Microbiology (2003), **149**, 3051–3072 **DOI** 10.1099/mic.0.26364-0

Review

Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella

Christopher R. Peabody,[1]† Yong Joon Chung,[1]‡ Ming-Ren Yen,[1]§ Dominique Vidal-Ingigliardi,[2] Anthony P. Pugsley[2] and Milton H. Saier, Jr[1]

Correspondence
Milton H. Saier, Jr
msaier@ucsd.edu

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

[2]Unité de Génétique Moléculaire, CNRS URA 2172, Institut Pasteur, 25 rue du Dr Roux, 75724 Paris, Cedex 15, France

Homologues of the protein constituents of the *Klebsiella pneumoniae* (*Klebsiella oxytoca*) type II secreton (T2S), the *Pseudomonas aeruginosa* type IV pilus/fimbrium biogenesis machinery (T4P) and the *Methanococcus voltae* flagellum biogenesis machinery (Fla) have been identified. Known constituents of these systems include (1) a major prepilin (preflagellin), (2) several minor prepilins (preflagellins), (3) a prepilin (preflagellin) peptidase/methylase, (4) an ATPase, (5) a multispanning transmembrane (TM) protein, (6) an outer-membrane secretin (lacking in Fla) and (7) several functionally uncharacterized envelope proteins. Sequence and phylogenetic analyses led to the conclusion that, although many of the protein constituents are probably homologous, extensive sequence divergence during evolution clouds this homology so that a common ancestry can be established for all three types of systems for only two constituents, the ATPase and the TM protein. Sequence divergence of the individual T2S constituents has occurred at characteristic rates, apparently without shuffling of constituents between systems. The same is probably also true for the T4P and Fla systems. The family of ATPases is much larger than the family of TM proteins, and many ATPase homologues function in capacities unrelated to those considered here. Many phylogenetic clusters of the ATPases probably exhibit uniform function. Some of these have a corresponding TM protein homologue although others probably function without one. It is further shown that proteins that compose the different phylogenetic clusters in both the ATPase and the TM protein families exhibit unique structural characteristics that are of probable functional significance. The TM proteins are shown to have arisen by at least two dissimilar intragenic duplication events, one in the bacterial kingdom and one in the archaeal kingdom. The archaeal TM proteins are twice as large as the bacterial TM proteins, suggesting an oligomeric structure for the latter.

Overview

Three related types of prokaryotic envelope protein complexes include (putative) prepilin proteins with highly similar hydrophobic N-terminal segments of approximately 20 amino acyl residues. These putative prepilins can assemble into filamentous structures which compose parts of (1) the type II secretion system (T2S), (2) the type IV piliation/fimbriation system (T4P) (both of Gram-negative bacteria) and (3) the flagellar system (Fla) of archaea. T2S, also called the type II secreton or the main terminal branch (MTB; TC #3.A.15) of the general secretory pathway (TC #3.A.5; Cao & Saier, 2003), represents the major pathway for exoprotein transport from the periplasm across the outer membrane in a wide variety of Gram-negative bacteria (Pugsley, 1993a). The type II secreton is composed of a core of around 12 proteins, some of which are not present in all type II secretons and others of which appear to be dispensable for secreton function (Filloux *et al.*, 1998; Pugsley, 1993a; Pugsley *et al.*, 1997; Sandkvist, 2001). In this review, secreton components will be referred to according to their specific, four-letter gene designation or by the designation T2S and the last letter of their gene designation. For example, the products of the *Klebsiella*

†Present address: Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA.

‡Present address: Department of Life Science, Jeonju University, Chonju, Korea.

§Present address: Institute of Molecular Biology, National Chung Hsing University, Taichung, 402, Taiwan, Republic of China.

The online version of this review (at http://mic.sgmjournals.org) contains a supplementary table with a list of representative T2S systems from proteobacterial and chlamydial species, and a supplementary bacterial 16S rRNA tree.

pneumoniae (*Klebsiella oxytoca*) *pulD* gene (d'Enfert *et al.*, 1989), the *Aeromonas hydrophila exeD* gene (Ast *et al.*, 2002) and the *Xanthomonas campestris xpsD* gene (Hu *et al.*, 1995) are all referred to as members of the T2SD family.

Surprisingly, the T2SD family of proteins (members of the secretin superfamily; TC #1.B.22) (Martinez *et al.*, 1998; Nguyen *et al.*, 2000; Thanassi, 2002; Yen *et al.*, 2002) are the only integral outer-membrane secreton components. Therefore, they are the only ones capable of forming channels in bacterial outer membranes to permit exoprotein efflux (Bitter *et al.*, 1998; Hardie *et al.*, 1996a; Nouwen *et al.*, 1999, 2000). The well-established multimeric state and low-resolution structures of secretins (Bitter *et al.*, 1998; Brok *et al.*, 1999; Collins *et al.*, 2001, 2003) are consistent with this idea. A role in pilus biogenesis has been proposed (Sauvonnet *et al.*, 2000).

The other secreton components include the following. (1) A peripheral outer-membrane lipoprotein (the T2SS protein or pilotin) (Hardie *et al.*, 1996a, b) that has so far been found only in a small number of secreton systems. (2) A peripheral plasma membrane protein (the T2SE protein), a putative ATP-binding protein that, in one case, is reported to be monomeric and to have both ATPase and autokinase activities (Sandkvist *et al.*, 1995). T2SE proteins have characteristic signature sequences, including a highly conserved region that is flanked by aspartate residues as well as an essential zinc-finger-like motif (Possot & Pugsley, 1994, 1997). They are part of a superfamily of ATPases that includes a subfamily of multimeric proteins (often referred to as the VirB11 subfamily) involved in type IV secretion/bacterial conjugation (Cao & Saier, 2001; Krause *et al.*, 2000; Yeo *et al.*, 2000). (3) Predicted integral plasma membrane proteins (T2SA, B, C, F, G, H, I, J, K, L, M, N and O). T2SG through K (the pseudopilins) have N-terminal domains that are similar to those of type IV pilins (Nunn, 1999; Pugsley, 1993a). According to modelling based on the structure of a type IV pilin (Parge *et al.*, 1995), they may mediate subunit interactions that lead to filament formation. T2SO is the prepilin peptidase that cleaves and then *N*-methylates pseudopilins/pilins at a conserved site N-terminal to the hydrophobic region (Bleves *et al.*, 1998; Nunn & Lory, 1992, 1993; Pugsley, 1993b; Pugsley *et al.*, 2001). T2SL is required for the T2SE protein to associate with the plasma membrane and is stabilized by T2SM (Michel *et al.*, 1998; Possot *et al.*, 2000; Py *et al.*, 1999, 2001; Sandkvist *et al.*, 1995, 1999, 2000).

The precise functions of the plasma membrane protein constituents of the secreton other than T2SO remain largely a matter of conjecture although, in view of the established similarity with the T4P systems, many of them are probably involved in the assembly of a pilus-like structure (see below). T2SC, T2SL and T2SM have relatively large periplasmic domains, leading to the notion that they might form part of a *trans*-periplasmic complex that controls the opening of the secretin channel and/or recognizes and directs the substrate exoproteins to this secretin (Possot *et al.*, 2000). Nevertheless, all three of these proteins are required for pilus formation by the T2S. Other proteins, such as the T2SE ATPase and/or a proton-channel-forming constituent (possibly T2SF), could be involved in energizing secreton/pseudopilus assembly or exoprotein transport through the outer membrane (Bleves *et al.*, 1999; Letellier *et al.*, 1997; Possot *et al.*, 1997, 2000). This latter suggestion is based, in part, on a superficial analogy between protein secretion and the import of bulky ligands (e.g. siderophores and cyanocobalamin) across the outer membrane of *Escherichia coli*. The latter process is driven by the proton-motive force (pmf) via an integral plasma membrane protein complex, the TonB/ExbBD complex (Postle & Kadner, 2003). However, it is also possible that ATP hydrolysis plays a direct role in the secretory process, especially in secretons that have two ATPases, like those in *Aeromonas* species (Schoenhofen *et al.*, 1998).

The long-recognized similarity between the T2S and T4P systems (Hobbs & Mattick, 1993; Pugsley, 1993a) was strengthened by the recent observation that increased expression of the major pseudopilin (T2SG) caused bacteria expressing secreton genes to assemble a pilus composed of this protein (Sauvonnet *et al.*, 2000). The similarities between the T2S and T4F systems extend beyond the pilins/pseudopilins and prepilin peptidase to include T2SD (secretin) (Bitter *et al.*, 1998; Collins *et al.*, 2001; Schmidt *et al.*, 2001) as well as T2SE and T2SF (Nunn *et al.*, 1990). In addition, a pilotin whose sequence is unrelated to that of identified T2S proteins is required for secretin assembly and stability in T4P systems (Drake *et al.*, 1997). However, some secreton components that are needed for pilus assembly by the T2S (e.g. T2SC; Sauvonnet *et al.*, 2000) appear to be absent from the T4P system. Additionally, certain T4P systems have unique components that are required for pilus assembly (see later). These observations probably reflect the ancient separation during divergent evolution of the T2S and T4P systems.

A uniform system of nomenclature for T4P system components remains to be established. In the following sections, we will refer extensively to three relatively well-characterized T4P systems. These are from *Pseudomonas aeruginosa* (Pil), *Neisseria* (Pil) and the *E. coli* EAF plasmid (Bfp) (see footnote 3 in Table 1 for nomenclature of major T4P components in these bacteria). Many T4P systems, including these three, have two or even three ATPases that are related to T2SE. In these bacteria, T4P systems cause 'twitching' motility by cycles of pilus extrusion (assembly) and retraction (disassembly) (Merz *et al.*, 2000; Skerker & Berg, 2001). PilT/BfpF have been proposed to be the force-generating proteins (Merz *et al.*, 2000). The pilus might span the outer membrane by passing though the centre of the secretin channel (Wolfgang *et al.*, 2000).

The Fla systems of archaea (Thomas *et al.*, 2001) are less well characterized than either the T2S or T4P systems. The most prominent component of the Fla system is the flagellin, which shares similarities near its N-terminal end

with type IV pilins and pseudopilins and is processed by an enzyme with similar substrate recognition properties to prepilin peptidase (Bardy & Jarrell, 2002). ATPases and TM proteins homologous to those found in T2S and T4P systems of bacteria can also be identified. The numbers of protein constituents in the archaeal flagellar organelles are comparable to those in the bacterial T2S and T4P systems.

In this paper, we identify recognizable homologues in the current databases of the protein constituents of a generic T2S system that includes all secreton components irrespective of the bacterium in which they were identified, the related T4P systems of *P. aeruginosa* and other Gram-negative bacteria, and the related archaeal flagellar systems of *Methanococcus voltae* and other archaea. The sequences of the most conserved of these proteins are analysed for structural and phylogenetic attributes, and the conclusions resulting from these analyses are presented. Tables of proteins as well as the corresponding multiple alignments and some supplementary phylogenetic trees can be found on our website (www-biology.ucsd.edu/~msaier/supmat).

Computer methods

Computer programs used were as follows. (1) The PSI-BLAST program (Altschul *et al.*, 1997) with iterations to convergence was used to screen the databases for homologues of the three systems (T2S, T4P and Fla) that represent the focus of this study. The query sequences were the Pul (*K. oxytoca*), Pil (*P. aeruginosa*) and Fla systems (*M. voltae*) (see our website www-biology.ucsd.edu/~msaier/supmat). The homologues found and reported in this review represent those proteins in the databases as of February 2002. (2) The Clustal X program (Thompson *et al.*, 1997) and (3) the TREE program (Feng & Doolittle, 1990) were used for multiple alignment of homologous sequences and derivation of phylogenetic trees with the aid of the BLOSUM30 scoring matrix and the TREEVIEW drawing program (Page, 1996; see Young *et al.*, 1999 for evaluation of these and other relevant programs). (4) The TMPred program (Hofmann & Stoffel, 1993) and (5) the TopPred2 program (von Heijne, 1992) were used for prediction of the integral membrane topologies of individual proteins. (6) The DAS program was used for prediction of secondary structure. (7) The WHAT program (Zhai & Saier, 2001b), with a sliding window of from 7 to 21 residues, was used to simultaneously predict hydropathy, amphipathicity (angle of 100° for α-helix; angle of 180° for β-strand), topology and secondary structure of individual proteins. (8) The AveHAS program (Zhai & Saier, 2001a) was used for plotting mean hydropathy, similarity and amphipathicity as a function of alignment position in the multiple alignments. These programs are available on our 'software' and 'biotools' websites (http://www-biology.ucsd.edu/~msaier/transport/ and http://www-biology.ucsd.edu/~yzhai/biotools.html, respectively). (9) The GAP program was used to establish homology (Devereux *et al.*, 1984; Saier, 1994).

