as an orthologue from man. Cluster IV includes HMG-CoA reductases from 3 different animals. Finally, Cluster V includes a group of 7 paralogues from *C. elegans*, none of which is functionally characterized.

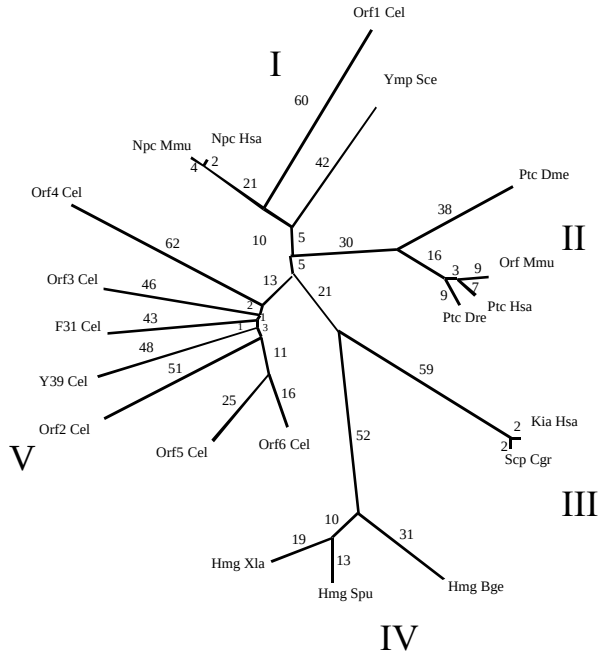

Figure 13. Phylogenetic Tree for the Proteins of the ESH Family of the RND Superfamily
Five clusters of proteins (I-V) are indicated as discussed in the text. The format of presentation and computer programs used were as described in the legend to Figure 5 (see Table 6 for protein abbreviations).

Because of the topological diversity noted above among proteins of the ESH family, the five clusters were separately analyzed with respect to their average hydropathy profiles. These are presented in Figure 14. Cluster I exhibits an average hydropathy plot similar in many respects to those generated by prokaryotic RND superfamily members except that they display an extra N-terminal domain with two distant hydrophobic peaks. Thus, 14 putative TMSs rather than 12 appear to be present, in agreement with the conclusion of Lange and Steck (1998). In other respects, the apparent topology is similar. Thus, a 1 + 3 + 2 + 1 + 3 + 2 TMS arrangement of the C-terminal 12 TMSs is apparent as noted also for the entirety of the proteins of the prokaryotic families analyzed in this paper.

Cluster II, the "patched" cluster , exhibits the expected 12 TMS topology with the same approximate spacing that is generally observed for RND superfamily members (1 + 3 + 2 + 1 + 3 + 2 TMSs). Cluster III (SCAP and its human orthologue) exhibit the typical RND superfamily topology for the first 600 residues (1 + 3 + 2 TMSs), but following this unit, four approximately equidistant peaks of hydrophobicity imply the presence of four additional transmembrane helical spanners. Lange and Steck (1998) have suggested that only two are present.

