Protein Science (1994), *3*:325–344. Cambridge University Press. Printed in the USA.

Sequence relationships between integral inner membrane proteins of binding protein-dependent transport systems: Evolution by recurrent gene duplications

W. SAURIN AND E. DASSA

Unité de Programmation Moléculaire et Toxicologie génétique, CNRS URA 1444, Institut Pasteur 25, Rue du Dr. Roux–F75645 Paris Cedex 15, France

(RECEIVED September 23, 1993; ACCEPTED November 22, 1993)

Abstract

Periplasmic binding protein-dependent transport systems are composed of a periplasmic substrate-binding protein, a set of 2 (sometimes 1) very hydrophobic integral membrane proteins, and 1 (sometimes 2) hydrophilic peripheral membrane protein that binds and hydrolyzes ATP. These systems are members of the superfamily of ABC transporters. We performed a molecular phylogenetic analysis of the sequences of 70 hydrophobic membrane proteins of these transport systems in order to investigate their evolutionary history. Proteins were grouped into 8 clusters. Within each cluster, protein sequences displayed significant similarities, suggesting that they derive from a common ancestor. Most clusters contained proteins from systems transporting analogous substrates such as monosaccharides, oligopeptides, or hydrophobic amino acids, but this was not a general rule. Proteins from diverse bacteria are found within each cluster, suggesting that the ancestors of current clusters were present before the divergence of bacterial groups. The phylogenetic trees computed for hydrophobic membrane proteins of these permeases are similar to those described for the periplasmic substrate-binding proteins. This result suggests that the genetic regions encoding binding protein-dependent permeases evolved as whole units. Based on the results of the classification of the proteins and on the reconstructed phylogenetic trees, we propose an evolutionary scheme for periplasmic permeases. According to this model, it is probable that these transport systems derive from an ancestral system having only 1 hydrophobic membrane protein. None of the proteins considered in this study display detectable sequence similarity to hydrophobic membrane proteins or domains from other ABC transporters such as bacterial polysaccharide export systems, bacterial toxin proteins exporters, and eukaryotic ABC proteins. It is likely that they constitute a specific subfamily within the superfamily of ABC transporters.

Keywords: ABC transporters; ATP-binding proteins; bacteria; binding protein-dependent permeases; computational analysis; evolution; integral membrane proteins; periplasmic space; phylogenetic relationships

In enteric bacteria, periplasmic binding protein-dependent transport systems or, for short, periplasmic permeases (Ames, 1988) participate in the transport of a wide variety of substrates and show a common global organization (Higgins et al., 1990). These multicomponent systems contain a periplasmic substrate-binding protein that is released from bacteria by cold osmotic shock. In many cases this protein serves also as a primary receptor for chemotaxis toward substrates (Hazelbauer, 1975). Transport depends on the presence of the periplasmic protein that binds the substrate with high affinity in the micromolar range. In addition to this soluble binding protein, these transport systems include 2 (sometimes 1) very hydrophobic integral membrane proteins and 1 (sometimes 2) hydrophilic peripheral membrane protein that form a complex in the cytoplasmic membrane. This complex mediates the ATP-dependent translocation of the substrate into the cytoplasm. The peripheral membrane proteins contain sequence motifs similar to the ATP-binding motif found in ATPases or kinases. Such proteins from the oligopeptides, histidine, and maltose transport systems bind ATP analogues (Higgins et al., 1990), and it was recently shown that the purified MalK ATP-binding protein hydrolyzes ATP (Walter et al., 1992). Remarkably, the comparison of the predicted sequences of ATP-

Reprint requests to: Elie Dassa, Unité de Programmation Moléculaire et Toxicologie génétique, CNRS URA 1444, Institut Pasteur 25, Rue du Dr. Roux–F75645 Paris Cedex 15, France; e-mail: elidassa@pasteur.fr.

binding proteins reveals long regions of extensive similarity (30% sequence identity over the entire protein sequences) that extend beyond the ATP-binding region.

Both prokaryotes and eukaryotes possess other transport systems that share the global organization of periplasmic permeases. In gram-negative bacteria, the systems that transport iron-bearing siderophores and vitamin-B12 constitute one such a class of transporters. However, they differ by the presence of an additional component, a substrate-specific high-affinity outer membrane receptor, through which substrates enter the periplasm in an energy-dependent fashion. The TonB protein, which is essential for the function of all these systems, may be an energy-transducer protein that couples the outer membrane transport to inner membrane energy-generating systems (Skare & Postle, 1991). Iron-bearing siderophore transport systems also contain a periplasmic substrate-binding protein (Köster & Braun, 1990) but, at least in the cobalamine transport system, the protein BtuE may be dispensable (Rioux & Kadner, 1989). Clearly, in these systems, the major recognition event takes place at the level of the outer membrane receptor.

Gram-positive bacteria (Gilson et al., 1988) and mycoplasma (Dudler et al., 1988) have transport systems with a genetic and/or structural organization similar to periplasmic permeases. They contain a set of hydrophobic and peripheral membrane proteins and specifically, a membrane lipoprotein displaying sequence similarity to certain gram-negative substrate-binding proteins. It is not known if these lipoproteins actually bind substrates as their periplasmic counterparts in enterobacteria.

In various bacteria, systems responsible for the secretion of drugs (Guilfoile & Hutchinson, 1991), polysaccharides (Cangelosi et al., 1989), or proteins into the medium (Glaser et al., 1988; Gilson et al., 1990; Létoffé et al., 1990; Guzzo et al., 1991; Possot et al., 1992) also contain 1 or 2 hydrophobic integral membrane proteins or domains and a hydrophilic peripheral membrane protein or domain that displays an ATP-binding motif highly similar to those of periplasmic permeases.

Finally, eukaryotes such as *Drosophila* (Dreesen et al., 1988), yeast (McGrath & Varchavsky, 1989), and mammals have proteins involved in the excretion of polypeptides, pigments, or small molecules that include domains with strong similarity to the ATP-binding proteins of periplasmic permeases. The most prominent representatives of the mammalian systems are the Mdr protein (Chen et al., 1986; Gros et al., 1986), the CFTR protein (Riordan et al., 1989), and proteins that are probably involved in peptide antigen presentation (Deverson et al., 1990; Monaco et al., 1990).

In all theses systems, the different subunits or domains may be arranged in various ways. In prokaryotes they are generally independent, whereas in eukaryotic systems they are fused into a single polypeptide chain that presents alternating transmembrane and ATP-binding domains. Other proteins, apparently not involved in transport, such as the *Escherichia coli* UvrA and FtsE proteins, which participate in DNA repair and cell division, respectively, share sequence similarities with the ATP-binding subunits of these transport systems (see Higgins [1992] for a review).

The relatedness of these systems in terms of organization and mechanism, and the high conservation of the ATP-binding proteins or domains strongly suggested that they have a common evolutionary origin. They may constitute a superfamily for which the names of "ABC-transporters" or "traffic ATPases" have been proposed (Ames et al., 1990; Higgins et al., 1990). The sequences of the ATP-binding proteins have been compared in several laboratories but their study awaits a more extensive phylogenetic analysis. Recently, Tam and Saier (1993) reported an evolutionary analysis of the substrate-binding components of periplasmic transport systems. They showed that these proteins may be organized into 8 families. To test whether all the components of such transport systems have a common evolutionary origin, we analyzed the sequence relationships among their hydrophobic membrane proteins. In contrast to the high sequence conservation of the ATP-binding subunits, hydrophobic inner membrane components are generally thought to display few sequence similarities. However, we have found that most of these proteins display a conserved peptide motif, with the consensus "EAA---G---------I-LP," located at about 90 residues from their C-terminus (Dassa & Hofnung, 1985). This work shows that integral inner membrane proteins of periplasmic transport systems can be grouped into 8 clusters of similar proteins, which generally correlate with those described by Tam and Saier (1993). Our data suggest that the genetic regions encoding periplasmic transport system evolved as units. We also describe a probable evolutionary scheme for binding protein-dependent transport systems by recurrent gene duplication.

Results

Bacterial integral inner membrane proteins from binding protein-dependent transport systems are not homologous to eukaryotic ABC transport systems nor to prokaryotic ABC excretion systems

When bacterial integral inner membrane protein sequences from binding protein-dependent transport systems were used to search protein sequence databases, no significant similarity was found to either eukaryotic proteins like the Mdr or the CFTR proteins, nor to prokaryotic ABC excretion systems like the hemolysin translocator (Gilson et al., 1990) and the envelope polysaccharide exporting systems (Frosch et al., 1991). This is in contrast to the ATP-binding proteins that are strongly conserved in all these systems. Therefore the phylogenetic study that follows was applied to the set of proteins that have significant similarity with at least 1 hydrophobic protein of a well-characterized bacterial binding protein-dependent transport system. The proteins were collected as described in the Materials and methods section and were listed in Table 1A. Interestingly, 2 proteins for which no function in transport has been demonstrated, the MPOMBPY open reading frame (ORF) from the chloroplast genome of *Marchantia polymorpha* and the CPANIFC protein from *Clostridium pasteurianum*, respectively similar to ECOCYST and ECOCHLJ, were found to fulfill this criterion. Table 1B reports sequences of proteins found after completion of this study.

Extensive searches of translated nucleic acid databases identified new binding protein-dependent transport systems

Each protein of the selected sequences was used to search nucleic acid databases translated in the 6 reading frames. We found new prokaryotic sequences in the databases, located near previously sequenced genes. For instance, ORFs ranging from 50

to 200 amino acids similar to ECOMALG were found near the *Clostridium thermosulfurogenes* β-amylase and β-galactosidase genes, the *Microbispora bispora* cellodextrinase gene, and the *Bacillus stearothermophilus* α-amylase gene. These observations suggest that sequences encoding binding protein-dependent transport systems are likely to reside in close vicinity to the genes of these oligosaccharide-degrading enzymes, and that they probably constitute operons. The complete set of sequences found during this search is reported in Table 2.

Classification of integral inner membrane proteins from periplasmic transport systems

The proteins described in Table 1A were grouped in clusters according to their sequence similarities as described in the Materials and methods section. Figure 1 displays a graphic representation of the results of the pairwise comparisons performed to build clusters. Eight major clusters were identified, in which the computed scores obtained from pairwise comparisons were equal or higher than the threshold score of 88. It appeared that hydrophobic membrane proteins from a given transport system fell in 2 categories, a major one where the 2 partners of the system were found in the same cluster and a minor one where the 2 partners belonged to different clusters such as the proteins from the OPP (oligopeptide) and the LIV (leucine, isoleucine, and valine) transport systems.

Cluster 1: A large cluster of transport systems with a wide diversity of substrates

It was not possible using Treealign (Hein, 1990) to generate a multiple alignment of all the protein sequences present in this cluster. To compute trees, we further divided the cluster into 3 subclusters.

Subcluster 1a: The phosphate, sulfate, molybdate, glycine-betaine, spermine, and putrescine transport systems. Most transport systems with a single hydrophobic membrane protein fall in this cluster as for instance ECOPROW, SMASFUB, ECOCHLJ, CPANIFC, and AVIMODC. Figure 2 shows the multiple alignment and the tree computed from this alignment. The tree has 2 major branches, one containing the proteins from the phosphate transport system, the second the proteins from the molybdate, sulfate, and polyamine transporters. The ECOPROW protein from the glycine-betaine transport system and the SMASFUB protein from a non-siderophore transport system in *Serratia marcescens* diverged early, near the root of the tree.

This subcluster contains an orf MPOMBPY, identified in the chloroplast genome of *M. polymorpha* (Ohyama et al., 1986), which is strongly similar to SYNCYST and ECOCYST (Laudenbach & Grossman, 1991). This fact, and the presence in the same genome of an orf MBPX similar to the CYSA ATP-binding protein from *E. coli*, favors the hypothesis of the occurrence of a binding protein-dependent transport system for sulfate in chloroplasts. However, a chloroplast analogous to the substrate-binding protein has not been detected yet. A chloroplast homologue of CYSW is also lacking. The late divergence of MPOMBPY from SYNCYST may indicate that the chloroplast homologue of CYSW was lost after the divergence of chloroplasts from cyanobacteria and probably during the reduction of the chloroplast genome. The possibility that the genes coding for proteins lacking from this system have been moved onto the plant nuclear genome cannot be excluded, and one may speculate on the possibility that they could be found in the chloroplast genomes from other plant cells.

Subclusters 1b and 1c: The di-, oligosaccharides, and β-glycerophosphate, transport systems. These subclusters contain proteins from the maltose and maltodextrins transport system in *E. coli* and *Streptococcus pneumoniae*, the putative starch-degradation products transport system in *C. thermosulfurogenes*, the multiple sugar (raffinose, melibiose) transport system from *Streptococcus mutans*, the β-glycerophosphate transport system from *E. coli*, and the lactose transport system from *Agrobacterium radiobacter*. These systems include 2 hydrophobic membrane proteins. Remarkably, none of these transport systems have their 2 proteins present in the same subcluster. Moreover, all the proteins whose genes are located upstream with respect to the transcriptional direction are grouped in one subcluster (the MalF-like proteins in Fig. 3A). Similarly, all downstream proteins are grouped in the other subcluster (MalG-like proteins in Fig. 3B). It should be noted that the trees of each subcluster have identical topology and that proteins from the same transport system are located in the same relative position within their respective trees. This suggests that evolutionary constraints act on the 2 genes as a single unit. The substrates transported by these systems are mainly disaccharides, in either α or β configurations. The transport system for β-glycerophosphate, not structurally related to disaccharides, falls in these subclusters.

Cluster 2: The histidine, glutamine, arginine, nopaline (N2-[1-D-dicarboxypropyl]-L-arginine), and octopine (N2-[1-D-dicarboxyethyl]-L-arginine) transport systems

The alignment and the tree for these proteins are shown in Figure 4. With the exception of the glutamine transport system, these permeases have 2 hydrophobic membrane proteins. Two sub-trees are evident, one with proteins similar to STYHISM and one to STYHISQ of the *Salmonella typhimurium* histidine transport system. The topologies of the sub-trees are identical. The ECOGLNP protein diverged early, at the root of the tree, suggesting that the systems in cluster 2 evolved from a common ancestor system with a single hydrophobic protein. The ECOGLNP protein might be reminiscent of this ancestral protein.

E. coli possesses an arginine-specific transport system distinct from the histidine system (Wissenbach et al., 1993). The structure of the tree suggests that these systems diverged from an ancestor system with 2 hydrophobic proteins.