In this paper, we use the WHAT (Zhai & Saier, 2001b) and AveHAS (Zhai & Saier, 2001a) programs in combination to predict transmembrane segments (TMSs). These programs combine several established programs to make structural predictions about transmembrane proteins. For example, the WHAT program examines individual proteins, using JNET (Cuff *et al.*, 1998) and MEMSAT (Jones *et al.*, 1994) for secondary structure and transmembrane topology prediction, respectively. Both of these programs are among the best available for these purposes. The AveHAS program first generates a multiple alignment for a collection of homologous sequences (Thompson *et al.*, 1997) and then averages (1) hydropathy, (2) amphipathicity and (3) similarity plots to provide structural information that is much more reliable than possible when evaluating a single protein sequence (Zhai & Saier, 2001a). Transmembrane β-strands can thus be accurately predicted because they exhibit (1) predicted β-structure using JNET, (2) increased hydrophobicity, relative to other portions of the polypeptide chain, and (3) increased amphipathicity when the angle is set at 180° as is appropriate for β-strands (Le *et al.*, 1999; Zhai & Saier, 2002). This method predicts transmembrane β-strands with about 80 % accuracy.

Supplementary material which can be found on our website (www-biology.ucsd.edu/~msaier/supmat) includes: (1) tables of all homologues of the different protein types included in this study, (2) the multiple alignments for these homologues, (3) the phylogenetic trees for these same families of proteins, (4) a 16S rRNA phylogenetic tree for all bacteria from which proteins included in this study were derived and (5) a tabulation of known protein constituents of all T2S systems for which homologues of all or most constituents of the secreton have been identified.

Complementation of the *pulF* deletion in the complete *pul* gene cluster was carried out using pBR322 derivatives by homologous genes under *lac*p control in a compatible plasmid, as described by Possot *et al.* (2000). *gspF* was amplified using specific primers that incorporated restriction endonuclease cleavage sites for cloning, as previously described (Possot *et al.*, 2000).

T2S, T4P and Fla system constituents

Table 1 presents a summary of the protein constituents of representative T2S, T4P and Fla systems. Four of the proteins in the T2S and T4P systems are demonstrably homologous. These are (1) the prepilin peptidase/*N*-methyltransferase, (2) the ATPase, (3) the secretin and (4) the multispanning transmembrane (TM) protein. Three of these, the ATPase, the TM protein and the prepilin peptidase, have been identified in Fla systems. The first two are clearly homologous to their T2S/T4P counterparts (Patenge *et al.*, 2001; Thomas & Jarrell, 2001). As indicated in Tables 2 (T2S) and 3 (T4P), these are the constituents of both the T2S and T4P systems with the largest numbers of recognizable homologues. Archaeal FlaK shows only weak similarity to established bacterial prepilin peptidases (Bardy

Table 1. Comparison of functionally equivalent T2S, T4P and archaeal Fla proteins

	T2S*	TMS†	T4P‡	TMS†	Fla§	TMS†	H‖
Major prepilins or flagellins	G	1	A	1	FlaB1	1	
					FlaB2	1	
Minor prepilins or flagellins	H	1	E	1			
	I	1	FimT	1	FlaA	1	
	J	1	FimU	1	FlaB3	1	
	K	1	V	1			
			W	1			
			X	1			
Peptidase/*N*-methylase	O	7	D	7	FlaK		√
ATPase	E	0	B	0	FlaI	1	√
			T	0			
			U	0			
Secretin	preD¶	1 (+ β-strands)	preQ¶	1 (+ β-strands)			√
TM protein	F	3	C	3	FlaJ	7–9	√
Pilotin	preS#	1	preP#	1			
Others	A	1	F	1	FlaC	0	
	B	1	M	2	FlaD	0	
	C	1	N	1	FlaE	0	
	L	1	O	1	FlaF	1	
	M	1	Y1	2	FlaG	1	
	N	1	Z	1	FlaH	0	

*Fourth letter protein (gene) designations are as in the *K. oxytoca* Pul system, except for A and B, which are from the *Aeromonas hydrophila* Exe system.

†Number of putative α-helical transmembrane segments, except for the secretins, which are believed to span the outer membrane 12 times as β-strands (Yen *et al.*, 2002).

‡Fourth letter protein (gene) designations are as in the *P. aeruginosa* Pil system, except for FimT and FimU, which are also from *P. aeruginosa*. The neisserial Pil and *E. coli* Bfp protein equivalents of the *P. aeruginosa* Pil proteins are respectively: PilE and BfpA for PilA; PilQ and BfpB for PilQ; PilP and BfpG for PilP; PilF and BfpD for PilB; PilT and BfpF for PilT; PilG and BfpE for PilC; PilD and BfpP for PilD (XcpA).

§Fla protein designations are those used for the *Methanococcus voltae* flagellum as well as many other archaeal flagella.

‖These protein constituents are demonstrably homologous between the bacterial Pul (T2S) and Pil (T4P) systems, and the ATPases and multispanning transmembrane (TM) proteins are also homologous with the archaeal Fla system constituents. Peptidase/*N*-methylase constituents and secretins have not been identified in the archaeal Fla systems.

¶There are an estimated 12 β-strands in secretins that anchor these proteins in the outer membrane in an oligomeric structure [10–20 subunits; 12 subunits for PilQ of *Neisseria meningitidis* (Collins *et al.*, 2001)]. The single TMS is the signal peptide.

#The one TM segment present in pilotins is the signal peptide.

& Jarrell, 2002). Otherwise, homology between constituents could not be demonstrated.

T2S systems

The online version of this review (at http://mic.sgmjournals.org) contains a supplementary table (Table S1) with a list of representative T2S systems from proteobacterial and chlamydial species. Homologues of most or all of the known T2S proteins could be identified in several of these systems. In most cases, the genes are closely linked, although their order and organization in transcription units vary considerably from system to system. Phylogenetic trees were derived using four large and well-conserved constituents (see Table 1) that represent some of the best-defined secreton constituents from functional standpoints (Fig. 1a–d). Fig. 1(a) presents the outer-membrane secretins (T2SD); Fig. 1(b) presents the cytoplasmic ATPases (T2SE); Fig. 1(c) presents the putative multispanning TM plasma membrane proteins (T2SF); and Fig. 1(d) presents the largest of the prepilin subunits (T2SK). Examination of these four trees reveals that, within experimental error, they all have the same configuration. Thus, proceeding from the top of these trees in the clockwise direction: (1) The same γ-proteobacterial proteins always cluster together at the top of the trees. (2) Next, two *P. aeruginosa* proteins cluster together. (3) These are followed by the *Caulobacter* homologue, Ccr. (4) The *Aquifex* homologues, Aae, are next (displaced in Fig. 1d). (5) These are followed by the closely related *Xylella* and *Xanthomonas* proteins. (6) The chlamydial clusters (absent in the tree shown in Fig. 1d; see Table 3) are together in Fig. 1(a–c) in the expected

Table 2. Protein constituents of the *Klebsiella pneumoniae* secreton (Pul) as well as additional T2S components from the *Aeromonas hydrophila* Exe system

Protein type	No. of proteins*	Size of protein (no. of amino acids)	Size range (no. of amino acids)	Mean size ± SD	Bacterial representation†	No. of putative TMSs	Secondary structure‡	Probable location§	Proposed function
Major prepseudopilin									
PulG	25	140	123–182	145 ± 13	γ1, γ3, γ4, α, β, Aqu, Chla	1	α-helix (9–31, 37–61, 68–75) β-strand (116–120)	PM/OM	Major component of the pilus-like structure
Minor prepseudopilins									
PulH	21	159	119–194	165 ± 18	γ1, γ3, γ4, α, β, Aqu	1	α-helix (10–31, 38–57, 122–126) β-strand (65–71, 74–80, 142–146, 166–171)	PM/OM	
PulI	18	121	119–173	128 ± 12	γ1, γ3, γ4, α, β, Aqu	1	α-helix (10–33, 39–58) β-strand (72–74, 95–101, 110–117)	PM/OM	
PulJ	18	198	173–237	208 ± 15	γ1, γ3, γ4, α, β, Aqu	1	α-helix (25–83, 163–171) β-strand (107–112, 138–144, 157–161, 192–197, 204–208)	PM/OM	
PulK	18	326	281–369	323 ± 24	γ1, γ3, γ4, α, β, Aqu	1	α-helix (15–72, 146–157, 163–178, 230–238, 245–251, 284–291, 301–306) β-strand (114–122, 329–337, 343–352, 356–364)	PM/OM	
Prepilin peptidase/ *N*-methyltransferase									
PulO	48	279	155–388	258 ± 44	γ1, γ2, γ3, γ4, α, β, δ, ε, LG+, HG+, Aqu, CC, Chla, Dei, Grb, Th	7	α-helix (2–24, 30–42, 52–75, 85–95, 100–122, 136–148)	PM	Prepilin peptidase/ *N*-methyltransferase that processes proteins G, H, I, J and K
ATPase									
PulE	167	497	312–1136	471 ± 122	γ1, γ2, γ3, γ4, α, β, ε, HG+, LG+, Arc, Aqu, CC, Chla, Chlo, Dei, Th	0	α-helix (56–65, 73–79, 184–193, 209–218, 258–268, 283–291, 309–317, 330–336, 382–390) β-strand (92–95, 145–151, 168–172, 225–229, 243–246, 274–277, 343–348)	C	Biogenesis or regulation (ATPase)
Secretin									
PulD	96	660	273–1285	609 ± 165	γ1, γ2, γ3, γ4, α, β, δ, ε, Aqu, CC, Chla, Chlo, Dei, Grb, Th	?	α-helix (27–34, 132–137, 310–320, 350–361, 441–454, 475–482, 549–559) β-strand (171–174, 192–195, 210–215, 241–248, 300–303, 331–334, 341–345, 432–437, 461–470, 487–491, 497–500, 537–543, 564–567, 570–574, 578–582, 586–594, 601–618, 624–631, 645–657, 662–670, 685–692, 703–712)	PM	Forms the outer-membrane pore (secretin)
Pilotin									
PulS	5	125	119–168	140 ± 21	γ1	1	α-helix (2–20, 38–53, 67–78, 89–101, 116–126)	OM	Promotes outer-membrane insertion of protein D

Table 2. cont.

Protein type	No. of proteins*	Size of protein (no. of amino acids)	Size range (no. of amino acids)	Mean size ± SD	Bacterial representation†	No. of putative TMSs	Secondary structure‡	Probable location§	Proposed function
Multispanning TM protein									
PulF	69	381	282–651	392 ± 59	γ1, γ2, γ3, γ4, α, β, LG+, Aqu, Arc, CC, Chla, Dei, Th	3	α-helix (2–9, 24–32, 65–76, 82–94, 99–113, 118–124, 131–142, 146–200, 208–221, 224–244, 250–259, 264–282, 288–298, 302–318, 321–328, 332–345, 350–405) β-strand (14–19)	PM	
Others									
ExeA	3	–	489–547	522 ± 30	γ1	1	α-helix (28–35, 56–64, 83–86, 102–118, 137–147, 166–182, 195–200, 214–224, 230–245, 254–291, 321–331, 335–341, 372–374, 451–462, 500–510) β-strand (43–58, 71–75, 123–127, 155–160, 382–387, 393–400, 407–411, 431–436)	PM	Assembly of D
ExeB	4	174	174–256	219 ± 34	γ1	1	α-helix (25–46, 72–79) β-strand (105–107, 114–118, 142–147)	PM	Assembly of D
PulC	20	285	112–508	290 ± 122	γ1, γ3, γ4, α, β, Chla	1	α-helix (28–42, 187–197, 230–235, 253–262) β-strand (45–50, 110–115, 121–127, 132–137, 146–151, 155–160, 163–167, 203–212, 219–223, 241–244, 271–276, 281–286)	PM/OM	Substrate recognition?
PulL	14	398	286–444	390 ± 38	γ1, γ3, α, β	1	α-helix (75–98, 121–133, 175–186, 215–221, 248–288, 302–311, 319–331, 360–374) β-strand (9–12, 57–62, 113–118, 162–165, 351–356)	PM	Attachment of PulE to PM
PulM	14	161	153–197	168 ± 11	γ1, γ3, α, β	1	α-helix (2–24, 28–77, 82–87, 99–109, 135–145, 152–155) β-strand (114–117, 125–128, 163–169)	PM	Interacts with and stabilizes PulL
PulN	6	252	219–252	246 ± 13	γ1, γ3	1	α-helix (3–30, 47–50, 140–146, 217–225, 237–246) β-strand (87–93, 188–193)	PM	No known role in secretion

*Number of proteins: number of protein homologues identified in the NCBI database as of February 2002.

†Greek letters refer to the subclasses of the proteobacteria represented (Yen *et al.*, 2002). Other abbreviations: LG+, low-G+C Gram-positive bacteria; HG+, high-G+C Gram-positive bacteria; Arc, archaea; Aqu, *Aquifex aeolicus*; CC, cyanobacteria/chloroplasts; Chla, chlamydia; Chlo, *Chlorobium*; Dei, *Deinococcus radiodurans*; Grb, green non-sulfur bacteria; Th, *Thermus aquaticus*.