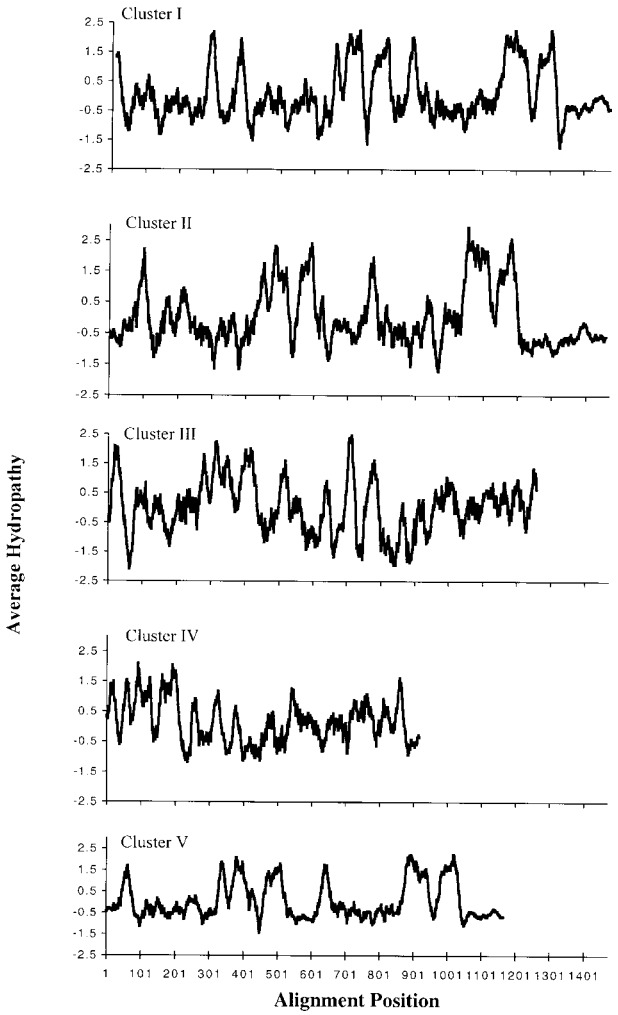

Figure 14. Hydropathy Plots for the ESH Family
Average hydropathy plots for the five clusters of the ESH family (see Figure 13). The format of presentation and computer program used were as described in the legend to Figure 4.

Cluster IV (HMG-CoA reductases) exhibits 6 peaks of hydrophobicity in a 1 + 3 + 2 arrangement in the N-terminal region of the alignment, and this hydrophobic domain is followed by a long hydrophilic tail. Lange and Steck (1998) have suggested an eight TMS topology. Finally, the average hydrophobicity plot for the seven Cluster V proteins of *C. elegans*, not analyzed by Lange and Steck, appears to be best interpreted in terms of the 1 + 3 + 2 + 1 + 3 + 2 topology that typifies the RND superfamily. The SCAP proteins (Cluster III) and the HMG-CoA reductases (Cluster IV) therefore appear to exhibit one 6 TMS repeat unit while the other three clusters exhibit two such units.

```
                    1         11        21        31
                                        **
Orf1 Afu-n  (122) SITRTGLAIVMALITTVIGFMSMLAPGMPAMAQFGII
Orf1 Afu-c  (493) TIERTGKAITTSALTMAGGFGSLMFSTFPIMQNFGFI
Orf2 Afu-n  (122) ALNHTRFPLFMAMATTVIGFASMCAPGIPSLFWFSFL
Orf2 Afu-c  (486) TVERTGKAVLTSALTTAGGFGALYFSTFPVLSNFGIL
Orf Bbu-n   (122) TIKKLKTPILLTSFTTAFGFLSLTTSSINAYKTMGIF
Orf Bbu-c   (493) SIPNVFNGIFANSISVGIGFLTLTFSSYKIISTLGAI
Orf Pho-n   (118) AISETGKALLGAALTTIAGFLALSLSILPSLKRLSIS
Orf Pho-c   (510) AMESVGPGILIGALTTAGGFLALLTGRLTAIHDFGKV
Orf Tpa-n   (123) AVDKIIQPVFLSALTTFVGFVSFCFTSVVPIFEFGVF
Orf Tpa-c   (562) TFYGSGRAILFNVLSVGSGFAVLMLSKFNVLADFGLL
Orf Mja     (118) AVVETGTAVMATTATTVVGFLALVLAPLPMMANLGKV

Consensus         ----TG-AI----LTT--GF--L-----P----FG--
```

Figure 15. Alignment of Repeat Unit Sequences from the HAE3 Family
A region of the multiple alignment of all repeat unit sequences derived from the proteins of the HAE3 family of the RND superfamily. See Table 7 for protein abbreviations. n: N-terminal half; c: C-terminal half. The format of presentation and computer program used were as described in the legend to Figure 3.

The (Largely Archaeal Putative) Hydrophobe/Amphiphile Efflux-3 (HAE3) Family (TC no. 2.6.7)

Table 7 lists the six proteins of the HAE3 family. All of these proteins were revealed by genome sequencing, and none has been functionally characterized. Four of the HAE3 family proteins are from the archaeal kingdom, two from *A. fulgidus*, one from *P. horikoshii* and one from *M. jannaschii*. However, two of the HAE3 proteins are from the spirochetes, *B. burgdorferi* and *T. pallidum*. Five of the six proteins are of about the same size (736-888 residues) but the sixth protein, the *M. jannaschii* protein, is half this size (388 residues). As will be shown below, the five full length proteins exhibit an internally duplicated unit, but the short *M. jannaschii* Orf, MJ1562, does not. Examination of the DNA flanking *orf* MJ1562 demonstrated that this gene is not duplicated as are the others. Thus, no sequence homologous to MJ1562 could be identified when the DNA sequences flanking this gene were translated in all six reading frames. We conclude that MJ1562 encodes an authentic half-sized member of the HAE3 family lacking a tandem repeat. It is possible that this Orf, if active, functions as a homodimer.