The NOC transport system (for nopaline, an arginine derivative) and the OCC transport system (for octopine, lysopine, histopine, and octopinic acid, which are, respectively, derivatives of arginine, lysine, histidine, and ornithine) are very closely related and are located on the large Ti plasmids of *Agrobacterium tumefaciens* (Valdivia et al., 1991; Zanker et al., 1992). Each system contains 2 hydrophobic proteins strongly related to the STYHISM and STYHISQ proteins. It is likely, as deduced from the topology of the tree, that the NOC and OCC systems evolved from a chromosomal homologue of the histidine transport system in *A. tumefaciens* (Krishnan et al., 1991). The corresponding *S. typhimurium* system transports histidine, lysine, arginine, and ornithine (Higgins et al., 1982)

Table 1. *Periplasmic binding protein-dependent transport systems*[a]

Organisms	Genes	Transported molecules	Kind	Proteins	Size	Accession number	Abbreviated name	Reference
A. Proteins considered in this study								
Gram-negative bacteria enterobacteria								
Escherichia coli	*artIQMJ*	Arginine	A2IP	ARTQ	238	embl:X67753	ECOARTQ	
				ARTM	222		ECOARTM	
	araFGH	Arabinose	PIA	ARAH	329	gb:X06191	ECOARAH	(Scripture et al., 1987)
	btuBCED	Vitamin B12	PIA	BTUC	326	gb:M14031	ECOBTUC	(Friedrich et al., 1986)
	chlJD	Molybdate	?	CHLJ	200	gb:M16182	ECOCHLJ	(Johann & Hinton, 1987)
	cysPTWAM	Sulfate/thiosulfate	P2IA	CYST	277	gb:M32101	ECOCYST	(Sirko et al., 1990)
				CYSW	291		ECOCYSW	(Hryniewicz et al., 1990)
	fecABCDE	Iron dicitrate	P2IA	FECC	332	gb:M26397	ECOFECC	(Staudenmaier et al., 1989)
				FECD	318		ECOFECD	
	fepE fepDCG fepB	Iron enterobactin		FEPD	334	gb:X57471	ECOFEPD	(Chenault & Earhart, 1991)
				FEPG	330		ECOFEPG	(Shea & Mcintosh, 1991)
	fhuACDB	Iron hydroxamate	API^2	FHUB	659	gb:X04319	ECOFHUB	(Koster & Braun, 1986)
	glnHPQ	Glutamine	PIA^2	GLNP	219	gb:X14180	ECOGLNP	(Nohno et al., 1986)
	livJ,livKHMGF	Leucine/isoleucine/valine	PII'AA'	LIVH	308	gb:J05516	ECOLIVH	(Adams et al., 1990)
				LIVM	424	gb:S47025	ECOLIVM	
	malEFG, malKLM	Maltooligosaccharides	P2IA	MALF	514	gb:J01648	ECOMALF	(Froshauer & Beckwith, 1984)
				MALG	296	gb:X02871	ECOMALG	(Dassa & Hofnung, 1985)
	mglABC	Galactose	PIA	MGLC	336	gb:M59444	ECOMGLC	(Hogg et al., 1991)
	phnA-Q	Phosphonates	?[b]	PHNE	276	gb:J05260	ECOPHNE	(Chen et al., 1990)
				PHNM	378		ECOPHNM	(Makino et al., 1991)
	potABCD	Spermidine-putrescine	P2IA	POTB	275	gb:M64519	ECOPOTB	(Furuchi et al., 1991)
				POTC	264		ECOPOTC	
	proVWX	Glycine-betaine	PIA	PROW	354	gb:K01992	ECOPROW	(Gowishankar, 1989)
	pstSABCU	Inorganic phosphate	P2IAA'	PSTA	296	gb:X02723	ECOPSTA	(Amemura et al., 1985)
				PSTC	309		ECOPSTC	(Surin et al., 1985)
	rbsDACBK	Ribose	PIA(?I')	RBSC	321	gb:M13169	ECORBSC	(Bell et al., 1986)
				RBSD	144		ECORBSD	
	ugpBAECQ	β-Glycerophosphate	P2IA	UGPA	293	gb:X13141	ECOUGPA	(Overduin et al., 1988)
				UGPE	281		ECOUGPE	
Enterobacter aerogenes	*malEFG, malKLM*	Maltooligosaccharides	P2IA	MALG	296	pir:S05333	EAEMALG	(Dahl et al., 1989)
				MALF	514	pir:S05332	EAEMALF	
Salmonella typhimurium	*argT,hisJQMP*	Histidine	P2IA	HISM	235	gb:J01805	STYHISM	(Higgins et al., 1982)
				HISQ	228		STYHISQ	
	malEFG, malKLM	Maltooligosaccharides	P2IA	MALG	296	gb:X05491	STYMALG	(Francoz et al., 1990)
				MALF	514		STYMALF	(Schneider et al., 1992)
	oppABCDE	Oligopeptides	PII'AA'	OPPB	305	gb:X05491	STYOPPB	(Hiles et al., 1987)
				OPPC	302		STYOPPC	
Serratia marcescens	*sfuABC*	Iron (Fe^{3+})	PI^2A	SFUB	527	gb:M33815	SMASFUB	(Angerer et al., 1990)
Other gram-negative bacteria								
Agrobacterium radiobacter	*lacI EFGKZ*	Lactose	P2IA	LACF	298	gb:S40378	ARALACF	(Williams et al., 1992)
				LACG	274		ARALACG	

Agrobacterium tumefaciens (pTi)	*occQMPJ*	Octopine	P2IA	OCCQ	237	gb:M77784	ATUOCCQ	(Valdivia et al., 1991)
				OCCM	246		ATUOCCM	
	nocPTQM	Nopaline	P2IA	NOCQ	236	gb:M77785	ATUNOCQ	(Zanker et al., 1992)
				NOCM	241		ATUNOCM	
Azotobacter vinelandii	*modABCD*	Molybdate	PIA	MODC	226	gb:X69077	AVIMODC	(Luque et al., 1993)
Vibrio anguillurum (pMJ1 plasmid)	*fatDCBA*	Iron anguibactin		FATC	317	gb:M74068	VANFATC	(Koster et al., 1991)
				FATD	314		VANFATD	
Pseudomonas aeruginosa	*braCDEFG*	Leucine/isoleucine/valine	PII'AA'	BRAD	307	gb:D90223	PAEBRAD	(Hoshino & Kose, 1990)
				BRAE	417		PAEBRAE	
Cyanobacteria								
Synechococcus sp.	*cysA,sbpAcysTRW*	Sulfate	P2IA	CYST	278	gb:M65247	SYNCYST	(Laudenbach & Grossman, 1991)
				CYSW	286		SYNCYSW	
Gram-positive bacteria								
Bacillus subtilis	*oppABCDE*	Oligopeptides	PII'AA'	OPPB	311	gb:X56347	BSUOPPB	(Perego et al., 1991)
				OPPC	305	gb:M57689	BSUOPPC	(Rudner et al., 1991)
	dciAABCDE	Dipeptides	PII'AA'	DCIAB	308	gb:X56678	BSUDCIAB	(Mathiopoulos et al., 1991)
				DCIAC	320		BSUDCIAC	
Clostridium pasteurianum	*nifC*	?	?	NIFC	286	gb:M34365	CPANIFC	(Wang et al., 1990)
Clostridium thermosulfuricum	*amyCD*	Starch degradation products	?2I?	AMYC	274	gb:S50264	CTHAMYC	(Bahl et al., 1991)
				AMYD	292	gb:X54982	CTHAMYD	
Streptococcus mutans	*msmEFG,gftAmsmK*	Melibiose/raffinose	P2IA	MSMF	290	embl:S83895	SMUMSMF	(Russell et al., 1992)
				MSMG	277		SMUMSMG	
Streptococcus pneumoniae	*amiACDEF*	Oligopeptides?	PII'AA'	AMIC	495	gb:X17337	SPNAMIC	(Alloing et al., 1990)
				AMID	308		SPNAMID	
	malXCD	Maltooligosaccharides	P2I?	MALC	430	embl:L08611	SPNMALC	(Puyet & Espinosa, 1993)
				MALD	276		SPNMALD	
Mycoplasma								
Mycoplasma hyorinis		? (Invasion)	PI^2A	P69	580	gb:M37339	MHYP69	(Dudler et al., 1988)
Chloroplasts								
Marchantia polymorpha		? (Sulfate)	IA	MBPY	288	gb:X04465	MPOMBPY	(Ohyama et al., 1986)
B. Proteins found after completion of this work								
Escherichia coli	*potFGHI*	Putrescine	P2IA	POTH	317	gb:M93329		(Pistocchi et al., 1993)
				POTI	281			
Salmonella typhimurium	*livB CAEFG*	Leucine/isoleucine/valine	PII'AA'	LIVA	308	gb:D12589		(Ohnishi et al., 1990)
				LIVE	428	gb:X54292		(Matsubara et al., 1992)
Clostridium perfrengens putative membrane protein		Homologous to ECOMALG	?	ORF	274	gb:S51418		(Holck & Blom, 1992)
Synechocystis sp.	*cysT,sbpA*	Sulfate	?	CYST	235	gb:X67911		–
Synechococcus sp.	*nrtABCD*	Nitrate		NRTB	279	gb:X61625		(Omata et al., 1993)

[a] The table is composed of 2 parts organized similarly. Part A lists the proteins that are considered in this work. Part B lists the proteins from periplasmic permeases described after completion of this study. The proteins are grouped according to the species in which they are found. Genes: The structure of the operon encoding the proteins is provided. A space between genes symbolizes a transcriptional stop. A comma between operons indicates that they are transcribed in opposite directions. Kind: The protein composition of the system is given in an abbreviated form. A = 1 ATP-binding protein. I = 1 hydrophobic inner membrane protein. P = 1 periplasmic binding protein. 2I = 2 hydrophobic membrane proteins falling in the same cluster. II' = 2 hydrophobic membrane proteins falling in different clusters. I^2 = 1 hydrophobic membrane protein with an internal duplication. Abbreviated name: It is composed of the name of the bacterium (first 3 letters) and the name of the protein (last 4 letters).

[b] The PHN system may be constituted by more than one transport system.

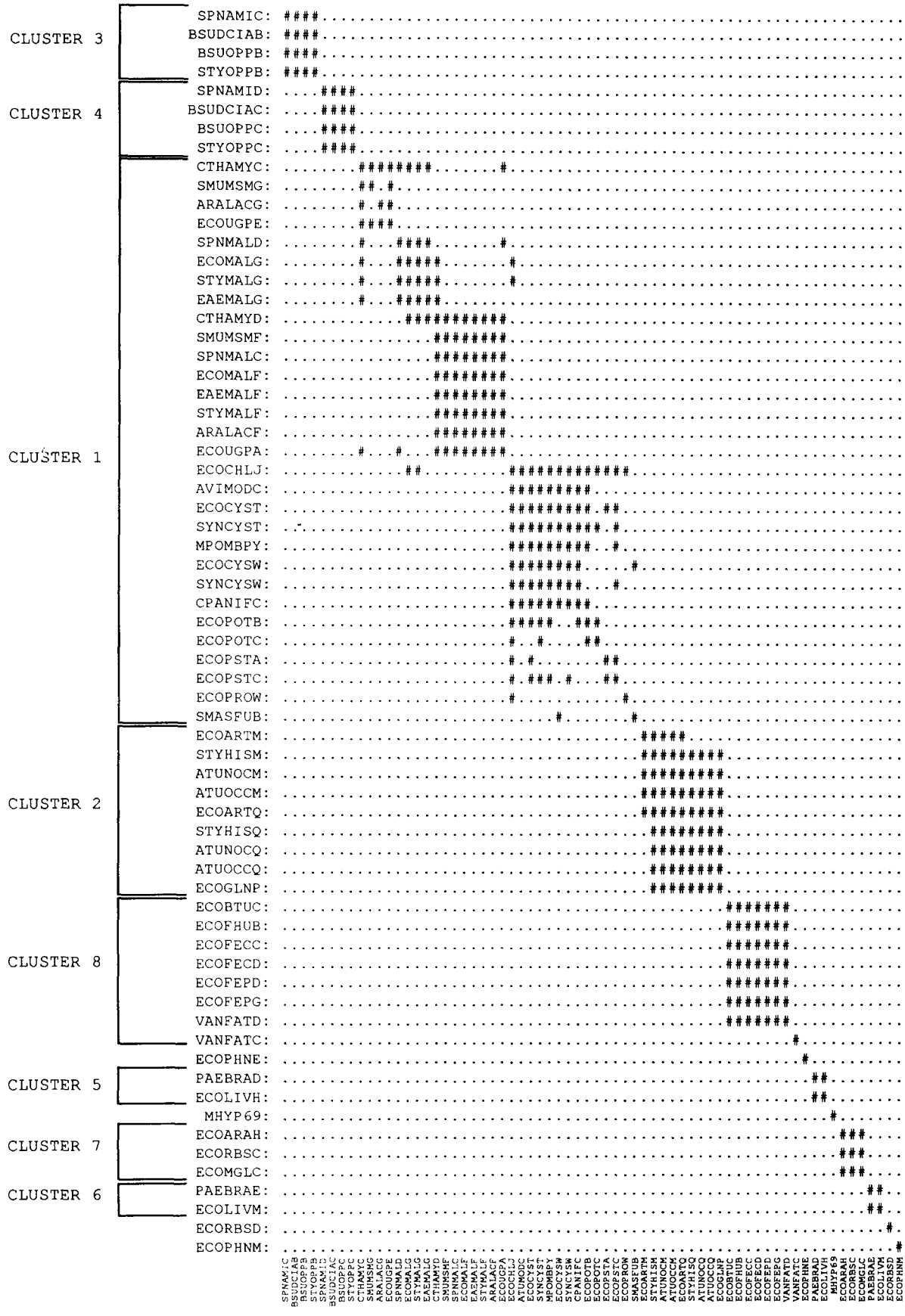

CLUSTER 3
CLUSTER 4
CLUSTER 1
CLUSTER 2
CLUSTER 8
CLUSTER 5
CLUSTER 7
CLUSTER 6
SPNAMIC: ####
BSUDCIAB: ####
BSUOPPB: ####
STYOPPB: ####
SPNAMID: ....####
BSUDCIAC: ....####
BSUOPPC: ....####
STYOPPC: ....####
CTHAMYC: ....########.....#
SMUMSMG: ....##.#
ARALACG: ....#.##
ECOUGPE: ....####
SPNMALD: ....#...####.....#
ECOMALG: ....#...#####.....#
STYMALG: ....#...#####.....#
EAEMALG: ....#...#####
CTHAMYD: ....##########
SMUMSMF: ....########
SPNMALC: ....########
ECOMALF: ....########
EAEMALF: ....########
STYMALF: ....########
ARALACF: ....########
ECOUGPA: ....#...#...########
ECOCHLJ: ....##.........##############
AVIMODC: ....#########
ECOCYST: ....##########.##
SYNCYST: ....##########.#
MPOMBPY: ....#########..#
ECOCYSW: ....########.....#
SYNCYSW: ....########...#
CPANIFC: ....#########
ECOPOTB: ....#####..###
ECOPOTC: ....#..#....##
ECOPSTA: ....#.#.....##
ECOPSTC: ....#.###.#...##
ECOPROW: ....#..........#
SMASFUB: ....#.....#
ECOARTM: ....#####
STYHISM: ....#########
ATUNOCM: ....#########
ATUOCCM: ....#########
ECOARTQ: ....#########
STYHISQ: ....########
ATUNOCQ: ....########
ATUOCCQ: ....########
ECOGLNP: ....########
ECOBTUC: ....#######
ECOFHUB: ....#######
ECOFECC: ....#######
ECOFECD: ....#######
ECOFEPD: ....#######
ECOFEPG: ....#######
VANFATD: ....#######
VANFATC: ....#
ECOPHNE: ....#
PAEBRAD: ....##
ECOLIVH: ....##
MHYP69: ....#
ECOARAH: ....###
ECORBSC: ....###
ECOMGLC: ....###
PAEBRAE: ....##..
ECOLIVM: ....##..
ECORBSD: ....#.
ECOPHNM: ....#

Table 2. *New open reading frames (ORFs) discovered by screening translated nucleic acid databases*[a]

Organisms, genetic region	Homologous to		Size	Accession number	Reference
Erwinia carotovora, *araC* gene	ECOARAH	ORF	89	gb:M11981	(Lei et al., 1985)
Mycobacterium tuberculosis, b antigen	ECOPSTA	ORF	201	gb:M30046	(Andersen et al., 1990)
Bacillus subtilis, subtilin operon	ECOPROW	ORF	202	gb:M99263	(Hansen & Chung, 1992)
Lactococcus lactis, lactose plasmid	STYOPPC	ORF	293	gb:M76471	
Clostridium thermosulfurogenes,					
β-amylase gene	ECOMALG	ORF	124	gb:M22471	(Kitamoto et al., 1988)
β-galactosidase gene	ECOMALG	ORF	50	gb:M57579	(Burchardt & Bahl, 1991)
Bacillus stearothermophilus,					
α-amylase gene	ECOMALG	ORF	18	gb:M36539	(Diderichsen & Christiansen, 1988)
Microbospora bispora,					
cellodextrinase gene	ECOMALG	ORF	141	gb:L06134	

[a] ORFs similar to hydrophobic membrane proteins of periplasmic permeases detected by screening nucleic acid databases. The conventions and nomenclature are the same as in Table 1.