‡Positions of secondary structural elements were determined by using the Clustal X alignment and the Jpred program. These secondary structural predictions are means for the entire protein family. Only structural predictions with a Jpred score of 4 or better are reported.

§OM, Outer membrane; PM, plasma membrane; C, cytoplasm.

Table 3. Protein constituents of the *P. aeruginosa* T4P (Pil) system and homologues

Protein type	No. of proteins*	Size of Pil protein (*P. aeruginosa*)	Size range	Mean size ± SD	Bacterial representation†	No. of putative TMSs	Secondary structure‡‖	Probable location§	Proposed function
Major pilins									
PilA	80	150	87–214	155 ± 16	γ1, γ2, γ3, γ4, β	1	α-helix (31–44, 49–56) β-strand (89–95, 101–106, 119–123)	PM/OM	The major subunits which form the extracellular filaments of type IV pili
PilE	12	141	125–169	143 ± 12	γ1, γ3, γ4, β	1	α-helix (16–36, 40–55) β-strand (7–14, 78–83, 89–92, 104–109)	PM/OM	The major subunits which form the extracellular filaments of type IV pili
Minor prepilins									
FimT and U	10	169, 168	148–222	182 ± 23	γ3, γ4, β	1	α-helix (14–28, 34–62) β-strand (68–72, 88–93, 119–122, 142–146, 153–156, 164–168)	PM/OM	Prepilin-like proteins
PilV	9	185	137–206	177 ± 25	γ1, γ3, γ4, β	1	α-helix (14–62, 105–112) β-strand (127–132)	PM/OM	Prepilin-like proteins
PilW	7	274	313–416	337 ± 48	γ3, γ4, β	1	α-helix (3–6, 10–33, 39–46, 52–65) β-strand (123–125, 171–182, 231–235, 237–243, 325–332, 348–351, 371–376)	PM/OM	Prepilin-like proteins
PilX	3	195	195–1161	517 ± 558	γ3, γ4	1	α-helix (9–24, 30–56, 156–168) β-strand (62–65, 114–116, 182–186)	PM/OM	Prepilin-like proteins
Peptidase/ *N*-methylase									
PilD	48	290	155–388	258 ± 44	γ1, γ2, γ3, γ4, α, β, δ, ε, LG+, HG+, Aqu, CC, Chla, Dei, Grb, Th	7	α-helix (2–24, 30–42, 52–75, 85–95, 100–122, 136–148)	PM	Bifunctional enzyme that carries out both cleavage and *N*-methylation of the PilA pilin subunit
ATPase									
PilB, T and U	167	566	312–1136	471 ± 122	γ1, γ2, γ3, γ4, α, β, ε, HG+, LG+, Arc, Aqu, CC, Chla, Chlo, Dei, Th	0	α-helix (56–65, 73–79, 184–193, 209–218, 258–268, 283–291, 309–317, 330–336, 382–390) β-strand (92–95, 145–151, 158–162, 225–229, 243–246, 274–277, 343–348)	PM	Nucleotide binding proteins – involved in regulating pilus retraction and extension
Secretin									
PilQ	96	714	273–1285	609 ± 165	γ1, γ2, γ3, γ4, α, β, δ, ε, Aqu, CC, Chla, Chlo, Dei, Grb, Th	?	α-helix (27–34, 132–137, 310–320, 350–361, 441–454, 475–482, 549–559)	OM	Forms the outer-membrane pore and a gated channel (secretin)

Table 3. cont.

Protein type	No. of proteins*	Size of Pil protein (*P. aeruginosa*)	Size range	Mean size ± SD	Bacterial representation†	No. of putative TMSs	Secondary structure‡‖	Probable location§	Proposed function
							β-strand (171–174, 192–195, 210–215, 241–248, 300–303, 331–334, 341–345, 432–437, 461–470, 487–491, 497–500, 537–543, 564–567, 570–574, 578–582, 586–594, 601–618, 624–631, 645–657, 662–670, 685–692, 703–712)		
TM protein									
PilC	69	406	282–651	392 ± 59	γ1, γ2, γ3, γ4, α, β, LG+, Aqu, Arc, CC, Chla, Dei, Th	4	α-helix (2–9, 24–32, 65–76, 82–94, 99–113, 118–124, 131–142, 146–200, 208–221, 224–244, 250–259, 264–282, 288–300, 302–318, 321–328, 332–345, 350–405) β-strand (14–19)	PM	A polytopic inner-membrane protein
Others									
PilF	6	252	179–253	225 ± 35	γ1, γ2, γ3, γ4, β	1	α-helix (3–7, 29–41, 50–60, 67–77, 87–96, 118–126, 153–164)	–	Involved in pilus biogenesis
PilM	76	354	156–678	405 ± 81	γ1, γ2, γ3, γ4, α, β, ε, HG+, LG+, Arc, Aqu, Cbf, Chla, Dei, Spi, Th	2	α-helix (54–58, 125–135, 256–264, 270–277, 313–332, 356–367) β-strand (6–10, 15–22, 28–32, 96–100, 105–108, 174–181, 225–229, 235–241, 295–298, 347–351)	–	Involved in the export of the pilus subunit, and is required for pilus assembly
PilN	10	198	196–373	253 ± 80	γ1, γ3, γ4, α, β	1	α-helix (3–7, 16–21, 30–95, 137–146) β-strand (128–133, 175–181)	–	Involved in the export of the pilus subunit, and is required for pilus assembly
PilO	7	207	199–245	216 ± 15	γ, β	1	α-helix (29–48, 54–75, 82–99, 110–117, 155–159) β-strand (175–178, 189–192)	–	Involved in the export of the pilus subunit, and is required for pilus assembly
PilP	7	174	172–181	176 ± 3	γ1, γ3, γ4, β	4	α-helix (4–8, 22–35) β-strand (102–106, 112–118, 122–126, 130–133, 138–142, 145–152, 165–168)	OM	Pilotin; required for secretin stability and pilus assembly
PilY1	14	1161	1033–1472	1129 ± 127	γ3, γ4, β	2?	α-helix (11–17, 26–35, 186–213, 221–224, 243–249, 259–265, 275–287, 359–365, 456–463, 475–489, 551–559, 590–594, 651–655, 667–673, 705–714, 811–817, 869–873, 913–921, 1041–1046)	–	

Table 3. cont.

Protein type	No. of proteins*	Size of Pil protein (*P. aeruginosa*)	Size range	Mean size ± SD	Bacterial representation†	No. of putative TMSs	Secondary structure‡‖	Probable location§	Proposed function
							β-strand (41–44, 52–54, 76–79, 100–103, 107–110, 138–142, 312–316, 406–412, 568–572, 679–682, 724–727, 757–761, 799–802, 827–833, 894–900, 934–938, 1002–1006, 1026–1031, 1050–1055, 1108–1110, 1139–1141)		
PilZ	3	118	116–118	117 ± 1	*γ*3, *γ*4, *β*	1	*α*-helix (21–32)	OM	Supports outer-membrane insertion of protein D
							β-strand (69–73, 85–87)		

*Number of proteins: number of protein homologues identified in the NCBI database as of February 2002.
†Abbreviations of the organisms are as indicated in footnote † of Table 2.
‡Positions of secondary structural elements were determined by using the Clustal X alignment and the Jpred program. Therefore, these secondary structural predictions are means for the entire protein family. Only structural predictions with a Jpred score of 4 or better are reported.
§OM, Outer membrane; PM, plasma membrane; C, cytoplasm; −, unknown.
‖Jpred only takes inputs up to 800 residues. Therefore, the secondary structural predictions for PilY1 are from the program PHD using only PilY1 (*P. aeruginosa*) as a representative of this family. Only structures with a reliability score of 5 or better are reported.

positions. (7) The loose cluster including the *P. aeruginosa* Hxc system constituents together with the tightly clustering Bce and Bps proteins are next observed. (8) Finally, the Lpn proteins can be found just to the left of the *γ*-proteobacterial clusters. The similar relative configurations and branch lengths exhibited in these four trees suggest that the constituents of these secreton systems have evolved in parallel from a single common ancestral system without shuffling of constituents between systems throughout most of evolutionary history.

An interesting phenomenon relating to certain secretins concerns their pilotin proteins, which are essential for the insertion of the secretin into the outer membrane. The prototypical representative of this family is the lipoprotein PulS. Homologues of PulS, which binds to the C-terminal regions of PulD homologues (Hardie *et al.*, 1996b), are absent from several species of bacteria that produce T2SD-type secretins. A complete phylogenetic tree of PulS homologues cannot be constructed because of the limited number of known examples, possibly due to their poor conservation which renders them non-identifiable by sequence comparisons (see www-biology.ucsd.edu/~msaier/supmat). One of the difficulties in determining whether PulS-like pilotins exist in other T2S systems is that the C-terminal regions of the secretins, to which they probably bind, are among the least conserved regions of these proteins, suggesting that the pilotins may also be poorly conserved. Other proteins, notably T2SA and T2SB (Ast *et al.*, 2002; Condemine & Shevchik, 2000), have been proposed to perform PulS-like functions. Interestingly, *E. coli* K-12 has genes for T2SA and T2SB proteins within its secreton gene cluster (Francetic *et al.*, 2000) as well as an unlinked gene, *yacC*, with limited similarity to *pulS*.

Examination of the phylogenetic trees for the other constituents of these systems (see Table 3 and www-biology.ucsd.edu/~msaier/supmat) revealed that they exhibit essentially the same configurations and relative branch lengths within experimental error. Thus we conclude that secreton systems have probably evolved by whole gene cluster duplication and by speciation without appreciable exchange of constituents between systems.

The T2SC family of proteins deserves special mention for two reasons. First, the *Xanthomonas campestris* gene designated *xpsN* (Lee *et al.*, 2000, 2001) was clearly misnamed, since it is similar to genes for T2SC proteins (and, therefore, should be called *xpsC*) and is unrelated to genes for T2SN proteins. This allows one to rationalize recent data showing that XpsC(N) is essential and interacts with proteins D, M and/or L (Lee *et al.*, 2000, 2001). PulN is not essential while T2SC proteins are essential and interact with T2SD, L and/or M proteins (Bleves *et al.*, 1999; Possot *et al.*, 2000). Second, the T2SC family of proteins can be divided into several distinct clusters depending on whether they possess (1) a coiled-coil segment, (2) a PDZ-type structure (Gerard-Vincent *et al.*, 2002; Pallen & Ponting, 1997) or (3) neither, close to the C-terminal ends of the proteins (Fig. 2). It is

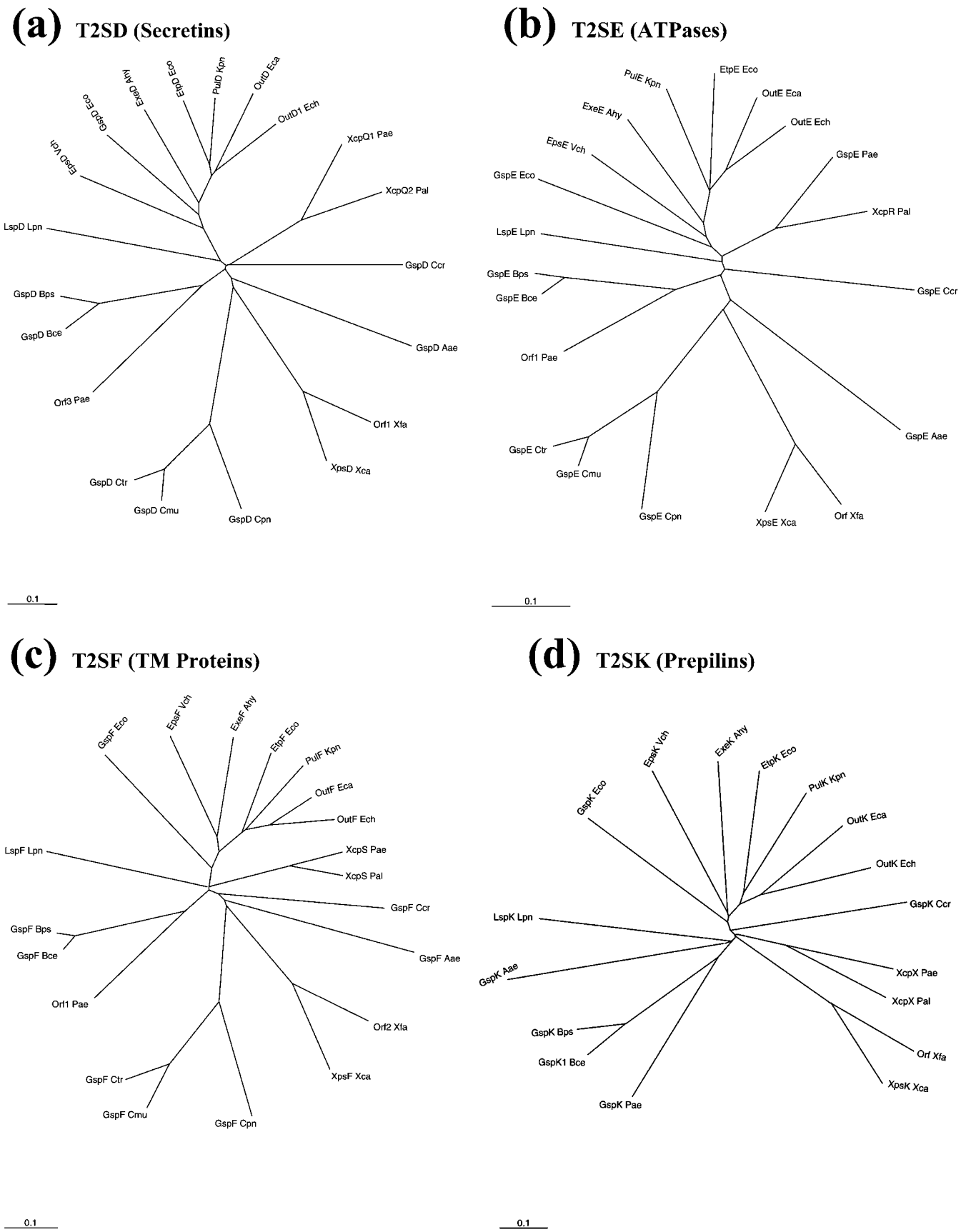

Fig. 1. Phylogenetic trees for representative proteins of the secreton: (a) T2SD secretins; (b) T2SE ATPases; (c) T2SF multispanning TM integral membrane proteins; and (d) T2SK prepilins. All trees shown here and in Figs 2–5 were generated using the Clustal X program. Protein abbreviations can be found on our website (www-biology.ucsd.edu/~msaier/supmat). These homologues are tabulated in supplementary Table S1 in the online version of this review (at http://mic.sgmjournals.org).