The five full length HAE3 family proteins were split into their two homologous halves (see Table 7), and the 10 halves were aligned with the half sized *M. jannaschii* homologue. A well conserved portion of the full alignment is shown in Figure 15. No gaps are observed in this partial alignment, suggesting that the divergence of two halves of the full length proteins did not involve many insertions or deletions. Although there are only two fully conserved residues near the central portion of the alignment, careful examination of the complete multiple alignment revealed that the alignment is probably correct throughout its length.

Table 7. Proteins of the (Archaeal and Spirochete) Hydrophobe/Amphiphile Efflux-3 (HAE3) Family (TC no. 2.6.7)

Abbreviation	Protein or description	Size (no. aas)	Segments analyzed n	Segments analyzed c	Source	Accession number
Orf1 Afu	Gene AF1229 protein	750	1-375	376-750	*Archaeoglobus fulgidus*	gbAE001019
Orf2 Afu	Gene AF0459 protein	736	1-368	369-736	*Archaeoglobus fulgidus*	gbAE001073
Orf Bbu	Gene BB0252 protein	767	1-383	384-767	*Borrelia burgdorferi*	gbAE001135
Orf Pho	Gene PH0287 protein	787	1-393	394-787	*Pyrococcus horikoshii*	gbAP000001
Orf Tpa	Gene TP0790 protein	888	1-444	445-888	*Treponema pallidum*	gbAE001250
Orf Mja	Gene MJ1562 protein	388	—	—	*Methanococcus jannaschii*	pirA64495

Two signature sequences were derived from the complete multiple alignment of the HAE3 family halves. These family-specific signature sequences are:

1. (L I V F Y) (R K) X_5 (L I V F) (L I V F A M) (P G M) (L I V A T) (L I V F M) (P L I V S A T) (L I V S A M) X_2 (L I V A) X_4 (G A T V) X (M L) X (L F) X_2 (L I V T S) X_2 (T S N D) X_5 (L I V A T M) X_2 (L I V M F) (L I V T S) (L I V M S) (G A S) (L I V M) (G A)(L I V C S)

2. (L I V) X_3 (L I V A F) (T S) X_3 G F (L I V M A G) (S A T V) (L M F) X_2 (P S T A G) X (L I V M F Y) X_2 (L I V M Y) X_2 (L I V M F) (G S) X (L I V F S) X_2 (L I V M F) (G A T) (L I V M)

An average hydropathy plot for the eleven half-length sequences is shown in Figure 16A. As expected for RND superfamily members, six peaks of hydropathy are observed, and these occur in a 1 + 3 + 2 arrangement with a large hydrophilic region between putative TMSs 1 and 2 and a small hydrophilic loop between putative TMSs 4 and 5. The average similarity plot shown in Figure 16B reveals that the homologous HAE3 family protein halves exhibit similar degrees of similarity throughout their lengths, but the C-termini are somewhat better conserved than the N-termini. Finally, the average amphipathicity plot (100° angle for α-helix) shown in Figure 16C reveals that the two largest amphipathic α-helical regions in these HAE3 family protein halves occur immediately before TMS 1 and shortly before TMS 2. Thus, as has been noted for several integral membrane transporters including other RND superfamily families, amphipathic regions precede the transmembrane regions. This fact may have biogenic implications (Yamada *et al.*, 1991).

Phylogenetic trees for the N-terminal halves (A), the C-terminal halves (B), and for all halves (C) of the HAE3