All the proteins within clusters 1 and 2 display a short region of maximal similarity located in their C-termini. This region contains a set of identical residues that matches almost perfectly the previously defined consensus sequence for hydrophobic proteins from binding protein-dependent transport systems (Dassa & Hofnung, 1985). An alignment of these very limited conserved regions, obtained by using program Pileup (Devereux et al., 1984), appears in Figure 5.

Clusters 3 and 4: OPPB and OPPC clusters for di- or oligopeptide transport systems

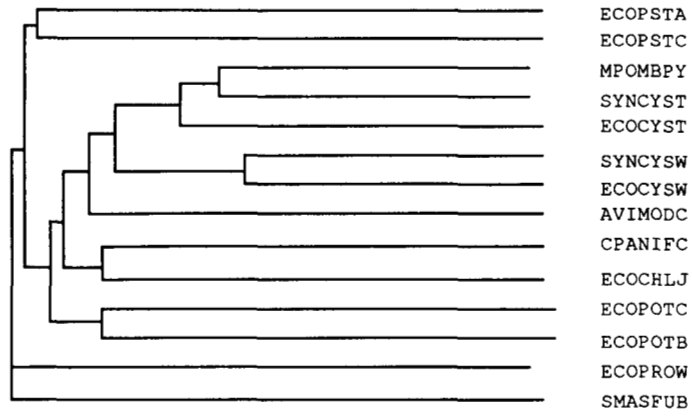
Di- or oligopeptide transport systems are found in gram-negative and gram-positive bacteria. In *Bacillus subtilis*, the integrity of the oligopeptide transport system is crucial for the establishment of sporulation (Perego et al., 1991), whereas in *S. pneumoniae*, *ami*-negative mutants show pleiotropic defects including a reduced ability to generate a membrane potential (Trombe et al., 1984). All these systems are composed of 2 hydrophobic membrane proteins that partition into 2 clusters. Cluster 3 contains proteins similar to the OPPB protein of *S. typhimurium*, whereas cluster 4 contains proteins similar to the OPPC protein. The distances between proteins from cluster 3 and cluster 4 are in the same range as the distance observed between completely unrelated clusters. Within each cluster, the proteins are very highly conserved (Fig. 6), 31 and 47 residues being identical in the respective alignments of clusters 3 and 4. The topologies of these trees are identical with proteins belonging to the same system being in the same locations, suggesting that both genes were constrained similarly during evolution. Interestingly, the AMI proteins diverged earlier from the group of oligopeptide and dipeptide-transporting proteins, probably before the separation of gram-negative and gram-positive bacteria. This might be due to a peculiar functional specialization of the AMI system suggested by the complex phenotypes of *ami*-negative mutants. The BSUOPPB and BSUOPPC proteins are more related to the BSUDCIAB and BSUDCIAC proteins than to the STYOPPB and STYOPPC proteins. Moreover Tam and Saier (1993) have shown that this was true for the respective substrate binding proteins. This may be interpreted as the consequence of the specialization of the oligopeptide and the dipeptide transport systems from an ancestral peptide transporter into *B. subtilis*. The sequences of the membrane proteins from the dipeptide transport system in *S. typhimurium* are not available. Their determination will help to decide whether such a specialization event occurred independently in *B. subtilis* and in *S. typhimurium*.

Clusters 5 and 6: LIVH and LIVM clusters for branched-chain amino acid transport systems

These systems, found in *E. coli*, *S. typhimurium*, and *Pseudomonas aeruginosa*, are involved in the transport of leucine, isoleucine, and valine. They contain 2 hydrophobic membrane proteins. The proteins from *S. typhimurium* are nearly identical to that of *E. coli* and were not considered in this analysis. As for the peptide transport systems, there is no significant sequence similarity between proteins from these 2 clusters. The trees and the alignments for these 2 clusters were not shown.

We therefore cannot assess whether the *livH* and *livM* genes on one hand, and the *oppB* and *oppC* genes on the other hand have diverged from a common ancestor. The lack of significant similarity between partners could be due either to a very early duplication event leading to the individualization of the 2 components, or alternatively, it might indicate that the 2 proteins are constrained differently.

Fig. 1 (*facing page*). Classification of hydrophobic membrane proteins of periplasmic binding protein-dependent transport systems. This figure summarizes graphically the results of the pairwise comparisons and the classification of the proteins. The names of proteins are as in Table 1. The order of the proteins in columns is the same as in the lines. At the intersection of lines and columns are represented symbols for the computed similarity scores. The symbols are the following: (#) indicates a score higher than or equal to the threshold score of 88; (.) indicates a score lower than the threshold value. Clusters are identified by brackets drawn at the left of the protein names.




```
AAVMSFPLMVRAIRLALEGVDVKLEQAARTLGAGRWRVFFTITLPLTLPGIIVGTVLAFARSLGEFGATITFVSNIPGET  149 ECOCHLJ
SVFYSLPFVVQPLQNAFEAIGERPLEVASTLRAGPWDTFFTVVVPLARPGFITAAILGFAHTVGEFGVVLMIGGNIPEKT  178 ATUMODC
MAFTSIPFVVRTVQPVLEELGPEYEEAAETLGATRWQSFCKVVLPELSPALVAGVALSFTRSLGEFGAVIFIAGNIAWKT  226 ECOCYST
MVFISLPFVVRTVEPLLLELEVEAEEAAASLGASPSETFWRVILPPILPGVLAGVAQGFSRAVGEFGSVVIISGNLPFDD  229 SYNCYST
MIFVSLPFVVRTIQPVLQNMEEDLEEAAWCLGASPWTTFWHILFPPLTPSLLTGTTLGFSRALGEYGSIVLIASNIPMKD  239 MPOMBPY
TIFVTCPFVVRELVPVMLSQGSQEDEAAILLGASGWQMFRRVTLPNIRWALLYGVVLTNARAIGEFGAVSVVSGSIRGET  232 ECOCYSW
TIFVSMPFVAREVIPNLEEIGTDAEEAASTLGANGWQTFWRVTLPSIKWSMLYGVVLTTARALGEFGAVSVVSGSITGKT  235 SYNCYSW
QFFVSSALYVRVLRDSVKSVPIELFEVSYVLGAGKIETIIKIMIPMLKKSIVSGLILAWIRSLGEFGATLMFAGNIIGKT  242 CPANIFC
LVYILLPFMVMPLYSSIEKLDKPLLEAARDLGASKLQTFIRIIIPLTMPGIIAGCLLVMLPAMGLFYVSDLMGGAKNL--  224 ECOPOTB
HITFCLPFVVVTVYSRLKGFDVRMLEAAKDLGASEFTILRKIILPLAMPAVAAGWVLSFTLSMDDVVVSSFVTGPSYE--  212 ECOPOTC
LALLQVPIVIRTTENMLKLVPYSLREAAYALGTPKWKMISAITLKASGSGIMTGILLAIARIAGET-APLLFTALSNQ--  236 ECOPSTA
LAIMIIPYIAAVMRDVFEQTPVMMKESAYGIGCTTWEVIWRIVLPFTKNGVIGGIMLGLGRALGETMAVTFIIGNTYQ--  254 ECOPSTC
TIIFALPPIIRLTILGINQVPADLIEASRSFGASPRQMLFKVQLPLAMPTIMAGVNQTLMLALSMVVIASMIAVGGLG--  291 ECOPROW
VLAY-FPFIYLPAAAVLRRLDPGIEDVATSLGSRPPAVFFRVVLPQLKLAVWGGSLLIALHLLAEYGLYAMIRFDTFT--  220 SMASFUB

RTIPSAMYTLIQTPGGESGAARLCII--SIALA----------MISLLISEWL-A---R--------I  193 ECOCHLJ
RTVAVQIFDHVEAMEYAQAHWLAGGM--VLFS----------FLVLFALY-S---S--------R  219 ATUMODC
EVTSLMIFVRLQEFDYPAASAIASVI--LAASL----------LLLFSINTLQ-S---R--------F  270 ECOCYST
LIAPVLIFERLEQYDYAGATVIGSVL--LLFSL----------VILFVINALQ-N-------------  271 SYNCYST
LVISVLLFQKLEQYDYKSATIIASFV--LIISF----------TALFFINKIQ-L-------------  281 MPOMBPY
LSLPLQIELLEQDYNTVGSFTAAALL--TLMAI----------ITLFLKSMLQ-W---R--------L  276 ECOCYSW
QTLPLFVEEAYKQYQTTLSYTAALLL--GGISL----------VTLVLKALLE-A---R--------T  279 SYNCYSW
RTIPLQIYTYMQDDIKMATAFATILY--IMTFV----------LL------LL-V---R--------L  280 CPANIFC
-LIGNVIKVQFLNIRDWPFGAATSIT--LTIVM----------GLMLLVYWRA-S---R-------LL  268 ECOPOTB
-ILPLKIYSMVKVGVSPEVNALATIL--LVLSL----------VMVIASQLIARD---K-------TK  257 ECOPOTC
-FWSTDMMQPIANLPVTIFKFAMSPF--AEWQQ--------LAWAGVLII----TLCVLLLNILA-R---V-------VF  290 ECOPSTA
-LDSASLYMPGNSITSALANEFAEAE--SGLHV--------AALMELGLILFVITFIVLAASKFM-I---M-------RL  312 ECOPSTC
-QMVLRGI-GRLDMGLATVGGVGIVI--LAIIL--------DRLTQAVGR----DSRSRGNRRWY-TTGPV-------GL  347 ECOPROW
-TAIFDQFQSTFNGPAANMLAGVLVLCCLGLLLLEAISRGRARYARVGSG----SARSQTPRRLS-PPLAALALLLPIAL  294 SMASFUB

SRERA--------GR  200 ECOCHLJ
RFKAG--------LS  226 ATUMODC
GRRVV--------GH  277 ECOCYST
WSSRY--------NG  278 SYNCYST
WKKTF--------HK  288 MPOMBPY
ENQEKRAQQEEHHEH  291 ECOCYSW
GRQSRI--------H  286 SYNCYSW
SI-RD--------DD  286 CPANIFC
NKKVE--------LE  275 ECOPOTB
GNTGD--------VK  264 ECOPOTC
AKNKH--------G-  296 ECOPSTA
AKNEG--------AR  319 ECOPSTC
LTRPF--------IK  354 ECOPROW
TALAL--------GV  301 SMASFUB
```

Fig. 2. Subcluster 1a for phosphate, sulfate, molybdate, glycine-betaine, spermine, and putrescine transport systems. This figure displays from top to bottom: the tree computed from distances determined by applying the UPGMA method to the multiple alignment of subsequences (see the Materials and methods section); multiple alignment of sequences generated by Treealign (Hein, 1990) in regions not needing long gaps to be aligned. The number at the right of the sequences is the residue number of the last amino acid in each line. Proteins are named as in Table 1. The branch lengths are proportional to evolutionary distances computed by the UPGMA method and are drawn to scale.

Remarkably, the branched-chain amino acid and the oligopeptide transport systems have the common characteristic of possessing 2 different genes for ATP-binding proteins (see Table 1) whose integrity is essential in the transport process. This observation suggests that each hydrophobic membrane protein recognized specifically a different ATP-binding protein and provides a possible explanation for the fact that protein partners are found in different clusters.

Cluster 7: Ribose, galactose, and arabinose transport systems (monosaccharides)

The arabinose and the galactose transport systems contain a single hydrophobic membrane protein ECOARAH and ECOMGLC, respectively (Scripture et al., 1987; Hogg et al., 1991). The transport system for ribose contains 2 proteins: ECORBSC and ECORBSD (Bell et al., 1986). ECORBSC is strongly similar to ECOARAH and ECOMGLC, whereas ECORBSD is not similar to any other protein of the sequence databases. ECORBSD is a short (139 residues) ORF, located 5′ proximal to the control region of the ribose transport operon. Because no experimental evidence exists for the translation of this ORF, nor for its involvement in ribose transport, and because it has no equivalent in the related arabinose and galactose transport systems, it is likely that ECORBSD does not function in transport. Therefore, we propose that this cluster contains systems with single hydrophobic membrane proteins. The strong similarity between the sequences of these proteins (Fig. 7) further suggests that they evolved from a common ancestor quite recently, perhaps within the enterobacterial family. The absence of similar sequences in other bacteria precludes a more precise dating of this event.

Cluster 8: Iron siderophore and cobalamine transport systems

Although the structures of the iron complexes transported by these systems are very different, the proteins are very similar. A multiple alignment of these sequences obtained by a different method was already published (Koster et al., 1991). The tree, computed from our alignment, showed that the 2 protein partners in the ferric enterobactin (FEP) and the ferric dicitrate (FEC) transport systems partition in 2 distinct sub-trees, the ECOFEPD-ECOFECC sub-tree and the ECOFEPG-ECOFECD sub-tree (Fig. 8). The cobalamine (BTU) transport system has 1 single hydrophobic protein ECOBTUC, which may be reminiscent of the common ancestor of these proteins. By contrast, the ECOFHUB protein, which is the single hydrophobic protein from the ferric-hydroxamate transport system, presents considerable internal similarity and results probably from the fusion of duplicated genes (Koster & Braun, 1986). The topology of the tree shows that the common ancestor of membrane proteins from the FEC, the FEP, and the BTU systems may derive from the duplicated ancestor of the C-terminal half of the ECOFHUB protein. The early divergence of the *E. coli* proteins from the *Vibrio anguillarum* proteins could be attributed to the phylogenetic distance between these 2 bacteria.