interesting to note that at least one member of the last class, HxcC from *P. aeruginosa*, is apparently functional (Ball *et al.*, 2002), indicating that neither the coiled-coil domain nor the PDZ domain is essential for secretion. Furthermore, the PDZ domain, predicted to exist in PulC (Pallen & Ponting, 1997), is also predicted to be a coiled-coil by the algorithms we used (Fig. 2).

Pilins (T2SG, H, I, J and K) and the pilin processing enzyme (T2SO)

As summarized in Table 1, homologues of the five prepilin constituents of T2S systems (T2SG, H, I, J and K) are found in similar numbers (18–25). Phylogenetic trees for these five families of homologues resembled those shown in Fig. 1(a–c) although only data for ~20 T2SK homologues are shown in Fig. 1(d). (The other trees can be viewed on our website at www-biology.ucsd.edu/~msaier/supmat). The tree for T2SO homologues, prepilin peptidase/*N*-methyltransferase that cleave and methylate prepseudopilins, revealed that they also cluster as shown in Fig. 1(a–d), within experimental error (see www-biology.ucsd.edu/~msaier/supmat). Thus sequence-related pseudopilins from different bacteria always fell into the same clusters, irrespective of the class of pseudopilin (G, H, I, J or K) analysed.

In addition to the authenticated Gram-negative bacterial T2SO proteins, members of the prepilin peptidase family are derived from both high- and low-G+C Gram-positive bacteria as well as very diverse Gram-negative bacteria (e.g. *Chlorobium*, *Deinococcus*, *Synechocystis* and *Thermatoga*) and might have different functions with related or unrelated substrate specificities. Moreover, a single organism may have multiple paralogues. For example, seven have been identified in *E. coli*, all very divergent in sequence, branching from points near the centre of the phylogenetic tree. They must have resulted from early gene duplication events or possibly were acquired by lateral transfer. On the other hand, only one prepilin peptidase gene is present in *P. aeruginosa* [PilD/XcpO(A)] and many other bacteria with fully sequenced genomes.

T2SE/T4PC/FlaI ATPase phylogeny

One hundred and sixty-seven T2SE homologues were identified, their sequences were multiply aligned, and from this alignment, a phylogenetic tree was derived. Because of the large size and the high degree of sequence similarity of these homologues, this tree is particularly reliable. The tree for the complete family, based on the multiple alignment shown at www-biology.ucsd.edu/~msaier/supmat, is presented in Fig. 3. This figure represents an expansion of two previously published trees exhibiting similar overall topology (Cao & Saier, 2001; Planet *et al.*, 2001). The tree includes 12 well-defined clusters. Cluster 1 has five subclusters designated A–E. All of these proteins likely function as secreton ATPases. Clusters 2, 3, 4, 5 and 6 include sequence divergent ATPases, all possibly concerned with bacterial pilus assembly/biogenesis/function. Clusters 7 and 8 A–C are all archaeal proteins; cluster 8B includes the recognized archaeal flagellar ATPases; clusters 9A–E and 10

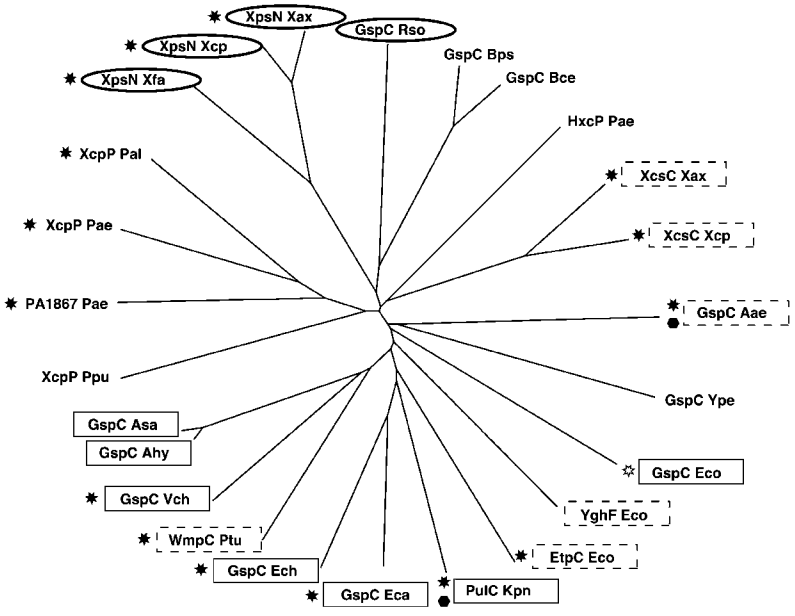

Fig. 2. Phylogenetic tree of secreton proteins in the T2SC family generated as in Fig. 1. Proteins with predicted PDZ or coiled-coil domains are indicated. The PROSITE signature pattern is: P-x(6)-F-x(4)-L-x(3)-D-[LIVM]-A-[LIVM]-x-[LIVM]-N-x-[LIVM]-x-L. Symbols are as follows: ▭, complete signature; ⬚, partial signature; ⬭, no signature; ✷, coil-coil at the end; ✡, coil-coil not at the end; ⬢, PDZ.

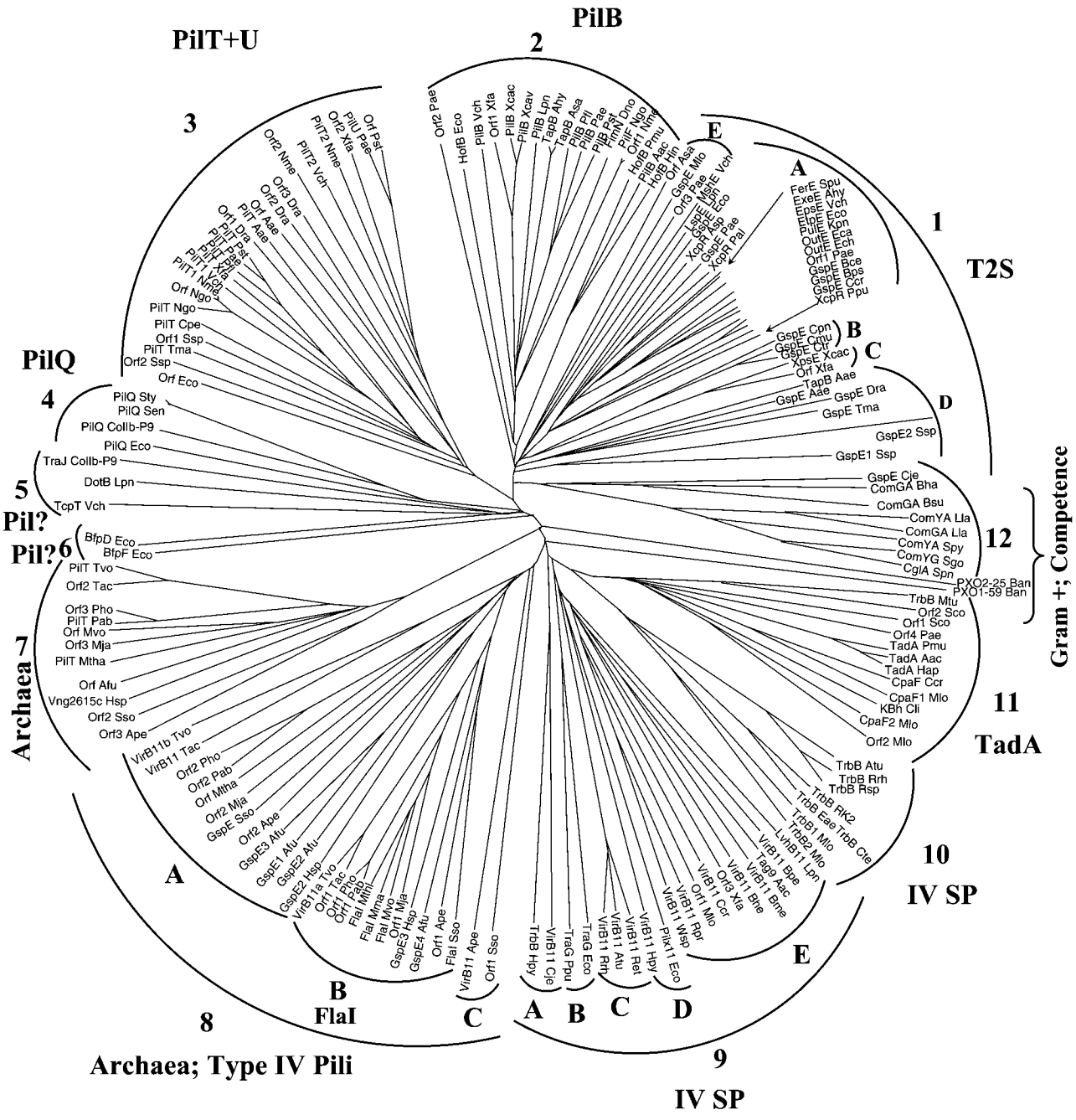

Fig. 3. Phylogenetic tree for the T2SE ATPases and their homologues. Clusters are labelled by number and family and are characterized further in Table 4.

are type IV protein secretion/conjugation ATPases (see Cao & Saier, 2001 and the Overview). Cluster 11 is a mixture of Gram-negative and high-G+C Gram-positive bacterial homologues including the TadA protein involved in tight adherence to surfaces in *Actinobacillus actinomycetemcomitans* (Kachlany *et al.*, 2000). Cluster 12 consists of competence-related proteins from low-G+C Gram-positive bacteria. Thus the primary clusters correspond to specific functional types, and most clusters are restricted to a particular type of organism (i.e. Gram-negative bacteria, Gram-positive bacteria or archaea). In many cases, the functions of the proteins are not known. Eukaryotic homologues were not found.

Table 4 summarizes the size differences (averaged, ± standard deviations) observed for the clusters depicted in Fig. 3. The proposed functions of these clustered proteins, usually based on functional assignments for just a few homologues, and their classification according to the organisms in which they are found, are also presented in

Table 4. Organismal representation, size variation and functional assignments for the phylogenetic clusters of the ATPase family

Cluster*	Organisms represented†	Mean size ± SD‡	(Putative) Function
1A	γ1, γ3, α, β	493 ± 15	Secreton/T2S
1B	Chla	499 ± 3	Secreton/T2S
1C	γ4	554 ± 33	Secreton/T2S
1D	Aqu, Dei, Th, CC	655 ± 150	Secreton/T2S
1E	γ1, γ3, α	551 ± 34	Secreton/T2S
2	γ1, γ2, γ3, γ4, β	542 ± 44	Pilus biogenesis (PilB)
3	γ1, γ3, γ4, β, Aqu, Dei, LG+	374 ± 51	Pilus biogenesis (PilT + U)
4	γ1	544 ± 31	Pilus biogenesis (PilQ)
5	γ1, γ3	420 ± 71	Pilus biogenesis (Pil?)
6	γ1	432 ± 144	Pilus biogenesis (Pil?)
7	Arc	582 ± 50	Archaea
8A	Arc	613 ± 213	Archaea; type IV pili
8B	Arc	562 ± 34	Archaea; type IV pili
8C	Arc	527 ± 74	Archaea; type IV pili
9A	ε	311 ± 7	Type IV secreton (IVSP)
9B	γ1, γ3	361 ± 30	Type IV secreton (IVSP)
9C	α	348 ± 16	Type IV secreton (IVSP)
9D	γ1	338 ± 11	Type IV secreton (IVSP)
9E	γ2, γ3, γ4, α	341 ± 12	Type IV secreton (IVSP)
10	α, γ1	323 ± 3	Type IV secreton (IVSP)
11	γ2, γ3, α, Chlo, HG+	452 ± 47	?
12	LG+	354 ± 69	Competence

*Cluster refers to the phylogenetic cluster shown in Fig. 3.