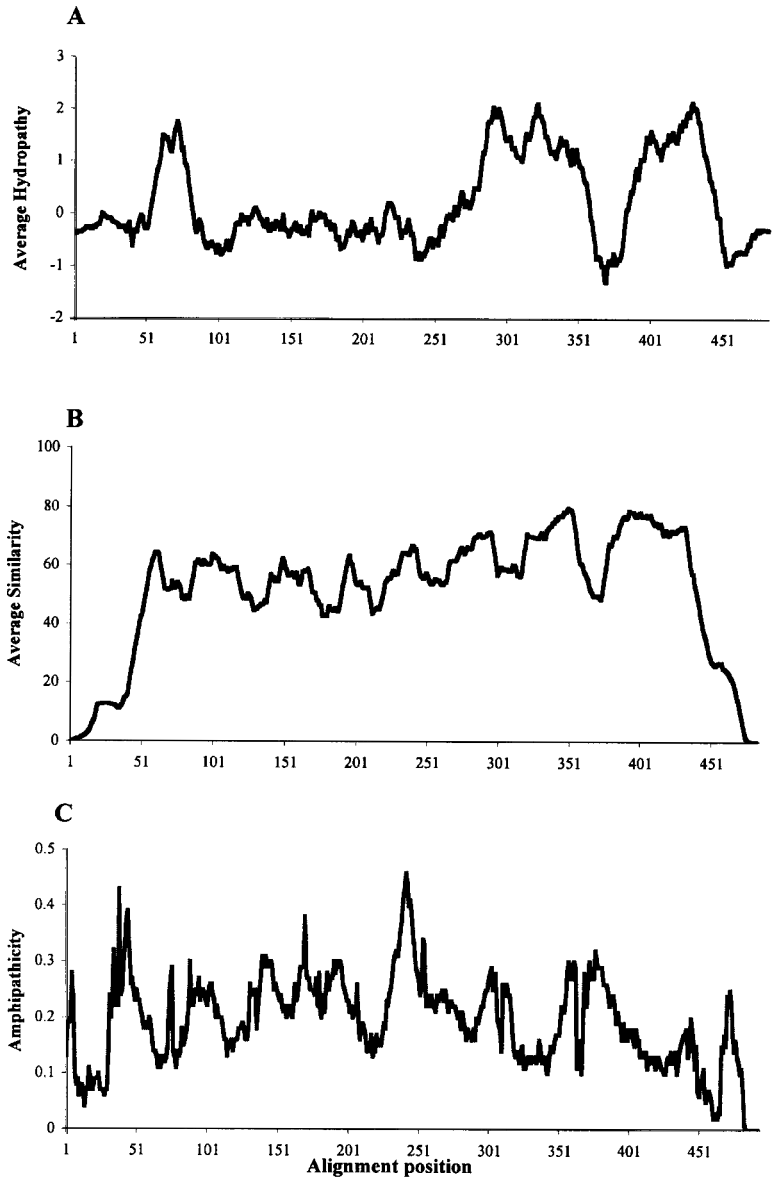

Figure 16. Hydropathy, Similarity, and Amphipathicity for the HAE3 Family
Average hydropathy (A), average similarity (B), and average amphipathicity (C) for the repeat units in the protein of the HAE3 family (see Table 7). The format of presentation and computer programs used were as described in the legend to Figure 4.