Proteins that do not fall into any cluster suggest the existence of new families of transport systems

Apart from ECORBSD, some proteins could not be related to the clusters described above. The MHYP69 protein is from a system of unknown function in *Mycoplasma hyorinis* (Dudler et al., 1988). ECOPHNM and ECOPHNE belong to a large operon (14 cistrons) involved in phosphonate transport and dissimilation in *E. coli* (Chen et al., 1990; Makino et al., 1991). There is no significant similarity between the 2 proteins, and it is probable that they belong to 2 different transport systems encoded in the same operon. This idea is substantiated by the fact that 3 putative ATP-binding proteins are found in this operon.

The existence of such orphan proteins suggests that new clusters of hydrophobic membrane proteins will be found in the course of the sequencing programs.

Discussion

This study establishes that hydrophobic membrane proteins from periplasmic transport systems partition into 8 clusters containing proteins that are related in terms of primary sequence. Within each cluster, the proteins very probably have a common evolutionary origin.

Proteins from widely diverse bacteria (Bacilli, Clostridia, Enterobacteriaceae, and Cyanobacteria) are found in a given cluster. This suggests that the ancestors of current clusters were present before the divergence of these bacterial groups.

No obligatory correlation exists between the distribution of the proteins into clusters and the nature of the transported substrates. For instance, proteins from cluster 1 belong to systems transporting oligosaccharides, glycerophosphate, organic, and inorganic anions. This finding suggests that the substrate recognition site of the proteins from this cluster may have progressively shifted from one specificity to another during evolution. By contrast, systems from clusters 3 and 4 transport similar substrates such as oligo- and dipeptides.

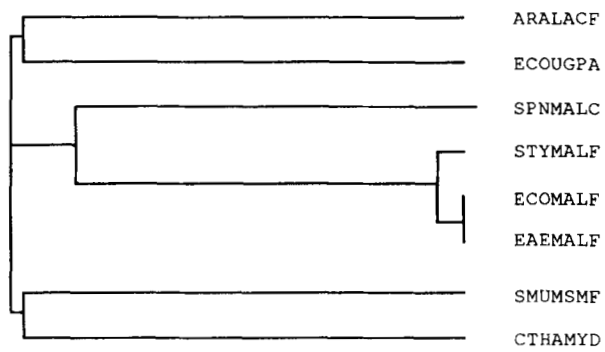
Recently, the study of sequence relationships among periplasmic substrate-binding proteins revealed that they also partition into 8 clusters (Tam & Saier, 1993), very similar to those found in this study. These authors describe clusters for oligosaccharides, inorganic polyanions, polar amino acids, oligopeptides, aliphatic amino acids, monosaccharides, and iron complexes, respectively, similar to clusters 1, 2, 3-4, 5-6, 7, and 8 found in this work. Some discrepancies may be explained by the fact that membrane protein partners in oligopeptide- and branched-chain amino acid transporters fall in distinct clusters. Within corresponding clusters, the topologies of phylogenetic trees for substrate-binding proteins and for hydrophobic membrane proteins are similar. This suggests that the genetic regions defining periplasmic transport systems rather than the individual genes are the target for evolutionary constraints.

A common modular organization for binding protein-dependent transport systems

Independently of the classification based on protein sequence similarities, transport systems may be classified according to the number and the size of inner membrane proteins.

Class I systems, such as the glutamine or the cobalamine transport systems, contain 1 hydrophobic protein (ECOGLNP, ECOBTUC) made of 200-300 amino acids. These systems have the PIA structure in Table 1. Secondary structure predictions indicate that the corresponding proteins have 5-6 potential transmembrane segments (Dassa & Saurin, unpubl. results).

Class II systems, such as the iron-hydroxamate transport system, also contain a single hydrophobic inner membrane protein (ECOFHUB), but this protein displays considerable internal




```
ALIGNMENT OF SEQUENCES:

STTNTGFWGLVIVTSWQMIGYVMVIYIAYIESIPTDLIEASKIDGANSWQQFRNVVFPLIAPAFTV-SLFITLSNS-FKL  227 CTHAMYD
GTANGAVIASIFVLLWQGVAMPIILFLSGLQSIPSEIVEAAAIDGADSKQTFWSVELPYLLPSISM-VFIMALKAG-LTA  227 SMUMSMF
TDPTWTKIALIMMQGWLGFPYIYVLTLGILQSIPNDLYEAAYIDGANAWQKFRNITFPMILAVAAP-TLISQYTFN-FNN  357 SPNMALC
SDPTTARTMLIIVNTWLGYPYMMILCMGLLKAIPDDLYEASAMDGAGPFQNFFKITLPLLIKPLTP-LMIASFAFN-FNN  440 ECOMALF
SDPTTARTMIIIVNTWLGYPYMMILCMGLLKAIPDDLYEASAMDGAGPFQNFFKITLPLLIKPLTP-LMIASFAFN-FNN  440 EAEMALF
SDPNTARAMVIIVNTWLGYPYMMILCMGLLKAIPDDLYEASAMDGAGPFQNFFKITLPLLIKPLTP-LMIASFAFN-FNN  440 STYMALF
TDPFWAKVLIIIAITWRWTGYNMIFYLAALQNIDRSIYEAAKIDGVPSWGRFAFLTIPMLKPVILFTTITSTIGTLQLFD  235 ARALACF
QNSGQAMFLVVFASVWKQISYNFLFFYAALQSIPRSLIEAAAIDGAGPIRRFFKIALPLIAPVSFF-LLVVNLVYAFFDT  229 ECOUGPA
              *                 *     **   **       *     *

FDQNLSLTAGAPG--NTTQMITLN------IYQTAFSAQEMAVGQAKAVIMFLIIAVISVIQVYLTQKREV----E-M  292 CTHAMYD
FDQIFALTGGGPN--NSTTSLGLL------VYNYAFKSNQYGYANAIALILFIIIGIVSVLQIKLSKK--F----E-V  290 SMUMSMF
FSIMYLFNGGGPG--SVGGGAGSTDILISWIYRLTTGTSPQYSMAAAVTLIISIIVISISMIAFKKLH--AFDMEE-V  430 SPNMALC
FVLIQLLTNGGPDRLGTTTPAGYTDLLVNYTYRIAFEGGGGQDFGLAAA-IATLIFLLVGALAIVNLK--ATRMKF-D  514 ECOMALF
FVLIQLLTNGGPDRLGTTTPAGYTDLLVSYTYRIAFEGGGGQDFGLAAA-IATLIFLLVGALAIVNLK--ATRMKF-D  514 EAEMALF
FVLIQLLTNGGPDRLGTTTPAGYTDLLVSYTYRIAFEGGGGQDFGLAAA-IATLIFLLVGALAIVNLK--ATRMKF-D  514 STYMALF
EVYNFTEGTGGPA--NSTLTLSLY------IYNLTFRFMPSFSYAATVSYVIVLMVAVLSFLQFYAAR--E----R-K  298 ARALACF
FPVIDAATSGGPV--QATTTLIYK------IYREGFTLDLASSAAQSVVLMFLVIVLTVVQFRYVESK--V----RYQ  293 ECOUGPA
         * *                  *

number of completely conserved sites: 11
```

Fig. 3. A: Subcluster 1b, the MalF subfamily for di-, oligosaccharides, and β-glycerophosphate transport systems. **B:** Subcluster 1c, the MalG subfamily for di-, oligosaccharides, and β-glycerophosphate transport systems. Presentation and conventions are the same as in Figure 2. Completely conserved positions are indicated by stars. (*Continues on facing page.*)

similarity. Its size is roughly twice the size of the proteins from the class I system, and it is made of 12 transmembrane helices (Koster & Braun, 1986). These systems have the PI^2A structure in Table 1.

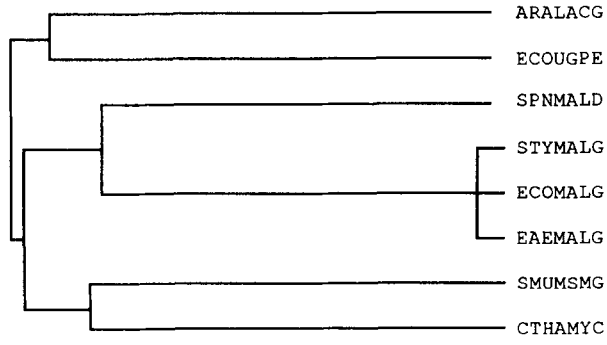
Class III systems have 2 hydrophobic inner membrane proteins with the size of the proteins from class I with few exceptions. These systems have the PII'A or the P2IA structure in Table 1.

The following observations strongly suggest that genes encoding protein partners from class III systems arise by tandem duplication of an ancestor gene encoding a class I protein: (1) Almost all protein partners of a given system are found in the same cluster or subcluster, showing a clear phylogenetic relatedness. (2) The genes for hydrophobic membrane protein partners are always contiguous within operons. (3) The genes for homologous protein partners in the same cluster or subcluster occupy the same relative positions within their respective operons. (4) Single hydrophobic membrane proteins are commonly found in the same clusters as systems with 2 proteins and diverged near or at the root of the tree.

Examples of such tandem duplication events are found in clusters 2 (histidine, glutamine, and opines) and 8 (iron-siderophores) and in subcluster 1a (ions).

The study of the distribution of hydrophobic protein partners from a given transport system within clusters provides additional information. The protein partners of the peptide transport systems are found in 2 different clusters, those from the maltose transport system are in the same cluster but in different subclusters, whereas those from the sulfate transporters are in the same subcluster. Because the degree of dissimilarity between 2 sequences may be related to the time at which they have diverged from a common ancestor, it is thus likely that these tandem duplication events occurred several times during evolution.

These observations suggest that current periplasmic transport systems may derive from an ancestral system comprising a single hydrophobic membrane protein (the basic module) made of




```
ALIGNMENT OF SEQUENCES:

GIPTSLDEAALIDGCSRFRIYWNIILPLLNPTTITLAVLDIMWIWNDYLLPSLVINKVGS-RTLPLMIFYF--FS-QYTK  236 CTHAMYC
SVPDSLDEAAEIDGADKLTTYRKIIFPMLKPMHATTLIINALWFWNDFMLPLLILNKDSSMWTLPLFQYNY--SG-QYFN  240 SMUMSMG
AFPTELRDAAKVDGLKEWQIFFYIYVPVMRSTYAAAFVIVFMLNWNNYLWPLIVLQSNDT-KTITLVVSSL--AS-AYSP  236 ARALACG
TLPDELVEAARIDGASPMRFFCDIVFPLSKTNLAALPVITFIYGWNQYLWPLLIITDVDL-GTTVAGIKGMIATG-EGTT  245 ECOUGPE
TVPMSLDESAKLDGAGHFRRFWQIVLPLVRPMVAVQALWAFMGPFGDYILSSFLLREKEY-FTVAVGLQTF--VNNAKNL  239 SPNMALD
TIDSSLEEAAALDGATPWQAFRLVLLPLSVPILAVVFILSFIAAITEVPVASLLLRDVNS-YTLAVGMQQY--LN-PQNY  258 ECOMALG
TIDSSLEEAAALDGATPWQAFRLVLLPLSVPILAVVFILSFIAAITEVPVASLLLRDVDS-YTLAVGMQQY--LN-PQNY  258 STYMALG
TIDGSLEEAAALDGATPWQAFRLVLVPLSVPILAVVFILSFIAAITEVPVASLLLRDVNS-YTLAVGMQQY--LN-PQNY  258 EAEMALG
     *   *  **            *                                      *

QWNLGMAGLTIAILPVVIFYFLAQRKLVTAIIAGAVKQ  274 CTHAMYC
DYGPSFASYIVGIITITIVYLIFQKHIIAGMSNGAVK-  277 SMUMSMG
EYGTVMIGTILATLPTLLVFFAMQRQFVQGMLGSV--K  272 ARALACG
EWNSVMVAMLLTLIPPVVIVLVMQRAFVRGLVDSE--K  281 ECOUGPE
KIAYFSAGAILIALPICILFFFLQKNFVSGLTSGGDKG  277 SPNMALD
LWGDFAAAAVMSALPITIVFLLAQRWLVNGLTAGGVKG  296 ECOMALG
LWGDFAAAAVLSAIPITLVFLLAQRWLVNGLTAGGVKG  296 STYMALG
LWGDFAAAAVLSAIPITVVFLLAQRWLVNGLTAGGVKG  296 EAEMALG
                        *

number of completely conserved sites: 7
```

Fig. 3. *Continued.*

200–300 residues with 5–6 transmembrane helices. Systems from class I are most similar in design to this ancestral system. The basic module is duplicated in systems from class III. Class II systems may derive from class III systems by the genetic fusion of already duplicated modules.

A model for the evolution of binding-dependent transport systems

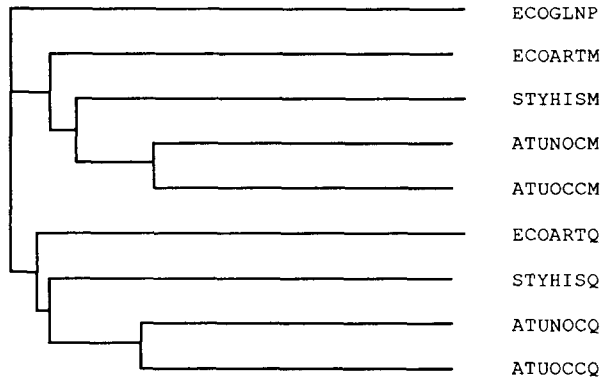
From this analysis of sequence relationships between hydrophobic membrane components, we propose a 3-step evolutionary scheme that accounts for the wide diversity of current binding protein-dependent transport systems (Fig. 9).

Periplasmic binding protein-dependent transport systems evolved by duplication of an ancestral transporter present in the eubacterial common ancestor

We propose that the genes encoding the class I ancestor proteins of each cluster could be themselves phylogenetically related. Such an idea is suggested by the fact that hydrophobic membrane proteins are homogeneous in terms of length and predicted secondary structure, that they are encoded in operons with a strikingly similar organization, and that they constitute systems with a common mechanism of energy coupling. Moreover, limited regions of similarity were found between hydrophobic proteins from different clusters as for instance between cluster 1 and 2. The relatively small size of our set of sequences and the limited set of bacterial species from which sequences are available do not allow us to demonstrate intercluster relationships. Therefore, the genes encoding these ancestor proteins could derive from a single ancestral gene by successive duplications. These duplications probably involved the whole genetic region containing the genes for the substrate-binding protein, the hydrophobic membrane protein, and the ATP-binding protein.