†All Greek letter entries refer to the category of proteobacteria from which the proteins are derived. Other abbreviations are: Aqu, *Aquifex*; Chla, *Chlamydia*; Dei, *Deinococcus*; Th, *Thermatoga*; CC, cyanobacteria/chloroplasts; Chlo, *Chlorobium*; LG+, low-G+C Gram-positive bacteria; HG+, high-G+C Gram-positive bacteria; Arc, archaea.

‡Sizes are expressed in numbers of amino acyl residues ± SD.

Table 4. Of greatest interest is the fact that most of the 12 clusters (subdivided in the cases of clusters 1, 8 and 9) consist of proteins with a characteristic and very restricted size range. The one exception is cluster 6, which consists of two sequence-divergent *E. coli* proteins that in actuality branch from points near the centre of the tree. It is therefore clear that both protein size and function consistently correlate with phylogeny, and often with organismal type as well.

The phylogenetic clusters shown in Fig. 3 were examined to determine if the phylogenies of the proteins within any of the clusters follow those of the 16S rRNAs. Most of these clusters include two or more paralogues from a single organism, and, in general, protein phylogeny does not follow that of the 16S rRNAs (see the online version of this review at http://mic.sgmjournals.org for the bacterial 16S rRNA tree). For example, in the secreton ATPases of cluster 1A, two paralogues are each observed for *E. coli* and *P. aeruginosa*. However, in clusters 7, 8A and 8B, each cluster displays only one protein from each archaeon with a fully sequenced genome, and these follow the phylogenies of the organisms fairly closely (see Fig. 4 for the archaeal 16S rRNA tree). Thus, with the possible exception of the archaeal homologues, no cluster can be considered to consist exclusively of orthologues. We predict that horizontal transfer as well as late gene duplication events were responsible for the configurations of the portions of the tree that include bacterial homologues.

T2SF/T4PC/FlaJ transmembrane (TM) protein phylogeny

The T2SF/T4PC/FlaJ integral membrane putative multispanning TM protein family tree is shown in Fig. 5. The proteins fall into seven clusters which are labelled according to the probable corresponding cluster designations presented in Fig. 3 for the ATPases. Cluster 1 includes almost all currently recognized secreton constituents, and most of these proteins are derived from proteobacteria of the α-, β- and γ-subclasses (cluster 1A). Paralogues in this cluster are found only for *E. coli* and *P. aeruginosa* as was observed for the ATPases (Fig. 3). Such observations suggest that each ATPase functions with a TM protein. The chlamydial proteins (cluster 1B) and the *Xylella fastidiosa* and *Xanthomonas campestris* proteins (cluster 1C) compose two additional but divergent groups, as observed for the

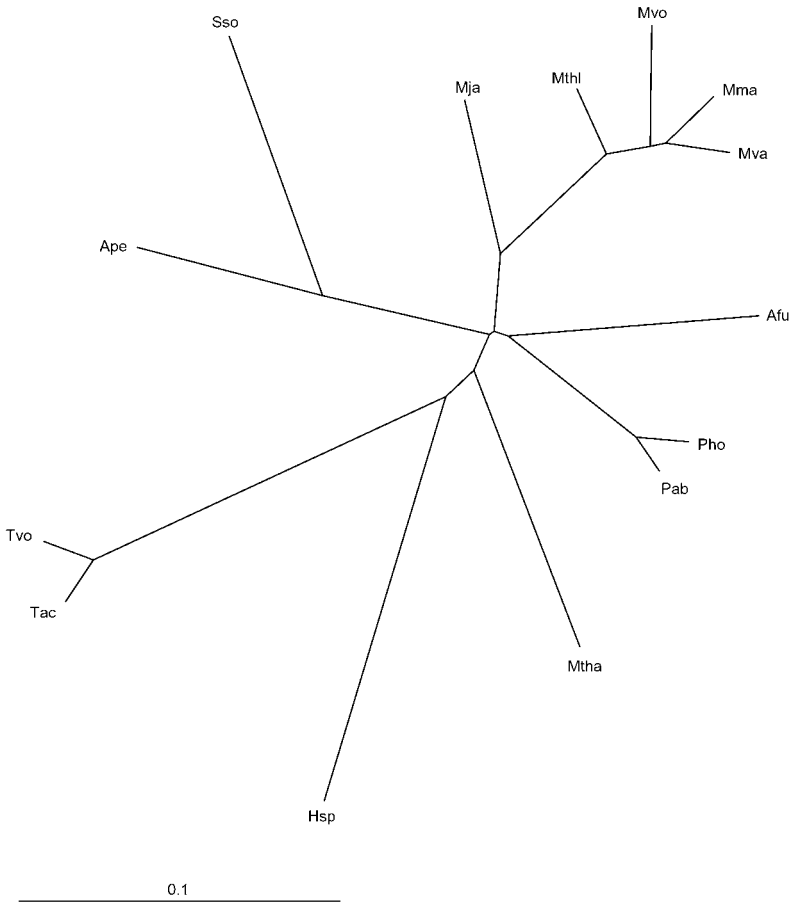

Fig. 4. Phylogenetic tree for 16S rRNAs from the archaea represented in Fig. 3.

ATPases shown in Fig. 3. Additional proteins included in cluster 1 are sequence divergent proteins from *Vibrio cholerae* and *P. aeruginosa*, both γ-proteobacteria (cluster E), and several proteins from other phylogenetically divergent bacteria (*Aquifex*, *Thermatoga*, *Deinococcus* and *Synechocystis*) (clusters D1 and D2). Corresponding proteins are found in cluster 1 of Fig. 3.

Cluster 2, to the left of cluster 1, is a large group of sequence divergent pilus-related proteins, all of which are from proteobacteria, and most of which are from the γ-subclass. Progressing around the tree in the clockwise direction, cluster 12 consists of competence proteins from low-G+C Gram-positive bacteria. Corresponding ATPases are found in cluster 12 of Fig. 3. Cluster 4 consists of a distinct group of γ-proteobacterial homologues, probably all concerned with pilus function or biogenesis. Clusters 11A and 11B represent two other primarily proteobacterial clusters with distant homologues from the high-G+C Gram-positive bacterium *Streptomyces coelicolor*. Included within these two clusters are the TadB and TadC homologues from the disease-causing Gram-negative bacterium *Actinobacillus actinomycetemcomitans*. Both of these two proteins are concerned with tight adherence to surfaces (Kachlany *et al.*, 2000). The corresponding ATPase, TadA, is found in cluster 11 in Fig. 3.

It is interesting to note that the two integral membrane proteins TadB and TadC apparently function with a single ATPase, TadA. This observation might suggest that these integral membrane constituents function as homo- or hetero-oligomeric (possibly dimeric) structures. A dimeric structure would be in agreement with the fact that the archaeal homologues are internally duplicated proteins, twice as large as the bacterial homologues with approximately twice as many TMSs (see next section).

Cluster 8 includes proteins derived exclusively from archaea. The functions of cluster 8A and cluster 8D proteins are unknown, and they may or may not have counterparts in Fig. 3. Cluster 8A in Fig. 3 includes only one protein per organism, except for *Archaeoglobus fulgidus* where three paralogues are found. However, in cluster 8A of Fig. 5, two sets of homologues are found for most represented organisms. Most of the proteins in cluster 8B are probably constituents of archaeal flagellar systems. These proteins are represented only once per organism, have counterparts in Fig. 3, and exhibit phylogenetic relationships that reflect those of the 16S rRNAs (compare Figs 3, 4 and 5). These proteins are therefore likely to be orthologues with a single ATPase per TM protein. It should be noted that many clusters of ATPases found in Fig. 3 are not represented in Fig. 5. These ATPases probably function in a process and by a mechanism that is independent of a multispanning TM protein homologue. Alternatively, they may act with

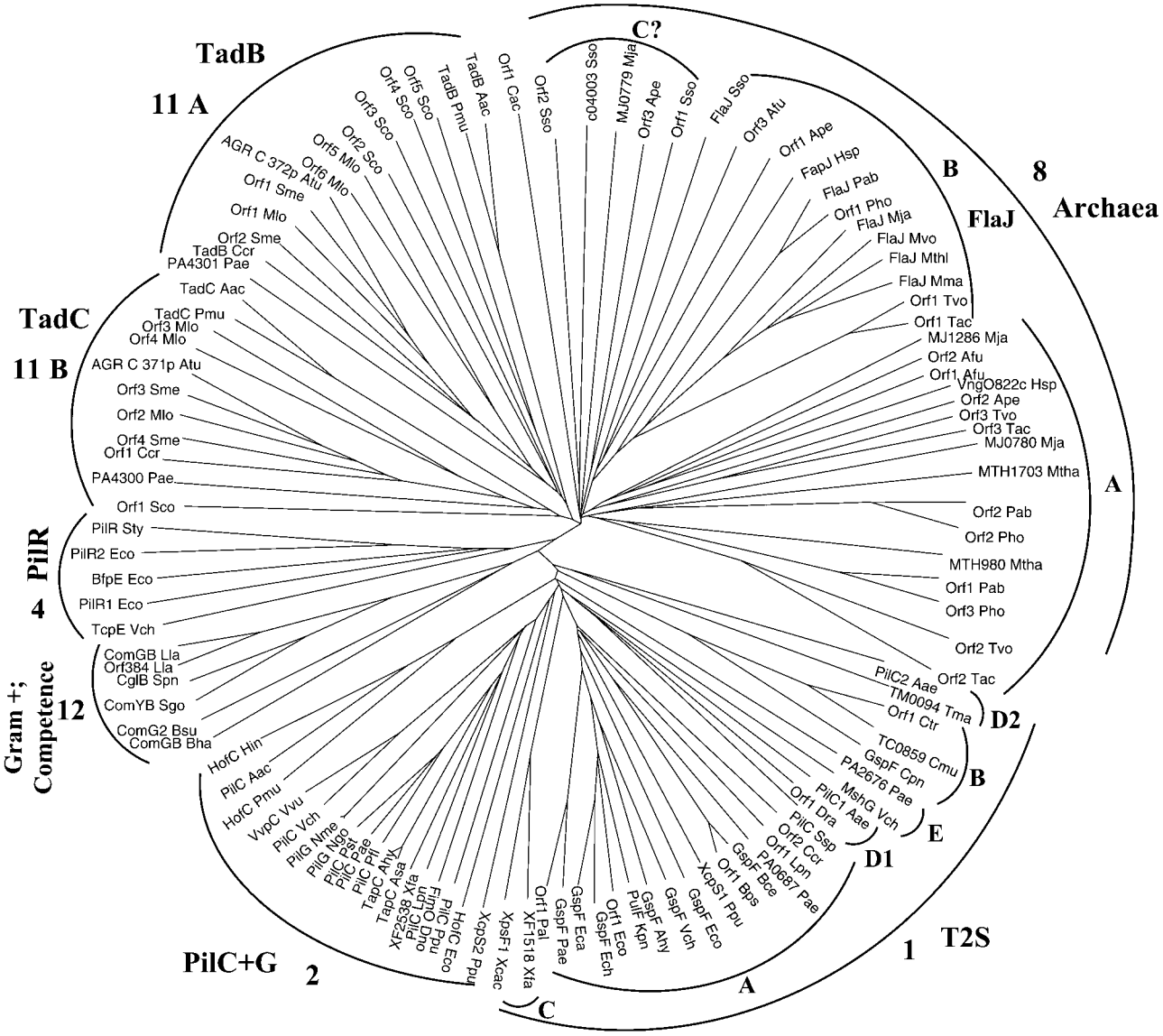

Fig. 5. Phylogenetic tree for the T2SF integral membrane (TM) proteins and their homologues. Clusters are labelled as in Fig. 3 and are characterized further in Table 5. These clusters are believed to correspond to the phylogenetic clusters of the same number and letter shown for the ATPases in Fig. 3.

multispanning TM proteins that are too divergent in sequence for us to recognize.

The data summarized in Table 5 reveal that, as for the ATPases, each major cluster of multispanning TM proteins exhibits its own characteristic size range. However, there is no direct or inverse size relationship between the ATPases tabulated in Table 4 and the membrane proteins tabulated in the corresponding clusters in Table 5.