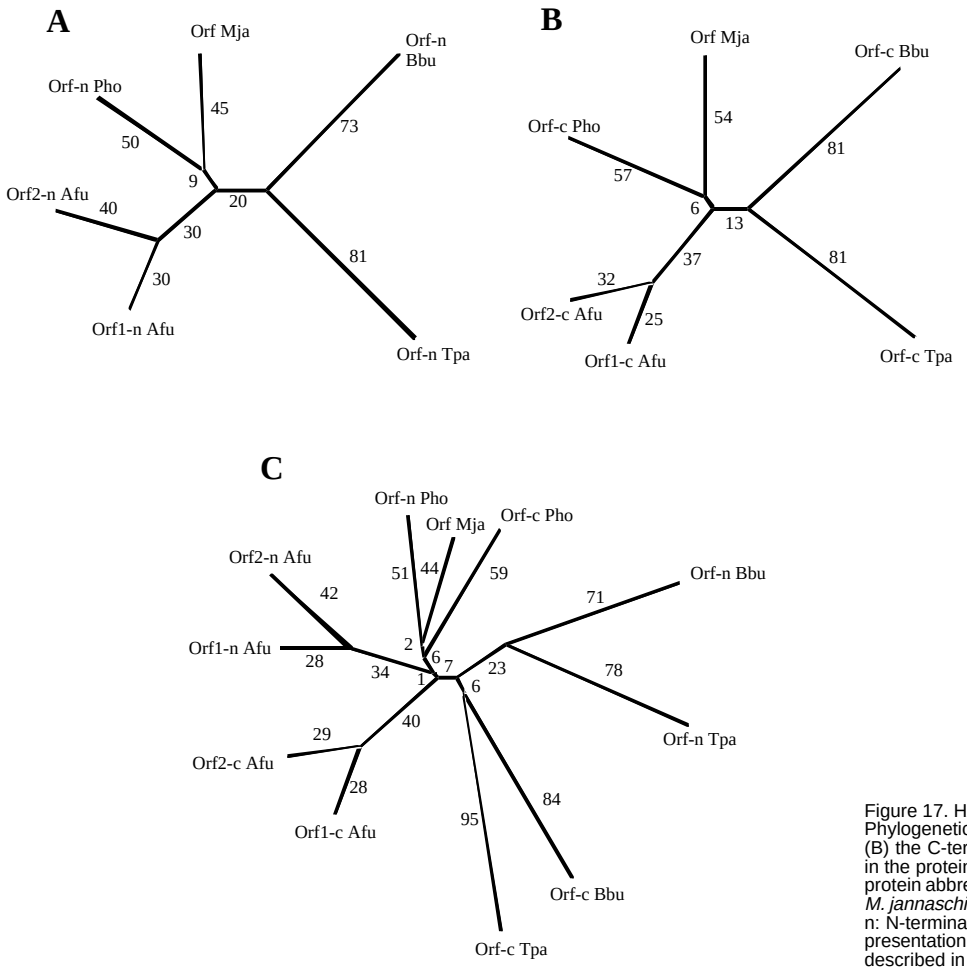

Figure 17. HAE3 Family Phylogenetic Trees
Phylogenetic trees for (A) the N-terminal halves, (B) the C-terminal halves, and (C) all repeat units in the proteins of the HAE3 family (see Table 7 for protein abbreviations). The single repeat unit of the *M. jannaschii* protein was included in all three trees. n: N-terminal half, c: Cterminal half. The format of presentation and computer programs used were as described in the legend to Figure 5.

family are presented in Figure 17. The half length *M. jannaschii* protein is included in all three trees. It is worth noting first, that the spirochete proteins cluster distantly from the archaeal proteins, second that the *A. fulgidus* proteins cluster together, and third that the *P. horikoshii* and *M. jannaschii* proteins cluster together. In Figure 17C, the N-terminal halves of both the spirochete and the *A. fulgidus* proteins cluster together as do the C-terminal halves. Interestingly, all halves emanate from points near the center of the tree. These observations suggest (1) that the tandem intragenic duplication event that gave rise to the full length spirochete proteins occurred before the *T. pallidum* lineage separated from the *B. burgdorferi* lineage, (2) that the intragenic duplication event that gave rise to both of the *A. fulgidus* proteins occurred before the extragenic duplication event that gave rise to these two paralogues from the ancestral protein, and (3) that both of these intragenic duplication events may have occurred after the archaeal kingdom split apart from the bacterial kingdom. These observations, plus the fact that the *M. jannaschii* homologue is of half length provides substantiation for the suggestion that tandem intragenic duplication events occurred more than once during the evolution of the RND superfamily.

Concluding Remarks and Perspectives

In the phylogenetic analysis of the RND superfamily reported here, we have greatly expanded upon the previously recognized RND family which had been thought to include members exclusively from Gram-negative bacteria (Saier *et al*., 1994). We now recognize that this superfamily is ubiquitous, being found in all major bacterial groups as well as archaea and eukaryotes. The functions of only a few of the Gram-negative bacterial proteins of the RND superfamily are currently known, but the partially characterized ActII3 protein of *Streptomyces coelicolor* suggests that it (and therefore its many homologues in pathogenic Mycobacteria) may either be drug efflux pumps or auxiliary proteins, facilitating drug efflux by other drug exporters as suggested by Férnandez-Moreno *et al.* (1991). We consider this second possibility to be less likely than the first. We similarly predict that the archaeal-spirochete RND family members will prove to transport drugs, but no experimental data are available to support or refute this suggestion. Finally, several of the eukaryotic RND family members have been implicated in sterol homeostasis and reception. While we suggest that the yeast YMP and human NPCI proteins are transporters, possibly pumps for the efflux of sterols and other amphipathic molecules from lysosomes, the available evidence suggests that “patched”

and SCAP are receptors. These possibilities have recently been noted independently by Sturley (1998). The dissimilar topologies of SCAP and HMG-CoA reductase are consistent with dissimilar functions for these homologues.

Within the RND superfamily, the eukaryotic RND family (ESH) and the archaeal/spirochete family (HAE3) proved to include members that did not arise by an intragenic tandem duplication event. The homologous regions of these proteins were in general half as large as those of other RND superfamily members, and they exhibited topologies typical of a single repeat unit. Because such single repeat elements were found in two of the seven RND superfamily families, we suggest that internal gene duplication occurred more than once during the evolution of the RND superfamily. The fact that the two halves of several RND superfamily proteins exhibit greater sequence similarity with each other than either do with either repeat unit in any protein of the other families of the superfamily clearly supports this contention. If intragenic duplication did occur multiple times during the evolutionary history of the RND superfamily, then a very significant functional benefit must result from inclusion of both repeat units within a single polypeptide chain, rather than having a homo- or heterodimeric permease. The fact that most (but not all) secondary carrier proteins exhibit 12 or close to 12 TMSs and that many of these families exhibit two internal repeat units clearly supports this notion. Exactly what the functional benefits of maintaining an internally duplicated polypeptide chain are remains to be ascertained. Restriction of domain mobility and/or division of functional labor between the two halves can be considered to be reasonable possibilities.