Hydrophobic membrane protein partners in class III transport systems are the result of a tandem duplication event

This tandem duplication event, discussed above, probably took place after the first round of long-range duplications leading to cluster-specific single hydrophobic membrane proteins. Alternately, if the ancestral gene had been tandemly duplicated before the long-range duplication events, then the class I trans-




```
M--------------F-E-YLP--------ELMKGL-----------HTSLTLTVASLIVALILALIFTIILTLKTPVLV   45 ECOARTM
M--------------I-E-IIQ--------EYWKSLLWTDGYRFTGVAITLWLLISSVVMGGLLAVILAVGRVSSNKFIR   56 STYHISM
M---------DIQLII-E-SFP--------KLLAAV-----------PTTLTLAFISLLIGFVVSVPVALMRLSKNRIVS   50 ATUNOCM
M-------PFDPAFLW-Q-TFV--------ALLSGI-----------PLALQLAVFSVALGTVLAFGLALMRVSRLWWLD   52 ATUOCCM
M--------------N-E-FFP---------LASAA-----------GMTVGLAVCALIVGLALAMFFAVWESAKWRPVA   44 ECOARTQ
M--------------LYG-FSG--------VILQGA-----------IVTLELALSSVVLAVLIGLVGAGAKLSQNRVTG   46 STYHISQ
M--------------DLT-LLQWGDAGWGDELARGA-----------MMTVVVAACSYFFGIIFGSLFAAAKLSRFWSLR   54 ATUNOCQ
M--------------DYSQLMGFGPDGWGYDMLRAT-----------AMTMAVAFSGFTIGLVFGCLGAAASLSSSGALQ   55 ATUOCCQ
MQFDWSAI-------W-P-AIP--------LLIEGA-----------KMTLWISVLGLAGGLVIGLLAGFARTFGGWIAN   52 ECOGLNP
*

WLVRGYITLFTGTPLLVRIFLIYYGPG-QF-PTLQEY--------------PALWHLLSEP-WLCALIALSVNSAAYTTQ  108 ECOARTM
FPIWLFTYIFRGTPLYVQLLVFYSGMY-TL-EIVKGT--------------DLLNAFFRSG-LNCTVLALTLNTCAYTTE  119 STYHISM
SLAYGYVYIIRSTPLLVQMFLIYYGSA-QF-RGVLSE--------------VGLWSSFREP-WFCAILALALNTAAYTSE  113 ATUNOCM
LPARFYIFAFRGTPLLVQIYIIYYGLS-QF-PDVRHS--------------F-IWPFLRDA-YWCAMAALALNTAAYTAE  114 ATUOCCM
WAGSALVTILRGLPEILVVLFIYFGSS-QL-LLTLSDGFTINLGFVQIPVQMDIENFDVSP-FLCGVIALSLLYAAYASQ  121 ECOARTQ
LIFEGYTTLIRGVPDLVLMLLIFYGLQIAL-NVVTDS--------------LGIDQIDIDP-MVAGIITLGFIYGAYFTE  110 STYHISQ
LLGDVYTTVVRGVPELLIIFLVFFGGGTLLRTIANGL--------------FGYEGYIEPPIFVIGVLCISVSAGAYATE  120 ATUNOCQ
AAASGYTTALRGIPDLLVIYLFYFGSSSVI-SNVASL--------------FGSSGFVGASTFLIGALAIGVVSGAYQTQ  120 ATUOCCQ
HVALVFIEVIRGTPIVVQVMFIYFAL----------P--------------MAFNDLRIDP-FTAAVVTIMINSGAYIAE  107 ECOGLNP
          •                                                              **

LFYGAIRAIPEGQWQSCSALGMSK-KDTLAILLPYAFKRSLSSYSNEVVLVFKSTSLAYTITLMEVMGYSQLLYGRTYD-  186 ECOARTM
IFAGAIRSVPHGEIEAARAYGFSSFKMYRCIILPSALRIALPAYSNEVILMLHSTALAFTATVPDLLKIARDINSATYQP  199 STYHISM
IIRGGIQSVSLGQIEAARAVGMSTFLQFRRIVFPIAIRQALPAYGNEVMLIIKSTSLASTITIVEVTGLAKQIISATYSP  193 ATUNOCM
IMRGGLLSVPAGQIEAAKACGMGRVKLFRRIVIPQAIRQMLPGYSNEVILMVKSTSLASTITIMEITGIAAKLISESYRT  194 ATUOCCM
TLRGALKAVPVGQWESGQALGLSKSAIFFRLVMPADVRHALPGLGNQWLVLLKDTALVSLISVNDLMLQTKSIATRTQEP  201 ECOARTQ
TFRGAFMAVPKGHIEAATAFGFTHGQTFRRIMFPAMMRYALPGIGNNWQVILKATALVSLLGLEDVVKATQLAGKSTWEP  190 STYHISQ
VIRAAVLAVPPGQIEAAKSIGMGPWLRLRRVLIPQAARFALPGLGNVWQFTLKDTSLISVVGLVEIMRTAAMGAGSTKQP  200 ATUNOCQ
VLRGAVLALNKGEIEAGRAYGMGALLLFRRIVLPQAARYALPGVGNVWQLVLKESALISVIGLVELMRQAQVGSGSTRQP  200 ATUOCCQ
ITRGAVLSIHKGFREAGLALGLSRWETIRTVILPLALRRMLPPLGNQWIISIKDTSLFIVIGVAELTRQGQEIIAGNFRA  187 ECOGLNP
          *       •           *     *   *         *

VMVFGAAGIIYLVVNGLLTLMMRLIE-----------RKAV---AFERR-N-  222 ECOARTM
FTAFGIAAVLYLLISYVLISLFRRAE-----------RRWL---QHVSS-K-  235 STYHISM
VEVFIVAGAIYLFITFVVSRLVMLAEWWLNPHMRARVGGTAP---KAAET-H-  241 ATUNOCM
VEVFSCAGAIYLILNFIVARLFTLLEWALWPERRNNRLTTDPVDRKGELH-A-  245 ATUOCCM
FTWYIVAAAIYLVITLLSQYILKRID-----------LRAT---RFERRPS-  238 ECOARTQ
FYFAVVCGLIYLVFTTVSNGVLLLLE-----------RRYS---VGVKRADL  228 STYHISQ
FTFYITAFVIFLLLSSVSNRGFLKAE-----------KWAN---RGV-RSQ-  236 ATUNOCQ
FSFYLTAAALYLLITFVSGQVFRLAE-----------TRSM---RGLQRGV-  237 ATUOCCQ
LEIWSAVAVFYLIITLVLSFILRRLE-----------RR--------MKI-L-  219 ECOGLNP
          *

number of completely conserved sites: 11
```

Fig. 4. Cluster 2 for the histidine, glutamine, arginine, nopaline, and octopine transport systems. Presentation and conventions are the same as in Figure 2.

```
            401                                                    450
ATUOCCM     lsVpagqiEA AkacGmgrvk lFRrIviPqa irqmlpgysn evilmvksts
ATUNOCM     qsVslgqiEA AravGmstFl qFRrIvFPia irqalpaygn evmliiksts
STYHISM     rsVphgeiEA ArayGfssFk mYRcIiLPsa lrialpaysn evilmlhsta
ECOARTM     raIpegqwqs csalGmskkd tL.aIlLPya fkrslssysn evvlvfksts
ATUOCCQ     lalnkgeiEA grayGmgaLl lFRrIvLPqa aryalpgvgn vWqlvlkesa
ATUNOCQ     laVppgqiEA AksiGmgpWl rLRrVliPqa arfalpglgn vWqftlkdts
STYHISQ     maVpkghiEA AtafGfthgq tFRrImFPam mryalpgign nWqvilkata
ECOARTQ     kaVpvgqwEs gqalGlsksa iFFrlvMPad vrhalpglgn qWlvllkdta
CTHAMYD     esIptdliEA skidGansWq qFRnVvFPli apaftvslfi tLs.ns....
SMUMSMF     qsIpseivEA AaidGadskq tFWsVeLPyl lpsismvfim aLk.ag....
ECOUGPA     qsIprsliEA AaidGagpir rFFkIaLPli apvsffllvv nLvyaf....
ARALACF     qnIdrsiyEA AkidGvpsWg rFafltiPml kpvi.lftti tstigt....
EAEMALF     kaIpddlyEA samdGagpFq nFFkItLPll ikpltplmia sFafnfnnfv
STYMALF     kaIpddlyEA samdGagpFq nFFkItLPll ikpltplmia sFafnfnnfv
ECOMALF     kaIpddlyEA samdGagpFq nFFkItLPll ikpltplmia sFafnfnnfv
SPNMALC     qsIpndlyEA AyidGanaWq kFRnItFPmi lavaaptlis qYtfnfnnfs
ECOMALG     etIdssleEA AaldGatpWq aFRlVlLPls vpilavvfil sFiaaitevp
STYMALG     etIdssleEA AaldGatpWq aFRlVlLPls vpilavvfil sFiaaitevp
EAEMALG     etIdgsleEA AaldGatpWq aFRlVlvPls vpilavvfil sFiaaitevp
SPNMALD     dtVpmsldEs AkldGaghFr rFWqIvLPlv rpmvavqalw aFmgpfgdyi
CTHAMYC     .gIptsldEA AlidGcsrFr iYWnIiLPll npttitlavl dimwiwndyl
SMUMSMG     lsVpdsldEA AeidGadkLt tYRkIiFPml kpmhattlii nalwfwndfm
ARALACG     kafptelrDA AkvdGlkeWq iFFyIyvPvm rstyaaafvi vFmlnwnnyl
ECOUGPE     mtlpdelvEA AridGaspmr fFcdIvFPls ktnlaalpvi tFiygwnqyl
ECOCYSW     lsqgsqedEA AillGasgWq mFRrVtLPni rwallygvvl tnaraigefg
SYNCYSW     eeIgtdaeEA AstlGangWq tFWrVtLPsi kwsmlygvvl ttaralgefg
SYNCYST     leleveaeEA AaslGaspse tFWrViLPpi lpgvlagvaq gFsravgefg
MPOMBPY     qnmeedleEA AwclGaspWt tFWhIlFPpl tpslltgttl gFsralgeyg
ECOCYST     eelgpeyeEA AetlGatrWq sFckVvLPel spalvagval sFtrslgefg
ECOCHLJ     egVdvkleqA ArtlGagrWr vFFtItLPlt lpgiivgtvl aFarslgefg
AVIMODC     eaIgerplEv AstlragpWd tFFtVvvPla rpgfitaail gFahtvgefg
CPANIFC     ksVpielfEv syvlGagkie tiikImiPml kksivsglil aWirslgefg
ECOPOTB     ekldkpllEA ArdlGaskLq tFirIiiPlt mpgiiagcll vmlpamglfy
ECOPOTC     kgfdvrmlEA AkdlGaseFt iLRkIiLPla mpavaagwvl sFtlsmddvv
ECOPSTC     eqtpvmmkEs AygiGcttWe viWrIvLPft kngviggiml gLgralgetm
ECOPSTA     klVpyslrEA AyalGtpkWk misaItLkas gsgimtgill aiariageta
ECOPROW     nqVpadliEA srsfGasprq mLFkVqLPla mptimagvnq tLmlalsmvv
Consensus   --I-----EA A---G---W- -FR-I-LP-- ---------- -F--------
```

Fig. 5. Alignment of the most conserved region between sequences of proteins from clusters 1 and 2. This alignment was generated using the program Pileup. The consensus was obtained for a plurality of 26 ensuring a 75% conservation. Amino acids satisfying the consensus are in capitals. The default amino acid equivalence table of the program was used.

port systems would be generated by deleting 1 of the 2 genes from class III transport systems. In those cases, class I proteins would be more related to one of the partners of class III proteins and would not display early divergence in trees. This was observed only in the case of the MPOMBPY chloroplast protein.

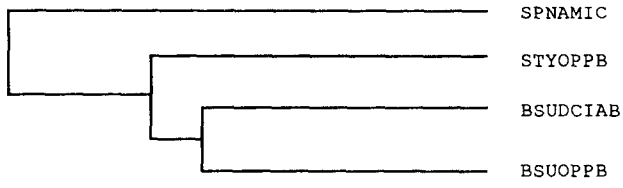
Further specialization toward substrates is achieved by a second round of duplications involving the whole transport region

Systems transporting different substrates in the same organism and falling in the same cluster very probably evolved by duplication of a parent system. This might be the case for the oligopeptide and the dipeptide transport systems in *B. subtilis*, for the FEP and FEC transport systems in *E. coli*.

Tandem duplications and the evolution of binding protein-dependent transport systems

Tandem duplication events are not limited to the hydrophobic membrane proteins. In the histidine transport system, the HISJ histidine-binding protein and the LAO-binding protein for lysine, arginine, and ornithine are contiguous and display strong similarity. This is also the case for the 2 ATP-binding proteins in the OPP and the LIV transport systems. The fact that pairs of genes for hydrophobic proteins or for ATP-binding proteins are contiguous suggests that the individual genes, rather than the operons underwent these tandem duplication events.

Gene duplications were described as playing a major role in adaptive evolution. The evolutionary advantage of such events was interpreted as an increase in the evolutionary potential of the organism in which it happens (Ohno, 1974; Rigby et al., 1974). In the case of dimeric systems such as those involved in binding protein-dependent transport, this evolutionary advantage might explain the frequent occurrence of tandem duplications. A molecular description of this evolutionary advantage might be the following. In the course of their specialization toward different substrates, genes encoding single hydrophobic membrane proteins are subjected to mutational events. Some of the fixed mutations altering the substrate specificity might have adverse effects on the function of the protein, on the stabil-




```
ALIGNMENT OF SEQUENCES:

SYGKDDPYTATESNYQYPSMIVSSAITGLIGLVLAYALAVPLGSAMARFKNTWIDSLSTGALTFLLALPTIALVYIVRLI  336 SPNAMIC
DFGPSIKKPSDSVNDMLERGFPVSFELGMTAIVIAVISGLVLGVIAALRRNGFLDYAAMSLAVLGISIPNFILATLLIQQ  153 BSUDCIAB
DFGPSFKYKDYTVNDLVAASFPVSAKLGAAAFLLAVIIGVSAGVIAALKQNTRWDYTVMGFAMTGVVIPSFVVAPLLVMV  153 STYOPPB
DFGPSFKYKGQSVNDLISSGFPVSFTLGAEAILLALALGVLFGVIAALYHNKWQDYTVAILTIFGISVPSFIMAAVLQYV  153 BSUOPPB
  *          *          *   *       *        *   *   *   *              *

GSSIALPDSFPILGAGDWRSYVLPAVILGLLGAPGTAIWIRRYMIDLQSQDFVRFARAKGLSEKEISNKHIFKNAMVPLV  416 SPNAMIC
FAVNLKLFPAATWTSPI--HMVLPTAALAVGPMAIIARLTRSSMVEVLTQDYIRTAKAKGLSPFKIIVKHALRNALMPVI  231 BSUDCIAB
FAITLQWLPGGGWNGGALKFMILPMVALSLAYIASIARITRGSMIEVLHSNFIRTARAKGLPMRRIIFRHALKPALLPVL  233 STYOPPB
FSMKLGLFPVAGWDSWA--YTFLPSIALASMPMAFIARLSRSSMIEVLNSDYIRTAKAKGLSAQRLQCGTPFETHFCRLL  231 BSUOPPB
                        **   *         *   *  *          * * ****

SGIPAAIIGVIGGATLTETVFAFPGMGKMLIDSVKASNNSMVVGLVFIFTCISIFSRLLGDIWMTIIDPRIKLTEK---G  493 SPNAMIC
TVLGTLVASILTGSFVIEKIFAIPGMGKYFVESINQRDYPVIMGTTVFYSVILIIMLFLVDLAYGLLDPRIKLHKK---G  308 BSUDCIAB
SYMGPAFVGIITGSMVIETIYGLPGIGQLFVNGALNRDYSLVLSLTILVGALTILFNAIVDVLYAVIDPKIRY-------  306 STYOPPB
HILGPMAAQVLTGSFIIETIFGIPGLGAHFVNSITNRDYTVIMGVTVFFSVILLLCVLIVDVLYGIIDPRIKLSKAKKGA  311 BSUOPPB
            *    *     ** *                                        *      ** *

GK  495 SPNAMIC
--  308 BSUDCIAB
--  306 STYOPPB
--  311 BSUOPPB

number of completely conserved sites: 31
```

Fig. 6. A: Cluster 3, the OPPB family for di- or oligopeptide transport systems. **B:** Cluster 4, the OPPC family for di- or oligopeptide transport systems. Presentation and conventions are the same as in Figure 2. (*Continues on facing page.*)

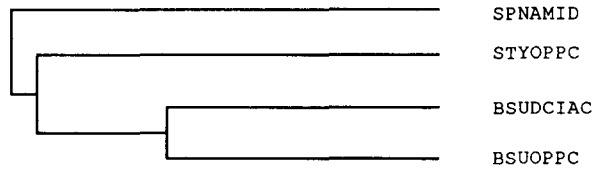
ity of its dimeric state, or on its interaction with the other proteins of the system. The evolutionary advantage provided by the duplication of these genes is clear if one considers that it enlarges the size of the target for mutational events compensating these adverse effects. When the 2 copies of the gene bear mutually compensating mutations, the duplicated state becomes irreversible.