Multispanning TM protein topologies and the occurrence of internal repeats

The multispanning TM proteins were examined with respect to their apparent topologies. At least three major hydrophobic peaks were identified in all proteins using several different algorithms. The experimentally determined topology of one member of the T2SF family, OutF from *Erwinia carotovora* (Thomas *et al.*, 1997), indicates three TM segments, as predicted by most hydrophobicity algorithms. However, it should be noted that the second predicted hydrophobic segment of these proteins is rather long and that some algorithms predict that it is extended to form two TM segments (TMSs), making four in all. One such protein is GspF from *E. coli* K-12 ($GspF^{Eco}$), which was predicted to have four TMSs by all algorithms tested. The distribution of positively charged amino acids at each end of the hydrophobic segments beyond TMS1 did not allow us to predict their orientation in the membrane according

Table 5. Organismal representation, size variation and functional assignments for the phylogenetic clusters of the PilC integral membrane (TM) protein family

Cluster*	Organisms represented†	Mean size ± SD‡	No. of TMS§	Function‖	Corresponding ATPase¶
1A	γ1, γ3, α, β, CC	403 ± 5	4	Secreton constituent	T2SE
1B	Chla	391 ± 0	4	Secreton constituent	T2SE
1C	γ4	398 ± 11	4	Secreton constituent	T2SE
$1D_1$	Aqu, Dei	407 ± 1	4	Secreton constituent	T2SE
$1D_2$	Aqu, Th	397 ± 4	4	Secreton constituent	T2SE
1E	γ1, γ3	379 ± 23	4	Secreton constituent	T2SE
2	γ1, γ2, γ3, γ4, β	406 ± 24	4	Pilus biogenesis (PilC + G)	Pilus biogenesis (PilB)
4	γ1	357 ± 12	4	Pilus biogenesis (PilR)	Pilus biogenesis (PilQ)
8A	Arc	380 ± 137	7–11	Function unknown	?
8B	Arc	551 ± 28	7–9	Archaeal flagellum (FlaJ)	FlaI
8C?	Arc	459 ± 196	3–10	Function unknown	?
11A	γ2, γ3, α, HG+	319 ± 16	4–6	Tight adherence to surface (TadB)	TadA
11B	γ2, γ3, α, HG+	311 ± 25	4–5	Tight adherence to surface (TadC)	TadA
12	LG+	320 ± 30	4	Competence	ComGA

*Cluster refers to the phylogenetic clusters shown in Fig. 5. These are believed to correspond to the ATPase clusters of the same number presented in Fig. 3 and Table 4. Several clusters shown in Fig. 3 (clusters 3, 5–7, 9 and 10) could not be conclusively identified in Fig. 5.

†All Greek letter entries refer to the category of proteobacteria from which the proteins are derived. Other abbreviations are as indicated in footnote † to Table 2.

‡Sizes are expressed in numbers of amino acyl residues ± SD.

§Number of putative α-helical transmembrane segments.

‖The probable functions of the proteins in each cluster are based on one or more characterized member(s) of these clusters.

¶Protein(s) in the corresponding phylogenetic cluster shown in Fig. 3.

to the 'positive-inside rule' (von Heijne & Gavel, 1988). Furthermore, the *gspF* gene exhibited very low complementation of a deletion of the *pulF* gene in the *pul* gene cluster (10–25 % restoration of pullulanase secretion in four independent experiments; data not shown) compared with $pulF^{Kox}$ and $outF^{Ech}$ (> 80 % restoration under identical conditions; Possot *et al.*, 2000). Whether or not $GspF^{Eco}$ differs from all other T2SF proteins remains to be determined. However, if $GspF^{Eco}$ topology is partially reversed compared to that of other T2SF proteins, a well-conserved segment predicted to be cytoplasmic in OutF (Thomas *et al.*, 1997) would be located in the periplasm in $GspF^{Eco}$, which seems unlikely.

The topology of T2SF homologues in the T4P system is equally unclear. The experimentally determined protein topology of one member of this group, the *E. coli* protein BfpE, a member of the PilR cluster of pilus-related proteins in Fig. 5, gives four TMSs, with TMSs 2 and 3 being nearly contiguous (Blank & Donnenberg, 2001). Indeed, most algorithms predict the same four TMSs in this protein (not shown). However, as with $GspF^{Eco}$, the topology of BfpE cannot be predicted from the positive-inside rule. Furthermore, most algorithms predict three TMSs for the closely related *V. cholerae* protein, TcpE, and three to five TMSs for other proteins in the PilR cluster. Topological predictions for the PilC/G cluster also indicate three and five TMSs, with three TMSs most frequently predicted for the archetypal protein of this cluster, *P. aeruginosa* PilC (Fig. 6). Once again, a large domain that is predicted to be periplasmic by the 3 TMS model is predicted to be cytoplasmic by the 4 TMS model, as discussed above for T2SF proteins. Although not as highly conserved as the corresponding segment of the T2SF proteins, this region of almost all proteins under consideration (including the T2SF proteins) contains several highly or absolutely conserved residues. The possible exceptions are all in the PilR cluster, including TcpE and BfpE, in which only some of these highly conserved residues are present. Furthermore, these regions of TcpE and BfpE are almost totally unrelated, which is in marked contrast, for example, to proteins in the T2SF cluster. Therefore, in contrast to $GspF^{Eco}$ noted above, it is quite conceivable that BfpE does have a topology different from the predicted 3 TM proteins such as PilC and TcpE.

Interestingly, TMS1 in proteins in the T2SF-PilC/G-PilR clusters includes an almost invariant motif: K– – –A – Y P – – – – – V A – V– – – $(L\ M)_2$ – – V V P. Note that the presence of an invariant lysine at the start (cytoplasmic side) of this motif/TMS is as expected from the positive-inside rule. Multiple alignments of these proteins showed that the short, downstream (periplasmically oriented) segment between TMS1 and the next hydrophobic segment is of identical length and includes a motif (LP–T) that is also almost universally conserved. However, the beginning

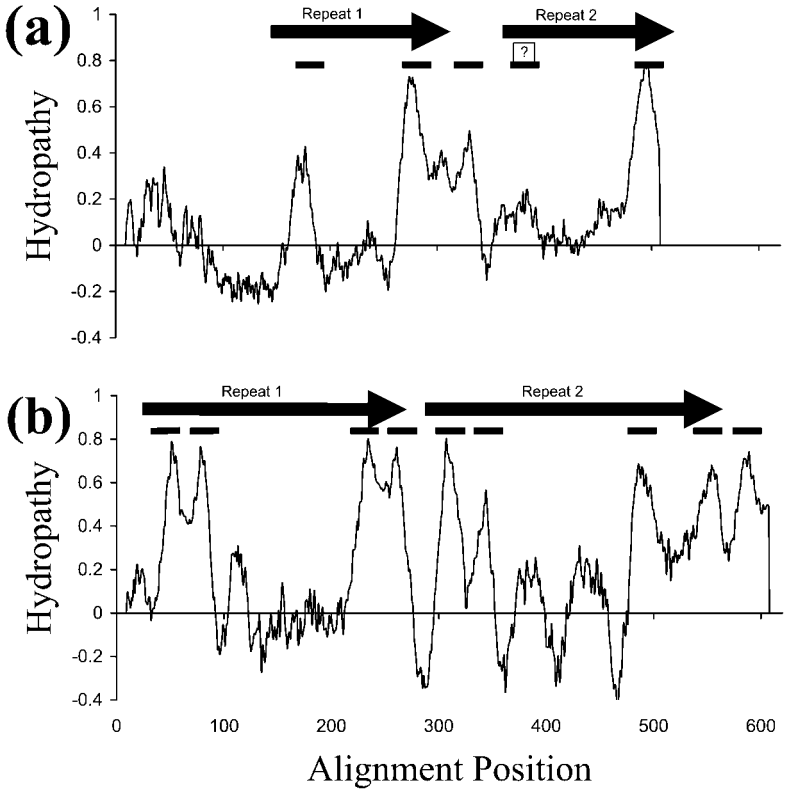

Fig. 6. Mean hydropathy plots for (a) the bacterial homologues of *P. aeruginosa* PilC and (b) the archaeal homologues of *M. voltae* FlaJ. The plots were derived using the AveHAS program (Zhai & Saier, 2001a). The black arrows above the plots show the positions of the repeat sequences while the bars below these arrows show the positions of putative transmembrane α-helical segments.

of the next hydrophobic segment (predicted to be TMS2 by all algorithms) is not highly conserved, although the C-terminal end of this hydrophobic segment (predicted as TMS3 in BfpE and $GspF^{Eco}$) is conserved.

The GAP program (Devereux *et al.*, 1984) predicted that regions within the first halves of the TM proteins are homologous to regions in the second halves. One such alignment for *P. aeruginosa* PilC is shown in Fig. 7(a). This 91 residue binary comparison shows 33 % identity and 55 % similarity with an e value of 3×10^{-8}. Comparison scores of 23–25 SD for these portions of the two halves of PilC and of 10 SD for corresponding portions of the two halves of PulF of *K. pneumoniae* were obtained. These values are sufficient to establish that the two halves of these proteins arose from a common origin, probably by an internal gene duplication event. Interestingly, part of this duplicated region includes a diagnostic motif for members of the T2SF-PilC/G-PilR protein clusters, while other residues are well conserved in all clusters (Fig. 7).

The corresponding hydropathy plot for the archaeal flagellar homologues (the FlaJ proteins) is shown in Fig. 6(b). Nine peaks of hydropathy (black bars) are evident. Peaks 1–4 appear similar in arrangement to peaks 5–8. Indeed, comparison scores using the GAP program suggested that these proteins also arose by an internal duplication event since, if TMS9 is omitted, the first halves appear to be homologous to the second halves. Thus, in comparing the most conserved portions of the first halves of these proteins with those of the second halves, comparison scores of up to 8 SD were obtained. A value of 8 SD corresponds to a probability of 10^{-15} that this degree of similarity arose by chance. This value is strongly suggestive of homology. Alignments of a bacterial TM protein (C372 of *Agrobacterium tumefaciens*) with the two halves of an archaeal TM protein (FlaJ of *M. voltae*) are shown in Fig. 7(b, c). These alignments were generated using the PSI-BLAST program with two iterations (Altschul *et al.*, 1997). The conserved residues in each of the archaeal halves thus represent a matrix of conserved motifs between each of the archaeal halves and the bacterial protein.

Based on these results, we suggest that the smaller bacterial proteins arose by one internal gene duplication event, and that a second internal gene duplication event gave rise to the larger archaeal homologues. Judging from the relative degrees of sequence similarity, however, the bacterial duplication event(s) probably occurred after the archaeal duplication event(s). This fact suggests that these duplication events have occurred more than once during the evolution of this family of TM proteins. A similar situation has been observed for other transmembrane protein families (Tseng *et al.*, 1999). It should be noted that the topologies of these proteins can not be deduced with certainty from the hydropathy plots alone.

Conclusions

In this review, we have attempted to provide phylogenetic data and conduct sequence comparisons that allow evolutionary conclusions that interrelate the bacterial T2S

(a)

```
                        m                                                m
                   trqlatl  ag pl                              v  g  l   l
PilC Pae1:  36 IALFTRQMATMMGAGVPLLQSFDIIGEGFDNPNMRKLVDEIKQEVSSGNSLANSLRKKPQ  95
               +A + R ++T   AGVPL+ + D +     N   +  V +IKQ+VS+G  L  S+R
PilC Pae2: 240 VARYARTLSTTFAAGVPLVDALDSVSGATGNIVFKNAVSKIKQDVSTGMQLNFSMR-TTS 298
                ar artl  l  agvpl  al        n             v  g      m

                f  ly   v age  g l   l   a y e
PilC Pae1:  96 YFDELYCNLVDAGEQSGALENLLDRVATYKEK 127
                F  +   +   GE+SG+L+ +L +VA+Y E+
PilC Pae2: 299 VFPNMAIQMTAIGEESGSLDEMLSKVASYYEE 330
                fp     m   ge sg ld ml   a
```

(b)

```
FlaJ Mvo:   4 DILPRVGLKPKDYFLKFVLPAVLASLFMVVLGFIYFDGITRLLVLALPLLLLGGALGYPY  63
              D L + GL       +F L +  A+     +  + +       +    +  LG
C372 Atu:  93 DKLVQAGLTIS--VRQFHLLSAGAAACAAFI-SLIYGLSLLTALPIAVVAALGLPRWVLA 149

FlaJ Mvo:  64 IELDSQKQKINERLHVYITKFGVLSITDLDNKALLELLSVEKEELGQLAEESRKIYVLVK 123
              L  +++K  E     +        + L     + L++ + +E   +  E +++    +
C372 Atu: 150 FLLKRRQKKFLEEFPNALDVMCRSIKSGLPLNDAVRLIASDGQE--PVKTEFQRVV-DAQ 206

FlaJ Mvo: 124 RWNQSLAGSCRFLANRTPSSQFGDFLDRMAYSIDSGQELKEFLAGEQDIVMDEYAGFYER 183
              +    +      +    P  +   F   +     +G  L E L+    ++ D      +
C372 Atu: 207 QVGIGIPQGIERMMLTMPLFEVSFFSIVINIQAQAGGNLSEALSNLSRVLRDRRKMRAKV 266

FlaJ Mvo: 184 ALYSLDNFKEMYVSAITSVSFFVTFAI----IAPFLLPYDFVTMVTVAIFIFMIIEVILI 239
                S++      +          +            L       ++  A  I+M+I + ++
C372 Atu: 267 NALSMEAKASAVIIGALPFIVMLLVHFTSPDYLSILFTDTRGHLILGASAIWMLIGIFIM 326
```