Establishment of a ubiquitous RND superfamily has numerous implications. Structural, functional and mechanistic information currently available for only a few Gram-negative bacterial proteins of the RND superfamily should be applicable to all members of the superfamily, and the degrees to which extrapolation of such information is applicable to other members of this superfamily should depend on the degrees of sequence similarity. Moreover, information obtained for **any** member of the superfamily will be relevant to **all** members. We predict that all transporters of the RND superfamily will function with outwardly directed polarity employing an H^+ antiport mechanism, but this possibility remains an open possibility. It is important to note that information obtained from the study of RND family members cannot be considered to be relevant to members of other families or to putative RND superfamily proteins which have not been proven to be homologous.

The work reported here clearly suggests that the RND superfamily is ancient, having arisen before the three major domains of life (bacteria, archaea and eukaryotes) separated from each other. Although the superfamily may have diverged extensively with respect to substrate recognition, RND transporters may not have diverged mechanistically as suggested above. This report appears to provide new evolutionary information concerning one of the largest and oldest families of transporters found in living organisms on Earth.

The analyses reported here illustrate the use of phylogenetic data to provide clues concerning the functions of proteins for which no functional data are available. They also provide useful guides for the molecular biologist interested in functional assignments. Such an approach is likely to prove applicable to thousands of proteins that are or will be revealed by genome sequencing. Many of these proteins will undoubtedly prove to be important in medicine, drug discovery and biotechnology.

Acknowledgements

We thank participants of the Bacterial Cell Surfaces Gordon Conference, held June 21-23, 1998 in New London, New Hampshire, for helpful discussions, Mary Beth Hiller for assistance in the preparation of this manuscript, and Milda Simoniaitis for assistance with the graphics. This work was supported by NIH grants 2R01 AI14176 from The National Institute of Allergy and Infectious Diseases and 9RO1 GM55434 from the National Institute of General Medical Sciences, as well as by the M.H. Saier, Sr. Research Fund (to MHS), by Deutsche Forschungsgemeinschaft (Ni262/2-1), Forschungsmittel LSA, Fonds der Chemischen Industrie (to DHN), and by Fonds National de la Recherche Scientifique and the Pôles d'Attraction Inter-Universitaire (to AG).

References

Baev, N., Endre, G., Petrovics, G., Banfalvi, Z., and Kondorosi, A. 1991. Six nodulation genes of *nod* box locus 4 in *Rhizobium meliloti* are involved in nodulation signal production: nodM codes for D-glucosamine synthetase. Mol. Gen. Genet. 228: 113–124.

Bolhuis, A., Broekhuizen, C.P., Sorokin, A., van Roosmalen, M.L., Venema, G., Bron, S., Quax, W.J., and van Dijl, J.M. 1998. SecDF of *Bacillus subtilis*, a molecular Siamese twin required for the efficient secretion of proteins. J. Biol. Chem. 273: 21217–21224.

Brennan, P.J., and Nikaido, H. 1995. The envelope of mycobacteria. Annu Rev. Biochem. 64: 29–63.

Carstea, E.D., Morris, J.A., Coleman, K.G. (and 35 others). 1997. Niemann-Pick C1 disease gene: Homology to mediators of cholesterol homeostasis. Science. 277: 228–231.

Dayhoff, M.O., Barker, W.C., and Hunt, L.T. 1983. Establishing homologies in protein sequences. Methods Enzymol. 91: 524–545.

Decottignies, A., and Goffeau, A. 1997. Complete inventory of the yeast ABC proteins. Nature Genetics. 15: 137–145.

Devereux, J., Hasberli, P., and Smithies, O. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12: 387–395.

Doolittle, R.F. 1986. Of Urfs and Orfs: A primer on how to analyze derived amino acid sequences. Mill Valley, CA, University Science Books.

Economou, A. 1998. Bacterial preprotein translocase: Mechanism and conformational dynamics of a processive enzyme. Mol. Microbiol. 27: 511–518.