Is there a common origin for ABC transporters?

ABC transporters constitute a superfamily of systems sharing a highly conserved domain involved in ATP binding and hydrolysis. According to this definition, periplasmic binding protein-dependent transport systems are members of this superfamily. The ATP-binding proteins or domains are highly conserved and it is very likely that they have a common phylogenetic origin. In contrast, it is remarkable that we did not detect any similarity between the group of proteins studied here and the hydrophobic components of other eukaryotic and prokaryotic ABC transporters. This lack of similarity can be explained by 2 hypotheses.

First, the hydrophobic components of ABC transporters might have a common ancestor. Because there is evidence that hydrophobic membrane proteins have recognition sites for substrates (Treptow & Shuman, 1985), they are therefore likely to be subjected to a wide variety of constraints in order to accommodate the various substrates. These constraints are clearly different from those that apply to the ATP-binding proteins that share the same substrate. Thus, the homology between hydrophobic membrane components may be scarcely detectable, hampering the phylogenetic reconstruction. Nevertheless, because ATP-binding domains are genetically linked to hydrophobic proteins, they would display the same phylogenetic pattern.

Alternatively, hydrophobic membrane components of ABC transporters may have several unrelated ancestors. It has been proposed that some ABC transporters, such as the systems catalyzing export of drugs and carbohydrates, form a subfamily ABC1 within the superfamily of ABC transporters (Reizer et al., 1992). The periplasmic binding protein-dependent transporters might constitute another subfamily. If these proposals are true, it is very likely that the ATP-binding module has been recruited during evolution to assume a similar role (e.g., ATP-binding or hydrolysis) in otherwise functionally different multimeric complexes. This hypothesis would be strengthened if the phylogenetic trees of ATP-binding proteins appeared to have topologies different from those of the hydrophobic proteins.




```
ALIGNMENT OF SEQUENCES:

MS--------------TIDKEKFQFVKRDDFASETIDAPAYSYWKSVFKQFMKKKSTVVMLGILVAIILISFIYPMFSKF   66 SPNAMID
MNLPVQTDERQPEQHNQVPDEWFVLNQEKNREADSVKRPSLSYTQDAWRRLKKNKLAMAGLFILLFLFVMAVIGPFLSPH   80 BSUDCIAC
MM--------------LSKKNSETLEN----FSEKLEVEGRSLWQDARRRFMHNRAAVASLIVLFLIALFVTVAPMLSQF   62 STYOPPC
MQ--------------NIPKNMFEPAAANAGDAEKISKKSLSLWKDAMLPFRSNKLAMVGLIIIVLIILMAIFAPMFSRY   66 BSUOPPC
*                                        *                 *             * *

DFNDVSKVNDFSVRYIKPNAEHWFGTDSNGKSLFDGVWFGARNSILISVIATVINLVIGVFVGGIWG-ISKSVDRVMMEV  145 SPNAMID
SVVRQSLTEQNLPP----SADHWFGTDELGRDVFTRTWYGARISLFVGVMAALIDFLIGVIYGGVAGYKGGRIDSIMMRI  156 BSUDCIAC
TYFDTDWGMMSSAPDM--ASGHYFGTDSSGRDLLERVAIGGRISLMVGIAAALVAVIVGTLYGSLSGYLGGKIDSVMDAF  140 STYOPPC
DYSTTNLLNADKPP----SKDHWFGTDDLGRDIFVRTWVGARISIFIGVAAAVLDLLIGVIWGSISGFRGGRTDEIMMRI  142 BSUOPPC
                      * **** *         * * *      *       *   *   *      * *

YNVISNIPPLLIVIVLTYSIGAGFWNLIFAMSVTTWIGIAFMIRVQILRYRDL---EYNLASRTLGTPTLKIVAKNIMPQ  222 SPNAMID
IEVLYGLPYLLVVILLMVLMGPGLGTIIVALTVTGWVGMARIVRGQVLQIKNY---EYVLASKTFGAKTFRIIRKNLLRN  233 BSUDCIAC
VEILNSFPFMFFVILLVTFFWA---EHSVDFRSPSAWSPGLIWRVSLWPNPNLKRKEFIEAAQVGGVSTASIVIRHIVPN  217 STYOPPC
ADILWAVPSLLMVILLMVVLPKGLFTIIIAMTITGWINMARIVRGQVLQLKNQ---EYVLASQTLGAKTSRLLFKHIVPN  219 BSUOPPC
       *    ** *                              *           *   *    *  *

LVSVIVTTMTQMLPSFISYEAFLSFFGLGLPITVPSLGRLISDYSQN-VTTNAYLFWIPLTTLVLVSLSLFVVGQNLADA  301 SPNAMID
TMGAIIVQMTLTVPAAIFAESFLSFLGLGIQAPFASWGVMANDGLPTILSGHWWRLFFPAFFISSTMYAFNVLGDGLQDA  313 BSUDCIAC
VLGVVVVYASLLVPSMILFESFLSFLGLGTQEPLSSWGALLSDGANS-MEVSPWLLLFPAGFLVVTLFCFKLYCDGLRDA  296 STYOPPC
AMGSILVTMTLTVPTAIFTEAFLSYLGLGVPAPLASWGTMASDGLPA-LTYYPWRLFFPAGFICITMFGFNVVGDGLRDA  298 BSUOPPC
             *  •  * ***  ***      * *    *               *                 * **

SDPRTHR  308 SPNAMID
LDPKLRR  320 BSUDCIAC
LDPK-DR  302 STYOPPC
LDPKLRK  305 BSUOPPC
 **

number of completely conserved sites: 47
```

Fig. 6. *Continued.*

Materials and methods

Proteins

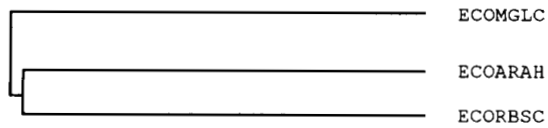
Hydrophobic membrane protein sequences from binding protein-dependent transporters were collected either by screening data banks or by the survey of the literature. In order to ensure that related proteins were not overlooked, each protein was used to search complete nonredundant nucleic acid databases translated in the 6 reading frames (Genbank Release 76.0, EMBL Data Library Release 35.0, and EST Data Library Release 35.0) by using the program tblastn (Altschul et al., 1990). Sequences giving a score better than 83 ($P < 0.0019$) were retained for this study. Table 1 describes the sequences that have been submitted to the phylogenetic analysis.

Computer methods

The rationale of the method was to group significantly similar protein sequences in clusters and to reconstruct their evolutionary history. To reconstruct these phylogenies, we had to define a multiple alignment of the sequences from each cluster. In spite of the similarities within a given cluster, this was not always possible. In this case, the cluster was further subdivided into subclusters with a higher internal similarity.

Computation of similarity scores between sequences

The classification process bears upon the similarity between sequences. We defined the similarity score between 2 sequences as the highest scoring pair of consecutive amino acid runs taken from each sequence, also called the MSP score (Altschul et al.,




```
ALIGNMENT OF SEQUENCES:

MMSSVSTSGSGAPKSSFSFGRIWDQYGMLVVFAVLFIACAIFVPNFATFINMKGLGLAISMSGMVACGMLFCLASGDFDL   80 ECOARAH
M-----TTQTVSGRRYFTKAWLMEQKSLIALLVLIAIVSTL-SPNFFTINNLFNILQQTSVNAIMAVGMTLVILTSGIDL   74 ECORBSC
M-----SALNK--KSFLTYLKEGGIYV--VLLVLLAIIIFQ-DPTFLSLLNLSNILTQSSVRIIIALGVAGLIVTQGTDL   70 ECOMGLC
*                                    *      * *    *        *     * *          **

SVASVIACAGVTTAVVINLTESL------WIGVAAGL------LLGVLCGLVNGFVIAKLKINALITTLATMQIVRGLAY  148 ECOARAH
SVGSLLALTGAVAAYIVGIEVNA------LVAVAAAL------ALGAAIGAVTGVIVAKGRVQAFIATLVMMLLLRGVTM  142 ECORBSC
SAGRQVGLAAVVAATLLQSMDNANKVFPEMATMPIALVILIVCAIGAVIGLINGLIIAYLNVTPFITTLGTMIIVYGINS  150 ECOMGLC
*            *                       *        *   *   *   *       * **  *    *

IISDGKAVG---IEDESFFA-LGYAN-WFGLPAPIWLTVACLIIFGL---LLNKTTFGRNTLAIGGNEEAARLAGVPVVR  220 ECOARAH
VYTNGSPVNTGFTENADLFGWFGIGR-PLGVPTPVWIMGIVFLAAWY---MLHHTRLGRYIYAVGGNEAATRLSGINVNK  218 ECORBSC
LYYDFVGASPISGFDSGFST-FAQGFVALGSFRLSYITFYALIAVAFVWVLWNKTRFGKNIFAIGGNPEAAKVSGVNVGL  229 ECOMGLC
                                 *                          *  *    * ***  *    *  *

TKIIIFVLSGLVSAIAGIILASRMTSGQPMTSIGYELIVISACVLGGVSLKGGIGKISYVVAGILILGTVENAMNLLNIS  300 ECOARAH
IKIIVYSLCGLLASLAGIIEVARLSSAQPTAGTGYELDAIAAVVLGGTSLAGGKGRIVGTLIGALILGFLNNGLNLLGVS  298 ECORBSC
NLLMIYALSGVFYAFGGMLEAGRIGSATNNLGFMYELDAIAACVVGGVSFSGGVGTVIGVVTGVIIFTVINYGLTYIGVN  309 ECOMGLC
       * *      *     *  *        ***  * * * ** *  ** *       *  *

PFAQYVVRGLILLAAVIFDRYKQKAKRTV  329 ECOARAH
SYYQMIVKAVVILLAVLVDNKKQ------  321 ECORBSC
PYWQYIIKGAIIIFAVALDSLKYA--RKK  336 ECOMGLC
   *          **  *  *
number of completely conserved sites: 56
```

Fig. 7. Cluster 7, ribose, galactose, and arabinose transport systems. Presentation and conventions are the same as in Figure 2.

1990). These scores were computed using a PAM120 matrix generated by the PAM program distributed with the blast package (Altschul et al., 1990). We computed MSP scores with program simil.c (Saurin, unpubl. and available upon request). By contrast with blast, which performs heuristic searches, simil.c performs an exact search of MSP scores between 2 sets of sequences.

Statistical significance of the similarity scores

The statistical significance of the scores was computed according to Karlin and Altschul (1990) using a C function kindly provided by these authors. We observed that scores greater or equal to 88 had a statistical significance greater than 3.26×10^{-7}, insuring a global significance greater than 7.86×10^{-4} for the 2,415 computed scores. We thus used the value of 88 as a threshold in the clustering process.

Clustering the sequences

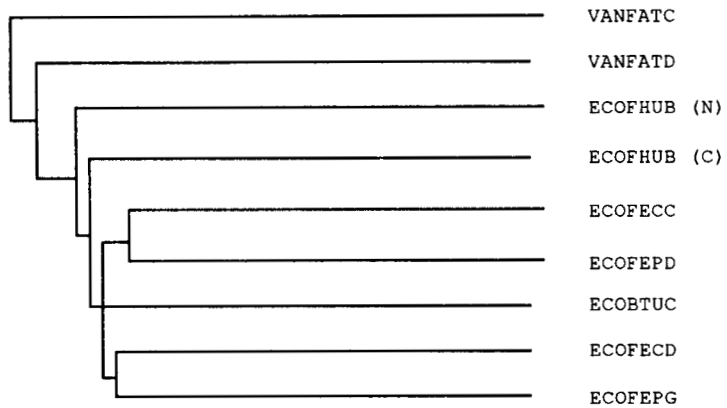
The sequences were grouped into clusters with the program linkage.c, an implementation of the single linkage classification algorithm (Johnson, 1967). This algorithm puts each sequence in a different cluster, then it iteratively joins 2 clusters into 1 when they display the highest similarity among all the possible pairs of clusters. The similarity between 2 clusters is the highest similarity observed between 2 sequences taken from each cluster. The process stops when all the similarities between clusters are less than a threshold value.

Multiple alignments of sequences

Program Treealign (Hein, 1990) was used to compute multiple alignments of the sequences from the previously defined clusters. The homology matrix built in the program was used and the gap penalty was set to 8 + 3L, where L is the length of the gap. We discarded the regions of the sequences presenting long gaps and realigned the remaining segments. In one case (cluster 1, see below) it was not possible to compute a multiple alignment for the whole cluster. Subclusters of cluster 1 were defined by increasing the threshold value.