(c)

```
FlaJ Mvo: 262 DRKLRKWLIISVGLTILASIVLFWGKYIYEAPQLLKIPYELIFSIAMTPLMLGGYMAQRE 321
              D+ ++  L ISV    L S                L +   L  ++     +    +A
C372 Atu:  93 DKLVQAGLTISVRQFHLLSAGAAACAAFISLIYGLSLLTALPIAVVAALGLPRWVLAFLL 152

FlaJ Mvo: 322 ESLVIRKENNFPDFLRSLGDSVSAKGGGTLESLGYLCTNDFGPLTKDLVALHRRLSIRIN 381
              +    +    FP+ L  +  S+   G    +++  + ++   P+  +   +     + I
C372 Atu: 153 KRRQKKFLEEFPNALDVMCRSI-KSGLPLNDAVRLIASDGQEPVKTEFQRVVDAQQVGIG 211

FlaJ Mvo: 382 GQKSWKYFGHDTCSYLIQLFSEMYERCTYLGGNSGQASHIIGKNFRKILRLRRSKYQNVN 441
                +  +        + +  FS +       GGN  +A   + +  R     RR     VN
C372 Atu: 212 IPQGIERMMLTMPLFEVSFFSIVINIQAQAGGNLSEALSNLSRVLRD----RRKMRAKVN 267

FlaJ Mvo: 442 QFAGVMYGLSGGMALTLFASYGVASMVNGLYSSLDIPDTMLSMVHVVAPSDFGFISYMMY 501
                +      +  +    F    +    +  Y S+   DT   ++   +      I   +
C372 Atu: 268 ALSMEAKASAVIIGALPFIVMLLVHFTSPDYLSILFTDTRGHLILGASAIWM-LIGIFIM 326
```

Fig. 7. Binary sequence alignments revealing the sequence similarities between (a) the two repeat sequences within a representative bacterial multispanning TM protein, the PilC protein of *P. aeruginosa*, (b) the first half of the archaeal FlaJ protein of *M. voltae* and second half of the bacterial C372 protein of *A. tumefaciens* and (c) the second half of the archaeal FlaJ protein of *M. voltae* and the bacterial C372 protein of *A. tumefaciens*. Residue number in each protein is provided at the beginning and end of each line. Residues identical in the two sequences are shown between the binary alignments while conservative substitutions are displayed with a '+'. In Fig. 7(a), residues indicated in lower-case characters above and below the alignment are amino acids that are highly conserved in the first and second halves, respectively, of most members of the T2SF-PilC/G-PilR clusters of proteins.

and T4P systems as well as the archaeal Fla type systems. All are believed to consist, in part, of type IV pili or pilus-like filaments. The following (tentative) conclusions resulted from our analyses. (1) Although the protein constituents of the T2S, T4P and Fla systems are widely distributed between the bacterial and archaeal domains, not a single homologue of any of these proteins was identified in a eukaryote. Thus, although horizontal transfer between bacterial species may have occurred repeatedly, no horizontal transfer has occurred between prokaryotes and eukaryotes. Moreover, since the archaeal proteins are always sequence divergent from the bacterial proteins, no horizontal transfer between bacteria and archaea has occurred recently, probably within the past 2 billion years. (2) Among the secreton systems of various bacteria, we provided evidence that the constituents of these systems have evolved from those of a single primordial

precursor system without shuffling of constituents between systems (see Fig. 1). This conclusion suggests the importance of strict protein–protein interactions for the maintenance of proper function. Similar conclusions have been reached for other protein secretion systems as well as other types of multicomponent transporters (Cao & Saier, 2001; Kuan *et al.*, 1995; Nguyen *et al.*, 2000). The implications of these findings have recently been discussed (Saier, 2003a). (3) The protein constituents of the various systems analysed have undergone sequence divergence at different rates, where the ATPases and multispanning TM proteins are among the largest and were the most slowly diverging. These facts allowed the most reliable phylogenetic analyses with these constituents and also facilitated homologue identification. (4) Between T2S and T4P systems, four proteins [the ATPases (T2SE homologues), the multispanning TM proteins (T2SF homologues), the secretins (T2SD homologues) and the prepilin peptidases (T2SO homologues)] proved to be demonstrably homologous. However, only three of these constituents proved to be demonstrably homologous between these bacterial systems and the archaeal Fla systems. Secretins are apparently lacking in the archaeal Fla systems as expected since archaea lack an outer membrane. (5) When the two most conserved families, those of the ATPases and the multispanning TM proteins, were analysed phylogenetically, it became clear that there is not a one-to-one correspondence between these two constituents. Some phylogenetic clusters of the ATPases appear to function with TM proteins while others either do not, or they utilize a TM protein that is too divergent in sequence to be recognized using the search tools employed. Nevertheless, we could identify clusters of ATPase that consistently function with TM proteins in both bacteria and archaea. (6) Size comparisons of proteins within both the ATPase clusters and the multispanning TM protein clusters (Tables 4 and 5) revealed that each phylogenetic cluster exhibits a characteristic size range. However, the sizes of the ATPase and TM protein homologues do not correlate with each other either directly or inversely. We propose that the members of each cluster serve a unified function. This last suggestion is substantiated by limited experimental data (see Tables 4 and 5). (7) Information about the evolution of the multispanning TM proteins was forthcoming. Thus the bacterial homologues clearly arose by an internal gene duplication event, and the same is probably true of the archaeal homologues. However, these duplication events were not the same, suggesting that gene duplication occurred at least twice during the evolution of this protein family. The same conclusion has been reached for some integral membrane transport protein families but not for others (Kuan & Saier, 1993; Pao *et al.*, 1998; Saier, 1999a, b, 2000, 2001a, b, 2003b).

This report presents sequence comparisons that allow us to establish relationships between several of the protein constituents of the T2S, T4P and Fla systems of bacteria and archaea. Further analyses will be required to establish the functional significance of many of the provocative observations made here.

Acknowledgements

Amplified *gspF* was generously provided by Guillaume Vignon. The work in the laboratory of M. H. S. was supported by NIH grants GM 64368 and GM 55434. Work in the laboratory of A. P. P. was supported in part by grant number HPRN-CT-2000-00075 from the EC. We thank Mary Beth Hiller for assistance in the preparation of this manuscript.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Ast, V., Schoenhofen, I., Langen, G., Stratilo, C., Chamberlain, M. & Howard, S. (2002). Expression of the ExeAB complex of *Aeromonas hydrophila* is required for the localization and assembly of the ExeD secretion port multimer. *Mol Microbiol* **44**, 217–231.

Ball, G., Durand, E., Lazdunski, A. & Filloux, A. (2002). A novel type II secretion system in *Pseudomonas aeruginosa*. *Mol Microbiol* **43**, 475–485.

Bardy, S. & Jarrell, K. (2002). FlaK of the archaeon *Methanococcus maripaludis* possesses preflagellin peptidase activity. *FEMS Microbiol Lett* **208**, 53–59.

Bitter, W., Koster, M., Latijnhouwers, M., de Cock, H. & Tommassen, J. (1998). Formation of oligomeric rings by XcpQ and PilQ, which are involved in protein transport across the outer membrane of *Pseudomonas aeruginosa*. *Mol Microbiol* **27**, 209–219.

Blank, T. & Donnenberg, M. (2001). Novel topology of BfpE, a cytoplasmic membrane protein required for type IV fimbrial biogenesis in enteropathogenic *Escherichia coli*. *J Bacteriol* **183**, 4435–4450.

Bleves, S., Lazdunski, A., Tommassen, J. & Filloux, A. (1998). The secretion apparatus of *Pseudomonas aeruginosa*: identification of a fifth pseudopilin, XcpX. *Mol Microbiol* **27**, 31–40.

Bleves, S., Gérard-Vincent, M., Lazdunski, A. & Filloux, A. (1999). Structure-function analysis of XcpP, a component involved in general secretory pathway-dependent protein secretion in *Pseudomonas aeruginosa*. *J Bacteriol* **181**, 4012–4019.

Brok, R., Van Gelder, P., Winterhalter, M., Ziese, U., Koster, A. J., de Cock, H., Koster, M., Tommassen, J. & Bitter, W. (1999). The C-terminal domain of the *Pseudomonas* secretin XcpQ forms oligomeric rings with pore activity. *J Mol Biol* **294**, 1169–1179.

Cao, T. B. & Saier, M. H., Jr (2001). Conjugal type IV macromolecular transfer systems of Gram-negative bacteria: organismal distribution, structural constraints and evolutionary conclusions. *Microbiology* **147**, 3201–3214.

Cao, T. B. & Saier, M. H., Jr (2003). The general protein secretory pathway: phylogenetic analyses leading to evolutionary conclusions. *Biochim Biophys Acta* **1609**, 115–125.

Collins, R. F., Davidsen, L., Derrick, J. P., Ford, R. C. & Tønjum, T. (2001). Analysis of the PilQ secretin from *Neisseria meningitidis* by transmission electron microscopy reveals a dodecameric quaternary structure. *J Bacteriol* **183**, 3825–3832.

Collins, R. F., Ford, R. C., Kitmitto, A., Olsen, R. O., Tonjum, T. & Derrick, J. P. (2003). Three-dimensional structure of the *Neisseria*

meningitidis secretin PilQ determined from negative-strain transmission electron microscopy. *J Bacteriol* **185**, 2611–2617.

Condemine, G. & Shevchik, V. (2000). Overproduction of the secretin OutD suppresses the secretion defect of an *Erwinia chrysanthemi outB* mutant. *Microbiology* **146**, 639–647.

Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). Jpred: a consensus secondary structure prediction server. *Bioinformatics* **14**, 892–893.

d'Enfert, C., Reyss, I., Wandersman, C. & Pugsley, A. P. (1989). Protein secretion by gram-negative bacteria. Characterization of two membrane proteins required for pullulanase secretion by *Escherichia coli* K12. *J Biol Chem* **264**, 17462–17468.

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

Drake, S. L., Sandstedt, S. A. & Koomey, M. (1997). PilP, a pilus biogenesis lipoprotein in *Neisseria gonorrhoeae*, affects expression of PilQ as a high-molecular-mass multimer. *Mol Microbiol* **23**, 657–668.

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

Filloux, A., Michel, G. & Bally, M. (1998). GSP-dependent protein secretion in Gram-negative bacteria: the Xcp system of *Pseudomonas aeruginosa*. *FEMS Microbiol Rev* **22**, 177–198.

Francetic, O., Belin, D., Badaut, C. & Pugsley, A. P. (2000). Expression of the endogenous type II secretion pathway in *Escherichia coli* leads to chitinase secretion. *EMBO J* **19**, 6697–6703.

Gerard-Vincent, M., Robert, V., Ball, G., Bleves, S., Michel, G., Lazdunski, A. & Filloux, A. (2002). Identification of XcpP domains that confer functionality and specificity to the *Pseudomonas aeruginosa* type II secretion apparatus. *Mol Microbiol* **44**, 1651–1665.

Hardie, K. R., Lory, S. & Pugsley, A. P. (1996a). Insertion of an outer membrane protein in *Escherichia coli* requires a chaperone-like protein. *EMBO J* **15**, 978–988.

Hardie, K. R., Seydel, A., Guilvout, I. & Pugsley, A. P. (1996b). The secretin-specific, chaperone-like protein of the general secretory pathway: separation of proteolytic protection and piloting functions. *Mol Microbiol* **22**, 967–976.

Hobbs, M. & Mattick, J. S. (1993). Common components in the assembly of type 4 fimbriae, DNA transfer systems, filamentous phage and protein-secretion apparatus: a general system for the formation of surface-associated protein complexes. *Mol Microbiol* **31**, 1596–1600.

Hofmann, K. & Stoffel, W. (1993). Tmbase – a database of membrane spanning protein segments. *Biol Chem* **347**, 166.

Hu, N. T., Hung, M. N., Liao, C. T. & Lin, M. H. (1995). Subcellular location of XpsD, a protein required for extracellular protein secretion by *Xanthomonas campestris* pv. *campestris*. *Microbiology* **141**, 1395–1406.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33**, 3038–3049.

Kachlany, S. C., Planet, P. J., Bhattacharjee, M. K., Kollia, E., DeSalle, R., Fine, D. H. & Figurski, D. H. (2000). Nonspecific adherence by *Actinobacillus actinomycetemcomitans* requires genes widespread in *bacteria* and *archaea*. *J Bacteriol* **182**, 6169–6176.