Erickson, R.P., Aviles, R.A., Zhang, J., Kozloski, M.A., Garver, W.S., and Heidenreich, R.A. 1997. High-resolution mapping of the *spm* (Niemann-Pick Type C) locus on mouse chromosome 18. Mamm. Genome. 8: 355–356.

Feng, D.-F., and Doolittle, R.F. 1990. Progressive alignment and phylogenetic tree construction of protein sequences. Methods Enzymol. 183: 375–387.

Fernández-Moreno, M.A., Caballero, J.L., Hopwood, D.A., and Malpartida, F. 1991. The act cluster contains regulatory and antibiotic export genes, direct targets for translational control by the *bldA* tRNA gene of streptomyces. Cell. 66: 769–780.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A., III, and Venter, J. C. 1995. The minimal gene complement of *Mycoplasma genitalium*. Science. 270: 397–403.

Gardel, C., Benson, S., Hunt, J., Michaelis, S., and Beckwith, J. 1987. *secD*, a new gene involved in protein export in *Escherichia coli*. J. Bacteriol. 169: 1286–1290.

Gardel, C., Johnson, K., Jacq, A., and Beckwith, J. 1990. The *secD* locus of *Escherichia coli* codes for two membrane proteins required for protein export. EMBO J. 9: 3209–3216.

Göttfert, M. 1993. Regulation and function of rhizobial nodulation genes. FEMS Microbiol. Rev. 104: 39–64.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and Herrmann, R. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res. 24: 4420–4449.

Hooper, J.E., and Scott, M.P. 1989. The Drosophila *patched* gene encodes a putative membrane protein required for segmental patterning. Cell. 59:

751–765.
Hua, X., Nohturfft, A., Goldstein, J.L., and Brown, M.S. 1996. Sterol resistance in CHO cells traced to point mutation in SREBP cleavage-activating protein. Cell. 87: 415–426.
Johnson, K., Murphy, C.K., and Beckwith, J. 1992. Protein export in *Escherichia coli*. Curr. Opin. Biotechnol. 3: 481–485.
Johnson, R.L., Rothman, A.L., Xie, J., Goodrich, L.V., Bare, J.W., Bonifas, J.M., Quinn, A.G., Myers, R.M., Cox, D.R., Epstein, E.H., Jr., and Scott, M.P. 1996. Human homolog of *patched*, a candidate gene for the basal cell nevus syndrome. Science. 272: 1668–1671.
Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Peterson, S., Reich, C.I., McNeil, L.K., Badger, J.H., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J.D., Weidman, J.F., McDonald, L., Utterback, T., Cotton, M.D., Spriggs, T., Artiach, P., Kaine, B.P., Sykes, S.M., Sadow, P.W., D'Andrea, K.P., Bowman, C., Fujii, C., Garland, S.A., Mason, T.M., Olsen, G.J., Fraser, C.M., Smith, H.O., Woese, C.R., and Venter, J.C. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390: 364–370.
Kyte, J., and Doolittle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157: 105–132.
Lange, Y., and Steck, T.L. 1998. Four cholesterol-sensing proteins. Curr. Opin. Struct. Biol. 8: 435–439.
Le, T., Tseng, T.-T., and Saier, M.H., Jr. 1999. Flexible programs for the estimation of average amphipathicity of multiply aligned homologous proteins: Application to integral membrane transport proteins. Mol. Mem. Biol., in press.
Loftus, S.K., Morris, J.A., Carstea, E.D., Gu, J.Z., Cummings, C., Brown, A., Ellison, J., Ohno, K., Rosenfeld, M.A., Tagle, D.A., Pentchev, P.G., and Pavan, W.J. 1997. Murine model of Niemann-Pick C disease: Mutation in a cholesterol homeostasis gene. Science. 277: 232–235.
Nakano, Y., Guerrero, I., Hidalgo, A., Taylor, A., Whittle, J.R.S., and Ingham, P.W. 1989. A protein with several possible membrane-spanning domains encoded by the *Drosophila* segment polarity gene *patched*. Nature. 341: 508–513.
Nelissen, B., DeWachter, R., and Goffeau, A. 1997. Classification of all putative permeases and other membrane plurispanners of the major facillator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*. FEMS Microbiol. Rev. 21: 113-134.
Neufeld, E.B., Cooney, A.M., Pitha, J., Dawidowicz, E.A., Dwyer, N.K., Pentchev, P.G., and Blanchette-Mackie, E.J. 1996. Intracellular trafficking of cholesterol monitored with a cyclodextrin. J. Biol. Chem. 271: 21604–21613.
Nies, D.H. 1995. The cobalt, zinc, and cadmium efflux system CzcABC from *Alcaligenes eutrophus* functions as a cation-proton antiporter in *Escherichia coli*. J. Bacteriol. 177: 2707–2712.
Pentchev, P.G., Comly, M.E., Kruth, H.S., Tokoro, T., Butler, J., Sokol, J., Filling-Katz, M., Quirk, J.M., Marshall, D.C., Patel, S., Vanier, M.T., and Brady, R.O. 1987. Group C Niemann-Pick disease: Faulty regulation of low-density lipoprotein uptake and cholesterol storage in cultured fibroblasts. FASEB J. 1: 40–45.
Pentchev, P.G., Vanier, M.T., Suzuki, K., and Patterson, M.C. 1995. Niemann-Pick disease type C: A cellular cholesterol lipidosis. In Metabolic and Molecular Bases of Inherited Disease, Vol. 2, Chapter 85 (C.R. Scriver, A.L. Beaudet, W.S. Sly, and D.D. Valle, Eds.), New York, McGraw-Hill, pp. 2625–2639.
Pogliano, K.J., and Beckwith, J. 1994. Genetic and molecular characterization of the *Escherichia coli secD* operon and its products. J. Bacteriol. 176: 804–814.
Rensing, C., Pribyl, T., and Nies, D.H. 1997. New functions for the three subunits of the CzcCBA cation-proton antiporter. J. Bacteriol. 179: 6871–6879.
Saier, M.H., Jr. 1994. Computer-aided analyses of transport protein sequences: Gleaning evidence concerning function, structure, biogenesis, and evolution. Microbiol. Rev. 58: 71–93.
Saier, M.H., Jr. 1998. Molecular phylogeny as a basis for the classification of transport proteins from bacteria, archaea and eukarya. In Advances in Microbial Physiology (R.K. Poole, Ed.), San Diego, Academic Press, pp. 81–136.
Saier, M.H., Jr., Werner, P.K., and Müller, M. 1989. Insertion of proteins into bacterial membranes: Mechanism, characteristics and comparisons with the eukaryotic process. Microbiol. Rev. 53: 333–366.
Saier, M.H., Jr., Tam, R., Reizer, A., and Reizer, J. 1994. Two novel families of bacterial membrane proteins concerned with nodulation, cell division and transport. Mol. Microbiol. 11: 841–847.
Saier, M.H., Jr., Paulsen, I.T., Sliwinski, M.K., Pao, S.S., Skurray, R.A., and Nikaido, H. 1998. Evolutionary origins of multidrug and drug-specific efflux pumps in bacteria. FASEB J. 12: 265–274.
Schatz, P.J., Bieker, K., Ottemann, K.M., Silhavy, T.J., and Beckwith, J. 1991. One of three transmembrane stretches is sufficient for the functioning of the SecE protein, a membrane component of the *E. coli* secretion machinery. EMBO J. 10: 1749–1757.
Sturley, S. 1998. A molecular approach to understanding human sterol metabolism using yeast genetics. Curr. Opin. Lipidol. 9: 85-91.
Sugai, M., and Wu, H.C. 1992. Export of the outer membrane lipoprotein is defective in *secD, secE*, and *secF* mutants of *Escherichia coli*. J. Bacteriol. 174: 2511–2516.
Vanier, M.T., Duthel, S., Rodriguez-Lafrasse, C., Pentchev, P., and Carstea, E.D. 1996. Genetic heterogeneity in Niemann-Pick C disease: A study using somatic cell hybridization and linkage analysis. Am. J. Hum. Genet. 58: 118–125.
Watari, H., Blanchette-Mackie, E.J., Dwyer, N.K., Glick, J.M., Patel, S., Neufeld, E.B., Brady, R.O., Pentchev, P.G., and Strauss, J.F., III. 1999. Niemann-Pick C1 protein: Obligatory roles for Nterminal domains and lysosomal targeting in cholesterol mobilization. Proc. Natl. Acad. Sci. USA. 96: 805–810.
Yamada, Y., Chang, Y.-Y., Daniels, G.A., Wu, L.-F., Tomich, J.M., Yamada, M., and Saier, M.H., Jr. 1991. Insertion of the mannitol permease into the membrane of *Escherichia coli*. Possible involvement of an N-terminal amphiphilic sequence. J. Biol. Chem. 266: 17863–17871.
Young, G.B., Jack, D.L., Smith, D.W., and Saier, M.H., Jr. 1999. The amino acid/auxin:proton symport permease family. Biochim. Biophys. Acta. 1415: 306–322.