Reconstruction of phylogenies

We used the UPGMA method (Sokal & Michener, 1958) to reconstruct a phylogenetic tree of each set of aligned sequences. The percentage of amino acids differing between 2 sequences in the aligned regions was taken as an estimation of the distance between these sequences.




```
ALIGNMENT OF SEQUENCES:

SPGDWFTPRGELFVW-QIRLPRTLAVLLVGAALAISGAVMQALFENPLAEPGLLGVSNGAGVGLIAAVLLG-QGLTP--N  114 ECOBTUC
AWSPDIDVIEQMIFH-YSLLPRLAISLLVGAGLGLVGVLFQQVLRNPLAEPTTLGVATGAQLGITVTTLWA-IPGAM--A  116 ECOFHUB(N)
HGWTWASG-ALLEDLMPWRWPRIMAALFAGVMLAVAGCIIQRLTGNPMASPEVLGISSGAAFGVVLMLFLVPGNAFG--W  448 ECOFHUB(C)
LLPGHTPTLPEALVQ-NLRLPRSLVAVLIGASLALAGTLLQTLTHNPMASPSLLGINSGAAWLWRYQRAES-DADCRLFS  120 ECOFECC
LLTDWQAGREHYYVLMEYRLPRLLLALFVGAALAVAGVLIQGIVRNPLASPDILGVNHAASLASVGALLLMPSLPVM--V  110 ECOFECD
AFSGTCQSADCTIVL-DARLPRTLAGLLAGGALGLAGALMQTLTRNPLADPGLLGVNAGASFAIVLGAALF-GYSSAQEQ  120 ECOFEPD
AALMGDAPRSMTMVVTEWRLPRVLMALLIGAALGVSGAIFQSLMRNPLGSPDVMGFNTGAWSGVLVAMVLF-GQDLT--A  117 ECOFEPG
SLLPTFNEKAWLPII-ASRLPRLVALILTGSGLAMCGVILQHIVRNRFVEPGTTGSLDAAKLGILVSIVML-PSSDK--L  102 VANFATD
AFIFINSGFDLEYII-PRRLIKLSAIIIGGSCVAISAVIFQALARNRILTPSIMG-YESIYLVWQALLLLF-VGTSG--S   95 VANFATC
                                *           *    *    *

WALGLCAIRGALIITL-----ILLRFARRHLSTSRLLLAGVALGIICSALMTWAIYFSTSVDLRQLMYWMMGGFGGVDWR  189 ECOBTUC
SQFAA-QAGACVVGLI-----VFGVAWGKRLSPVTLILAGLVVSLYCGAINQLLVIF-HHDQLQSMFLWSTGTLTQTDWG  189 ECOFHUB(N)
LLPAG-SLGAAVTLLI-----IMIAAGRGGFSPHRMLLAGMALSTAFTMLLMMLQASGDPRMAQVLTWISGSTYNATDAQ  522 ECOFHUB(C)
VVIAACGGGVSWLLVM-----TAGGGFRHTHDRNKLILAGIALSAFCMGLTRITLLLAEDHASYGIPYWLAGGVSHARWQ  195 ECOFECC
LPLLA-FAGGMAGLIL-----LKMLA--KTHQPMKLALTGVALSACWASLTDYLMLSRPQDVNNALLWLTGSLW-GRDWS  181 ECOFECD
LAMAFAGALVASLIVA-----FTGSQGGGQLSPVRLTLAGVALAAVLEGLTSGIALLNPD-VYDQLRFWQAGSLDIRNLH  194 ECOFEPD
IALSA-MVGGIVTSLL-----VWLLAWRNGIDTFRLIIIGIGVRAMLVAFNTWLLLKASLETALTAGLWNAGSLNGLTWA  191 ECOFEPG
ERMFF-AVLFCFAAGL-----VYIAIIRKVKFSNTAL-VPVIGLMFGSVLSALAEFYAYQNNILQSMSGWLMGDFSKVVQ  175 VANFATD
AVLGV-VGNFVVSAVLILLYSFVIQFWVLKRFQHDMHQVLLIGFVLTMVLTTVAQFIQIRISPGEFSIFQGLSYTSFERA  174 VANFATC


QSW-LMLALIPVLLWICC---QSRPMNMLALGEISARQLGLPLWFWRNVLVAATGWMVGVSVALAGAIGFIGLVIPHIL-  264 ECOBTUC
GVERLWPQLLGGVMLTLL---LLRPLTLMGLDDGVARNLGLALSLARLAALSLAIVISALLVNAVGIIGFIGLFAPLLA-  265 ECOFHUB(N)
VWRTGIVMVIL-LAITPL---CRRWLTILPLGGDTARAVGMALTPTRIALLLLAACLTATATMTIGPLSFVGLMAPHIA-  597 ECOFHUB(C)
DVWQLLPVVVTAVPVVLL---LANQLNLLNLSDSTAHTLGVNLTRLRLVINMLVLLLVGACVSVAGPVAFIGLLVPHLA-  271 ECOFECC
FVKIAIPLMILFLPLSLS---FCRDLDLLALGDARATTLGVSVPHTRFWALLLAVAMTSTGVAACGPISFIGLVVPHMM-  257 ECOFECD
TLKVVLIPVLIAGATALL---LSRALNSLSLGSDTATALGSRVARTQLIGLLAITVLCGSATAIVGPIAFIGLMMPHMA-  270 ECOFEPD
KTSPSAPIIILMLIAAAL---LVRRMRLLEMGDDTACALGVRLERSRLLMMLVAVVLTAAATALAGPISFIALVAPHIA-  267 ECOFEPG
EHYEIIFLILPITLLTYL---YAHRFTVMGMGEDIASNLGISYAMTAALGLILVSITVAVTVVTVGAIHFVGLVIPNLV-  251 VANFATD
KPSTLLFAGTVLSILALFANKWVSELDVIGLGRDQAMSLGLNDAHYIPKYFSVIAILVAISTSLIGPTAFMGVFIANIAY  254 VANFATC
                                  *   *                          *   *

RLCGLTDHRVLLPGCALAGASALLLAD-IVARLALA-AAELPIGVVTATLGAPVFIWLL---LK--A----G-R  326 ECOBTUC
KMLGARRLLPRLMLASLIGALILWLSDQIILWLTRV-WMEVSTGSVIALIGAPLLLWLL---PRL-R----S-I  329 ECOFHUB(N)
RMMGFRRTMPHIVISALVGGLLLVFAD-WCGRMVLF-PFQIPAGLLSTFIGAPYFIYLL---RK--Q-----SR  659 ECOFHUB(C)
RFWAGFDQRNVLPVSMLLGATLMLLAD-VLAR-ALAFPGDLPAGAVLALIGSPCFVWLV---RR--R----G--  332 ECOFECC
RSITGGRHRRLLPVSALTGALLLVVAD-LLARIIHP-PLELPVGVLTAIIGAPWFVWLL---VR--M------R  318 ECOFECD
RWLVGADHRWSLPVTLLATPALLLFAD-IIGR-VIV-PGELRVSVVSAFIGAPVLIFLV---RR--KTRGGA--  334 ECOFEPD
RRISGT-ARWGLTQAALCGALLLLAAD-LCAQQLFM-PYQLPVGVVTVSLGGIYLIVLLIQESR--K------K  330 ECOFEPG
ALKYGDHLKNTLPIVALGGASLLIFCD-VISRVVLF-PFEVPVGLTASAVGGVMFLAFL---LKG-A----K-A  314 VANFATD
SITGSPQYRHTLP-VACTIAIVMFLTA-QLMVEHFF-NYKTTVSILVNVLCGGYFLIIT---MRARS----Q-L  317 VANFATC

number of completely conserved sites: 9
```

Fig. 8. Cluster 8, iron-siderophore, and cobalamine transport systems. Presentation and conventions are the same as in Figure 2.

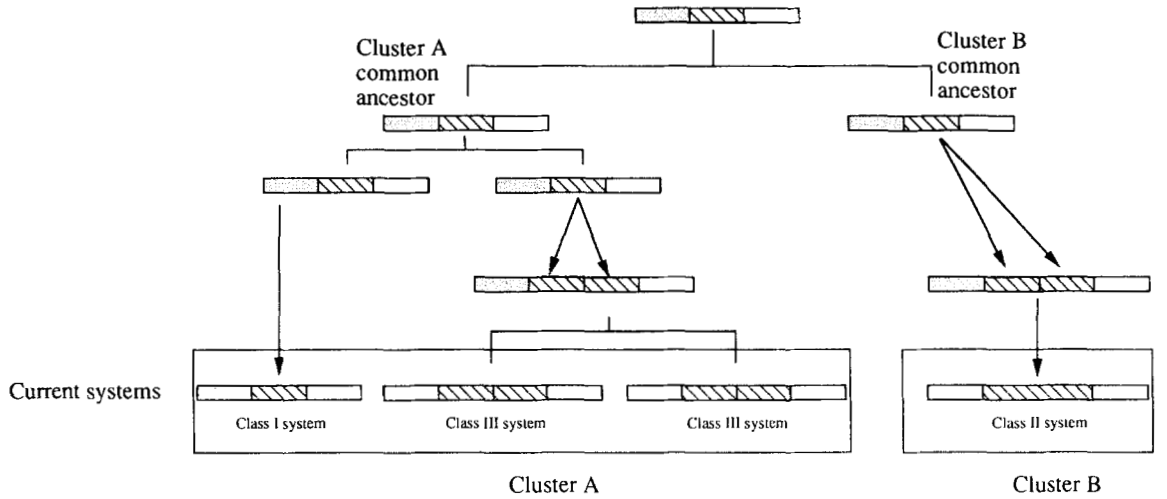


Fig. 9. A scheme for the evolution of binding protein-dependent transport systems. A simplified diagram is shown for 2 clusters. Starting from a putative primordial system (class I), the ancestors of the systems from different clusters are generated by duplicating the whole genetic region. The genes coding for the substrate-binding proteins are symbolized by stippled boxes, those for hydrophobic membrane proteins by hatched boxes, and those for ATP-binding proteins by open boxes. These duplication events are shown as dendrograms. Class III transport systems are generated by a local tandem duplication of the gene encoding the hydrophobic membrane protein. Tandem duplication events are indicated by oblique arrows. Class II systems may derive from class III systems by fusion of the genes encoding the hydrophobic membrane proteins.

Acknowledgments

We thank Maurice Hofnung for his constant support throughout this work and for helpful discussions. We thank Sophie Bachellier, Kalle Gehring, Philip Klebba, and Philippe Marlière for reviewing the manuscript. This work was supported by grants from the Ligue Nationale contre le Cancer and the Fondation pour la Recherche Médicale (FRM) to Maurice Hofnung.

References

Adams MD, Wagner LM, Graddis TJ, Landick R, Antonucci TK, Gibson AL, Oxender DL. 1990. Nucleotide sequence and genetic characterization reveal six essential genes for the LIV-I and LS transport systems of *Escherichia coli*. *J Biol Chem 265*:11436-11443.

Alloing G, Trombe MC, Claverys JP. 1990. The ami locus of the gram-positive bacterium *Streptococcus pneumoniae* is similar to binding protein-dependent transport operons of gram-negative bacteria. *Mol Microbiol 4*:633-644.

Altschul SF, Gish W, Miller W, Myers AW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Amemura M, Makino K, Shinagawa H, Kobayashi A, Nakata A. 1985. Nucleotide sequence of the genes involved in phosphate transport and regulation of the phosphate regulon in *Escherichia coli*. *J Mol Biol 184*:241-250.

Ames GFL. 1988. Structure and mechanism of bacterial periplasmic permeases. *J Bioenerg Biomembr 20*:1-18.

Ames GFL, Mimura C, Shyamala V. 1990. Bacterial periplasmic permeases belongs to a family of transport proteins operating from *Escherichia coli* to human traffic ATPases. *FEMS Microbiol Rev 75*:429-446.

Andersen AB, Ljunqgvist L, Olsen M. 1990. Evidence that protein antigen b of *Mycobacterium tuberculosis* is involved in phosphate metabolism. *J Gen Microbiol 136*:477-480.

Angerer AM, Gaisser S, Braun V. 1990. Nucleotide sequence of the *sfuA*, *sfuB*, and *sfuC* genes of *Serratia marcescens* suggests a periplasmic binding protein-dependent iron transport system. *J Bacteriol 172*:572-578.

Bahl H, Burchhardt G, Wienecke A. 1991. Nucleotide sequence of two *Clostridium thermosulfurogenes* EM1 genes homologous to *Escherichia coli* genes encoding integral membrane components of binding protein-dependent transport systems. *FEMS Microbiol Lett 81*:83-88.

Bell AW, Buckel SD, Groarke JM, Hope JN, Kingsley DH, Hermodson MA. 1986. The nucleotide sequences of the *rbsD rbsA* and *rbsC* genes of *Escherichia coli* K12. *J Biol Chem 261*:7652-7658.

Burchardt G, Bahl F. 1991. Cloning and analysis of the beta-galactosidase-encoding gene from *Clostridium thermosulfuricum* EM1. *Gene 106*:13-19.

Cangelosi GA, Martinetti G, Leigh JA, Lee CC, Theines C, Nester E. 1989. Role of *Agrobacterium tumefaciens* ChvA protein in export of beta-1,2-glucan. *J Bacteriol 171*:1609-1615.

Chen CJ, Chin JE, Ueda K, Clark DP, Pastan I, Gottesman MM, Roninson IB. 1986. Internal duplication and homology with bacterial transport proteins in the *mdr1* (P-glycoprotein) gene from multidrug-resistant human cells. *Cell 47*:381-389.

Chen CM, Ye QZ, Zhu Z, Wanner BL, Walsh CT. 1990. Molecular biology of carbon-phosphorus bond cleavage. *J Biol Chem 265*:4461-4471.

Chenault SS, Earhart CF. 1991. Organization of genes encoding membrane proteins of the *Escherichia coli* ferrienterobactin permease. *Mol Microbiol 5*:1405-1413.

Dahl MK, Francoz E, Saurin W, Boos W, Manson MD, Hofnung M. 1989. Comparison of sequences from the *malB* regions of *Salmonella typhimurium* and *Enterobacter aerogenes* with *Escherichia coli* K12: A potential new regulatory site in the interoperonic region. *Mol Gen Genet 218*:199-207.

Dassa E, Hofnung M. 1985. Sequence of *malG* gene in *E. coli* K12: Homologies between integral membrane components from binding protein-dependent transport systems. *EMBO J 4*:2287-2293.

Devereux J, Haeberli P, Smithies O. 1984. A comprehensive set of sequence analyses for the VAX. *Nucleic Acids Res 12*:387-395.

Deverson EV, Gow IR, Coadwell WJ, Monaco JJ, Butcher GW, Howard JC. 1990. MHC class II region encoding proteins related to the multidrug resistance family of transmembrane transporters. *Nature 348*:738-741.

Diderichsen B, Christiansen L. 1988. Cloning of maltogenic alpha-amylase from *Bacillus stearothermophilus*. *FEMS Microbiol Lett 56*:53-60.

Dreesen TD, Johnson DH, Henikoff S. 1988. The brown protein of *Drosophila melanogaster* is similar to the white protein and to components of active transport systems. *Mol Cell Biol 8*:5206-5215.

Dudler R, Schmidhauser C, Parish RW, Wettenhall REH, Schmidt T. 1988. A *Mycoplasma* high-affinity transport system and the in vitro invasiveness of mouse sarcoma cells. *EMBO J 7*:3963-3970.