Krause, S., Bárcena, M., Pansegrau, W., Lurz, R., Carazo, J.-M. & Lanka, E. (2000). Sequence-related protein export NTPases encoded by the conjugative transfer region of RP4 and by the *cag* pathogenicity island of *Helicobacter pylori* share similar hexameric ring structures. *Proc Natl Acad Sci U S A* **97**, 3067–3072.

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

Kuan, G., Dassa, E., Saurin, W., Hofnung, M. & Saier, M. H., Jr (1995). Phylogenic analyses of the ATP-binding constituents of bacterial extracytoplasmic receptor-dependent ABC-type nutrient uptake permeases. *Res Microbiol* **146**, 271–278.

Le, T., Tseng, T.-T. & 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 Membr Biol* **16**, 173–179.

Lee, H.-M., Wang, K.-C., Liu, Y.-L., Yew, H.-Y., Chen, L.-Y., Leu, W.-M., Chen, D. C. & Hu, N.-T. (2000). Association of the cytoplasmic membrane protein XpsN with the outer membrane protein XpsD in the type II protein secretion apparatus of *Xanthomonas campestris* pv. campestris. *J Bacteriol* **182**, 1549–1557.

Lee, H.-M., Tyan, S.-W., Leu, W.-M., Chen, L.-Y., Chen, D. C. & Hu, N.-T. (2001). Involvement of the XpsN protein in formation of the XpsL-XpsM complex in *Xanthomonas campestris* pv. campestris type II secretion apparatus. *J Bacteriol* **183**, 528–535.

Letellier, L., Howard, S. P. & Buckley, T. J. (1997). Studies on the energetics of proaerolysin secretion across the outer membrane of *Aeromonas* sp: evidence for requirement for both the protonmotive force and ATP. *J Biol Chem* **272**, 11109–11113.

Martinez, A., Ostrovsky, P. & Nunn, D. N. (1998). Identification of an additional member of the secretin superfamily of proteins in *Pseudomonas aeruginosa* that is able to function in type II protein secretion. *Mol Microbiol* **28**, 1235–1246.

Merz, A. J., So, M. & Sheetz, M. P. (2000). Pilus retraction powers bacterial twitching motility. *Nature* **407**, 98–101.

Michel, G., Bleves, S., Ball, G., Lazdunski, A. & Filloux, A. (1998). Mutual stabilization of the XcpZ and XcpY components of the secretory apparatus in *Pseudomonas aeruginosa*. *Microbiology* **144**, 3379–3386.

Nguyen, L., Paulsen, I. T., Tchieu, J., Hueck, C. J. & Saier, M. H., Jr (2000). Phylogenetic analyses of the constituents of type III protein secretion systems. *J Mol Microbiol Biotechnol* **2**, 125–144.

Nouwen, N., Ranson, N., Saibil, H., Wolpensinger, B., Engel, A., Ghazi, A. & Pugsley, A. P. (1999). Secretin PulD: association with pilot protein PulS, structure and ion-conducting channel formation. *Proc Natl Acad Sci U S A* **96**, 8173–8177.

Nouwen, N., Stahlberg, H., Pugsley, A. P. & Engel, A. (2000). Domain structure of secretin PulD revealed by limited proteolysis and electron microscopy. *EMBO J* **19**, 2229–2236.

Nunn, D. (1999). Bacterial type II protein export and pilus biogenesis: more than just homologies? *Trends Cell Biol* **9**, 402–408.

Nunn, D. & Lory, S. (1992). Components of the protein excretion apparatus of *Pseudomonas aeruginosa* are processed by the type IV prepilin peptidase. *Proc Natl Acad Sci U S A* **89**, 47–51.

Nunn, D. N. & Lory, S. (1993). Cleavage, methylation and localization of the *Pseudomonas aeruginosa* export proteins XcpT, -U, -V and -W. *J Bacteriol* **175**, 4375–4382.

Nunn, D., Bergman, S. & Lory, S. (1990). Products of three accessory genes, *pilB*, *pilC*, and *pilD*, are required for biogenesis of *Pseudomonas aeruginosa* pili. *J Bacteriol* **172**, 2911–2919.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* **12**, 357–358.

Pallen, M. J. & Ponting, C. P. (1997). PDZ domains in bacterial proteins. *Mol Microbiol* **26**, 411–413.

Pao, S. S., Paulsen, I. T. & Saier, M. H., Jr (1998). The major facilitator superfamily. *Microbiol Mol Biol Rev* **62**, 1–32.

Parge, H. E., Forest, K. T., Hickey, M. J., Christensen, D. A., Getzoff, E. D. & Tainer, J. A. (1995). Structure of the fibre-forming protein pilin at 2·6 Å resolution. *Nature* **378**, 32–38.

Patenge, N., Berendes, A., Engelhardt, H., Schuster, S. & Oesterhelt, D. (2001). The *fla* gene cluster is involved in the biogenesis of flagella in *Halobacterium salinarum*. *Mol Microbiol* **41**, 653–663.

Planet, P. J., Kachlany, S. C., DeSalle, R. & Figurski, D. H. (2001). Phylogeny of genes for secretion NTPases: identification of the widespread *tadA* subfamily and development of a diagnostic key for gene classification. *Proc Natl Acad Sci U S A* **98**, 2503–2508.

Possot, O. & Pugsley, A. P. (1994). Molecular characterization of PulE, a protein required for pullulanase secretion. *Mol Microbiol* **12**, 287–299.

Possot, O. & Pugsley, A. (1997). The conserved tetracysteine motif in the general secretory pathway component PulE is required for efficient pullulanase secretion. *Gene* **192**, 45–50.

Possot, O., Letellier, L. & Pugsley, A. P. (1997). Energy requirement for pullulanase secretion by the main terminal branch of the general secretory pathway. *Mol Microbiol* **24**, 457–464.

Possot, O., Vignon, G., Bomchil, N., Ebel, F. & Pugsley, A. P. (2000). Multiple interactions between pullulanase secreton components involved in stabilization and cytoplasmic membrane association of PulE. *J Bacteriol* **182**, 2142–2152.

Postle, K. & Kadner, R. J. (2003). Touch and go: tying TonB to transport. *Mol Microbiol* **49**, 869–882.

Pugsley, A. P. (1993a). The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev* **57**, 50–108.

Pugsley, A. P. (1993b). Processing and methylation of PulG, a pilin-like component of the general secretory pathway of *Klebsiella oxytoca*. *Mol Microbiol* **9**, 295–308.

Pugsley, A. P., Francetic, O., Possot, O. M., Sauvonnet, N. & Hardie, K. R. (1997). Recent progress and future directions in studies of the main terminal branch of the general secretory pathway in Gram-negative bacteria – a review. *Gene* **192**, 13–19.

Pugsley, A. P., Bayan, N. & Sauvonnet, N. (2001). Disulphide bond formation in secreton component PulK provides a possible explanation for the role of DsbA in pullulanase secretion. *J Bacteriol* **183**, 1312–1319.

Py, B., Loiseau, L. & Barras, F. (1999). Assembly of the type II secretion machinery of *Erwinia chrysanthemi*: direct interaction and associated conformational change between OutE, the putative ATP-binding component and the membrane protein OutL. *J Mol Biol* **289**, 659–670.

Py, B., Loiseau, L. & Barras, F. (2001). An inner membrane platform in the type II secretion machinery of Gram-negative bacteria. *EMBO Rep* **2**, 244–248.

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 (1999a). Genome archeology leading to the characterization and classification of transport proteins. *Curr Opin Microbiol* **2**, 555–561.

Saier, M. H., Jr (1999b). Evolutionary origins of transmembrane transport systems. In *Transport of Molecules Across Microbial Membranes* (Society for General Microbiology Symposium no. 58), pp. 252–274. Edited by J. K. Broome-Smith, S. Baumberg, C. J. Stirling & F. B. Ward. Cambridge: Cambridge University Press.

Saier, M. H., Jr (2000). Vectorial metabolism and the evolution of transport systems. *J Bacteriol* **182**, 5029–5035.

Saier, M. H., Jr (2001a). Evolution of transport proteins. In *Genetic Engineering. Principles and Methods*, vol. 23, pp. 1–10. Edited by J. K. Setlow. New York: Kluwer Academic/Plenum Press.

Saier, M. H., Jr (2001b). Families of transporters: a phylogenetic overview. In *Microbial Transport Systems*, pp. 1–22. Edited by G. Winkelmann. Weinheim: Wiley.

Saier, M. H., Jr (2003a). Answering fundamental questions in biology with bioinformatics. *ASM News* **69**, 175–181.

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

Sandkvist, M. (2001). Biology of type II secretion. *Mol Microbiol* **40**, 271–283.

Sandkvist, M., Bagdasarian, M., Howard, S. P. & DiRita, V. J. (1995). Interaction between the autokinase EpsE and EpsL in the cytoplasmic membrane is required for extracellular secretion in *Vibrio cholerae*. *EMBO J* **14**, 1664–1673.

Sandkvist, M., Hough, L. P., Bagdasarian, M. M. & Bagdasarian, M. (1999). Direct interaction of the EpsL and EpsM proteins of the general secretion apparatus in *Vibrio cholerae*. *J Bacteriol* **181**, 3129–3135.

Sandkvist, M., Keith, J. M., Bagdasarian, M. & Howard, S. P. (2000). Two regions of EpsL involved in species-specific protein-protein interactions with EpsE and EpsM of the general secretion pathway in *Vibrio cholerae*. *J Bacteriol* **182**, 742–748.

Sauvonnet, N., Vignon, G., Pugsley, A. P. & Gounon, P. (2000). Pilus formation and protein secretion by the same machinery in *Escherichia coli*. *EMBO J* **19**, 2221–2228.

Schmidt, S. A., Bieber, D., Ramer, S. W., Hwang, J., Wu, C.-Y. & Schoolnik, G. (2001). Structure-function analysis of BfpB, a secretin-like protein encoded by the bundle-forming-pilus operon of enteropathogenic *Escherichia coli*. *J Bacteriol* **183**, 4848–4859.

Schoenhofen, I. C., Stratilo, C. & Howard, S. P. (1998). An ExeAB complex in the type II secretion pathway of *Aeromonas hydrophila*: effect of ATP-binding cassette mutations on complex formation and function. *Mol Microbiol* **29**, 1237–1247.

Skerker, J. M. & Berg, H. C. (2001). Direct observation of extension and retraction of type IV pili. *Proc Natl Acad Sci U S A* **98**, 6901–6904.

Thanassi, D. G. (2002). Ushers and secretins: channels for the secretion of folded proteins across the bacterial outer membrane. *J Mol Microbiol Biotechnol* **4**, 11–20.

Thomas, J. D., Reeves, P. J. & Salmond, G. P. C. (1997). The general secretion pathway of *Erwinia carotovora* subsp. *carotovora*: analysis of the membrane topology of OutC and OutF. *Microbiology* **143**, 713–720.

Thomas, N. & Jarrell, K. (2001). Characterization of flagellum gene families of methanogenic archaea and localization of novel flagellum accessory proteins. *J Bacteriol* **183**, 7154–7164.

Thomas, N. A., Bardy, S. L. & Jarrell, K. F. (2001). The archaeal flagellum: a different kind of prokaryotic motility structure. *FEMS Microbiol Rev* **25**, 147–174.

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

Tseng, T.-T., Gratwick, K. S., Kollman, J., Park, D., Nies, D. H., Goffeau, A. & Saier, M. H., Jr (1999). The RND permease superfamily: an ancient, ubiquitous and diverse family that includes human disease and development proteins. *J Mol Microbiol Biotechnol* **1**, 107–125.

von Heijne, G. (1992). Membrane protein structure prediction, hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225**, 487–494.

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

Wolfgang, M., van Putten, J. P. M., Hayes, S. F., Dorward, D. & Koomey, M. (2000). Components and dynamics of fiber formation define a ubiquitous biogenesis pathway for bacterial pili. *EMBO J* **19**, 6408–6418.

Yen, M. R., Peabody, C. R., Partovi, S. M., Zhai, Y., Tseng, Y. H. & Saier, M. H. (2002). Protein-translocating outer membrane porins of Gram-negative bacteria. *Biochim Biophys Acta* **1562**, 6–31.

Yeo, H.-J., Savvides, S. N., Herr, A. B., Lanka, E. & Waksman, G. (2000). Crystal structure of the hexameric ATPase of the *Helicobacter pylori* type IV secretion system. *Mol Cell* **6**, 1461–1472.

Young, G. B., Jack, D. L., Smith, D. W. & Saier, M. H., Jr (1999). The amino acid/auxin : proton symport permease family. *Biochim Biophys Acta* **1415**, 306–322.

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

Zhai, Y. & Saier, M. H., Jr (2001b). A web-based program (WHAT) for the simultaneous prediction of hydropathy, amphipathicity, secondary structure and transmembrane topology for a single protein sequence. *J Mol Microbiol Biotechnol* **3**, 501–502.

Zhai, Y. & Saier, M. H., Jr (2002). The β-barrel finder (BBF) program, allowing identification of outer membrane-barrel proteins encoded within prokaryotic genomes. *Protein Sci* **11**, 2196–2207.