Francoz E, Schneider E, Dassa E. 1990. The sequence of the *malG* gene from *Salmonella typhimurium* and its functional implications. *Res Microbiol 141*:633-644.

Friedrich MJ, DeVeaux LC, Kadner RJ. 1986. Nucleotide sequence of the *btuCED* genes involved in vitamin B12 transport in *Escherichia coli* and homology with components of periplasmic binding protein-dependent transport systems. *J Bacteriol 167*:928-934.

Frosch M, Edwards U, Bousset K, Krausse B, Weisgerber C. 1991. Evidence for a common molecular origin of the capsule gene loci in gram-negative bacteria expressing group II capsular polysaccharides. *Mol Microbiol 5*:1251–1263.

Froshauer S, Beckwith J. 1984. Nucleotide sequence of the gene for MalF protein, an inner membrane component of the maltose transport system. *J Biol Chem 259*:10896–10903.

Furuchi T, Kashiwagi K, Kobayashi H, Igarashi K. 1991. Characteristics of the gene for a spermidine and putrescine transport system that maps at 15-min on the *Escherichia coli* chromosome. *J Biol Chem 266*:20928–20933.

Gilson E, Alloing G, Schmidt T, Claverys JP, Dudler R, Hofnung M. 1988. Evidence for high affinity binding protein-dependent transport systems in gram-positive bacteria and in *Mycoplasma*. *EMBO J 7*:3971–3974.

Gilson L, Mahanty HK, Kolter R. 1990. Genetic analysis of an MDR-like export system – The secretion of colicin-V. *EMBO J 9*:3875–3884.

Glaser P, Sakamoto H, Bellalou J, Ullmann A, Danchin A. 1988. Secretion of cyclolysin, the calmodulin-sensitive adenylate cyclase-haemolysin bifunctional protein of *Bordetella pertussis*. *EMBO J 7*:3997–4004.

Gowishankar J. 1989. Nucleotide sequence of the osmoregulatory *proU* operon of *Escherichia coli*. *J Bacteriol 171*:1923–1931.

Gros P, Croop J, Housman D. 1986. Mammalian multidrug resistance gene: Complete cDNA sequence indicates strong homology to bacterial transport proteins. *Cell 47*:371–380.

Guilfoile PG, Hutchinson CR. 1991. A bacterial analog of the mdr gene of mammalian tumor cells is present in *Streptomyces peucetius*, the producer of daunorubicin and doxorubicin. *Proc Natl Acad Sci USA 88*:8553–8557.

Guzzo J, Duong F, Wandersman C, Murgier M, Lazdunski A. 1991. The secretion genes of *Pseudomonas aeruginosa* alkaline protease are functionally related to those of *Erwinia chrysanthemi* proteases and *Escherichia coli* alpha-haemolysin. *Mol Microbiol 5*:447–453.

Hansen JN, Chung YJ. 1992. Determination of the sequence of *spaE* and identification of a promoter in the subtilin (*spa*) operon in *Bacillus subtilis*. *J Bacteriol 174*:6699–6705.

Hazelbauer GL. 1975. The maltose chemoreceptor of *Escherichia coli*. *J Bacteriol 122*:206–214.

Hein J. 1990. Unified approaches to alignment and phylogenies. *Methods Enzymol 188*:626–644.

Higgins CF. 1992. ABC transporters: From microorganisms to man. *Annu Rev Cell Biol 8*:67–113.

Higgins CF, Hagg PD, Nikaido K, Ardeshir F, Garcia G, Ferro-Luzzi Ames G. 1982. Complete nucleotide sequence and identification of membrane components of the histidine transport operon of *Salmonella typhimurium*. *Nature 298*:723–727.

Higgins CF, Hyde SC, Mimmack MM, Gilaedi U, Gill DR, Gallagher MP. 1990. Binding protein-dependent transport systems. *J Bioenerg Biomembr 22*:571–592.

Hiles ID, Gallagher MP, Jamieson DJ, Higgins CF. 1987. Molecular characterization of the oligopeptide permease of *Salmonella typhimurium*. *J Mol Biol 195*:125–142.

Hogg RW, Voelker C, Voncarlowitz I. 1991. Nucleotide sequence and analysis of the *mgl* operon of *Escherichia coli* K12. *Mol Gen Genet 229*:453–459.

Holck AL, Blom H. 1992. The nucleotide sequence of a putative membrane transport gene from *Clostridium perfringens*. *DNA Seq 3*:191–194.

Hoshino T, Kose K. 1990. Cloning, nucleotide sequences and identification of products of the *Pseudomonas aeruginosa* PAO *bra* genes, which encode the high-affinity branched-chain amino acid transport system. *J Bacteriol 172*:5531–5539.

Hryniewicz M, Sirko A, Palucha A, Bock A, Hulanicka D. 1990. Sulfate and thiosulfate transport in *Escherichia coli*-K12: Identification of a gene encoding a novel protein involved in thiosulfate binding. *J Bacteriol 172*:3358–3366.

Johann S, Hinton SM. 1987. Cloning and nucleotide sequence of the *chlD* locus. *J Bacteriol 169*:1911–1916.

Johnson SC. 1967. Hierarchical clustering schemes. *Psychometrika 32*: 241–254.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2268–2268.

Kitamoto N, Yamagata H, Kato T, Tsukagoshi N, Udaka S. 1988. Cloning and sequencing of the gene encoding thermophilic beta-amylase of *Clostridium thermosulfurogenes*. *J Bacteriol 170*:5848–5854.

Köster W, Braun V. 1986. Iron hydroxamate transport of *Escherichia coli*: Nucleotide sequence of the *fhuB* gene and identification of the protein. *Mol Gen Genet 204*:435–442.

Köster W, Braun V. 1990. Iron(III) hydroxamate transport into *Escherichia coli*: Substrate binding to the periplasmic FhuD protein. *J Biol Chem* 21407–21410.

Köster WL, Actis LA, Waldbeser LS, Tolmasky ME, Crosa JH. 1991. Molecular characterization of the iron transport system mediated by the pJM1-plasmid in *Vibrio anguillarum* 775. *J Biol Chem 266*:23829–23833.

Krishnan M, Burgner JW, Chilton WS, Gelvin SB. 1991. Transport of nonmetabolizable opines by *Agrobacter tumefaciens*. *J Bacteriol 173*:903–905.

Laudenbach DE, Grossman AR. 1991. Characterization and mutagenesis of sulfur-regulated genes in a cyanobacterium – Evidence for function in sulfate transport. *J Bacteriol 173*:2739–2750.

Lei SP, Lin HC, Heffernan L, Willcox G. 1985. *AraB* and the nucleotide sequence of the *araC* gene of *Erwinia carotovora*. *J Bacteriol 164*:717–722.

Létoffé S, Delepelaire P, Wandersman C. 1990. Protease secretion by *Erwinia chrysanthemi*: The specific secretion functions are analogous to those of *Escherichia coli* alpha-haemolysin. *EMBO J 9*:1375–1382.

Luque F, Mitchenall LA, Chapman M, Christine R, Pau RN. 1993. Characterization of genes involved in molybdenum transport in *Azotobacter vinelandii*. *Mol Microbiol 7*:447–459.

Makino K, Kim SK, Shinagawa H, Amemura M, Nakata A. 1991. Molecular analysis of the cryptic and functional *phn* operons for phosphonate use in *Escherichia coli* K-12. *J Bacteriol 173*:2665–2672.

Mathiopoulos C, Mueller JP, Slack FJ, Murphy CG, Patankar S, Bukusoglu G, Sonenshein AL. 1991. A *Bacillus subtilis* dipeptide transport system expressed early during sporulation. *Mol Microbiol 5*:1903–1913.

Matsubara K, Ohnishi K, Kiritani K. 1992. Nucleotide sequences and characterization of *liv* genes encoding components of the high-affinity branched-chain amino acid transport system in *Salmonella typhimurium*. *J Biochem (Tokyo) 112*:93–101.

McGrath JP, Varchavsky A. 1989. The yeast STE6 gene encodes a homologue of the mammalian multidrug resistance P-glycoprotein. *Nature 340*:400–404.

Monaco JJ, Cho S, Attaya M. 1990. Transport protein genes in the murine MHC: Possible implications for antigen processing. *Science 250*:1723–1726.

Nohno T, Saito T, Hong JS. 1986. Cloning and complete nucleotide sequence of the *Escherichia coli* glutamine permease operon (*glnHPQ*). *Mol Gen Genet 205*:260–269.

Ohnishi K, Nakazima A, Matsubara K, Kiritani K. 1990. Cloning and nucleotide sequence of *livB* and *livC* the structural genes encoding proteins of the high-affinity branched-chain amino acid transport in *Salmonella typhimurium*. *J Biochem 107*:202–208.

Ohno S. 1974. *Evolution by gene duplication*. New York: Springer-Verlag.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota SI, Inokuchi H, Ozeki H. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature 322*:672–574.

Omata T, Andriesse X, Hirano A. 1993. Identification and characterization of a gene cluster involved in nitrate transport in the cyanobacterium *Synechococcus* sp. PCC7942. *Mol Gen Genet 236*:193–202.

Overduin P, Boos W, Tommassen J. 1988. Nucleotide sequence of the *ugp* genes of *Escherichia coli* K-12: Homology to the maltose system. *Mol Microbiol 2*:767–775.

Perego M, Higgins CF, Pearce SR, Gallagher MP, Hoch JA. 1991. The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. *Mol Microbiol 5*:173–185.

Pistocchi R, Kashiwagi K, Miyamoto S, Nukui E, Sadakata Y, Kobayashi H, Igarashi K. 1993. Characteristics of the operon for a putrescine transport system that maps at 19 minutes on the *Escherichia coli* chromosome. *J Biol Chem 268*:146–152.

Possot O, Denfert C, Reyss I, Pugsley AP. 1992. Pullulanase secretion in *Escherichia coli* K-12 requires a cytoplasmic protein and a putative polytopic cytoplasmic membrane protein. *Mol Microbiol 6*:95–105.

Puyet A, Espinosa M. 1993. Structure of the maltodextrin-uptake locus of *Streptococcus pneumoniae*. Correlation to the *Escherichia coli* maltose regulon. *J Mol Biol 230*:800–811.

Reizer J, Reizer A, Saier MH. 1992. A new subfamily of bacterial ABC-type transport systems catalyzing export of drugs and carbohydrates. *Protein Sci 1*:1326–1332.

Rigby PWJ, Burleigh BDJ, Hartley BS. 1974. Gene duplication in experimental enzyme evolution. *Nature 251*:200–204.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielinski J, Lok S, Plavsic N, Chou JL, Drumm ML, Iannuzzi MC, Collins FS, Tsui LC. 1989. Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science 245*:1066–1072.

Rioux C, Kadner RJ. 1989. Vitamin B12 transport in *Escherichia coli* K12 does not require the *btuE* gene of the *btuCED* operon. *Mol Gen Genet 217*:301–308.

Rudner DZ, Ledeaux JR, Ireton K, Grossman AD. 1991. The *spoOK* locus of *Bacillus subtilis* is homologous to the oligopeptide permease locus and is required for sporulation and competence. *J Bacteriol 173*:1388–1398.

Russell RRB, Aduseopoku J, Sutcliffe IC, Tao L, Ferretti JJ. 1992. A binding protein-dependent transport system in *Streptococcus mutans* responsible for multiple sugar metabolism. *J Biol Chem 267*:4631–4637.

Schneider E, Francoz E, Dassa E. 1992. Completion of the nucleotide sequence of the maltose B region in *Salmonella typhimurium* – The high conservation of the *malM* gene suggests a selected physiological role for its product. *Biochim Biophys Acta 1129*:223–227.

Scripture JB, Voelker C, Miller S, O'Donnell RT, Polgar L, Rade J, Horazdovsky BF, Hogg RW. 1987. High-affinity L-arabinose transport operon. Nucleotide sequence and analysis of gene products. *J Mol Biol 197*:37–46.

Shea CM, Mcintosh MA. 1991. Nucleotide sequence and genetic organization of the ferric enterobactin transport system – Homology to other periplasmic binding protein-dependent systems in *Escherichia coli*. *Mol Microbiol 5*:1415–1428.

Sirko A, Hryniewicz M, Hulanicka D, Bock A. 1990. Sulfate and thiosulfate transport in *Escherichia coli* K-12: Nucleotide sequence and expression of the *cys TWAM* gene cluster. *J Bacteriol 172*:3351–3357.

Skare JT, Postle K. 1991. Evidence for a TonB-dependent energy transduction complex in *Escherichia coli*. *Mol Microbiol* 2883–2890.

Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull 28*:1409–1438.

Staudenmaier H, Van Hove B, Yaraghi Z, Braun V. 1989. Nucleotide sequences of the *fecBCDE* genes and locations of the proteins suggest a periplasmic binding protein-dependent transport mechanism for iron(III) dicitrate in *Escherichia coli*. *J Bacteriol 171*:2626–2633.

Surin B, Rosenberg H, Cox GB. 1985. Phosphate-specific transport system of *Escherichia coli*: Nucleotide sequence and gene–polypeptide relationships. *J Bacteriol 161*:189–198.

Tam R, Saier MH. 1993. Structural, functional and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol Rev 57*:320–346.

Treptow NA, Shuman HA. 1985. Genetic evidence for substrate and binding protein recognition by the MalF and MalG proteins, cytoplasmic membrane components of the *Escherichia coli* maltose transport system. *J Bacteriol 163*:654–660.

Trombe MC, Laneelle G, Sicard AM. 1984. Characterization of *Streptococcus pneumoniae* mutant with altered electric transmembrane potential. *J Bacteriol 158*:1109–1114.

Valdivia RH, Lu W, Winans SC. 1991. Characterization of a putative periplasmic transport system for octopine accumulation encoded by *Agrobacterium tumefaciens* Ti plasmid pTiA6. *J Bacteriol 173*:6398–6405.

Walter C, Bentrup KHZ, Schneider E. 1992. Large scale purification, nucleotide binding properties, and ATPase activity of the MalK subunit of *Salmonella typhimurium* maltose transport complex. *J Biol Chem 267*: 8863–8869.

Wang SZ, Chen JS, Johnson JL. 1990. A nitrogen fixation gene (*nifC*) in *Clostridium pasteurianum* with sequence similarity to *chlJ* of *E. coli*. *Biochem Biophys Res Commun 169*:1122–1128.

Williams SG, Greenwood JA, Jones CW. 1992. Molecular analysis of the *lac* operon encoding the binding protein-dependent lactose transport system and beta-galactosidase in *Agrobacterium radiobacter*. *Mol Microbiol 6*:1755–1768.

Wissenbach U, Keck B, Unden G. 1993. Physical map location of the new *artPIQMJ* genes of *Escherichia coli*, encoding a periplasmic arginine transport system. *J Bacteriol 175*:3687–3688.

Zanker H, Vonlintig J, Schroder J. 1992. Opine transport genes in the octopine (*occ*) and nopaline (*noc*) catabolic regions in Ti plasmids of *Agrobacterium tumefaciens*. *J Bacteriol 174*:841–849.