University of Groningen

Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis

Lolkema, JS; Slotboom, DJ

Published in:
Journal of Molecular Biology

DOI:
[10.1016/S0022-2836\(03\)00214-6](https://doi.org/10.1016/S0022-2836(03)00214-6)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2003

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Lolkema, JS., & Slotboom, DJ. (2003). Classification of 29 families of secondary transport proteins into a single structural class using hydropathy profile analysis. *Journal of Molecular Biology*, 327(5), 901-909. [https://doi.org/10.1016/S0022-2836\(03\)00214-6](https://doi.org/10.1016/S0022-2836(03)00214-6)

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

COMMUNICATION

Classification of 29 Families of Secondary Transport Proteins into a Single Structural Class using Hydropathy Profile Analysis

Juke S. Lolkema* and Dirk-Jan Slotboom

*Molecular Microbiology
Biomolecular Sciences and
Biotechnology Institute
University of Groningen
Kerklaan 30, 9751NN Haren
The Netherlands*

A classification scheme for membrane proteins is proposed that clusters families of proteins into structural classes based on hydropathy profile analysis. The averaged hydropathy profiles of protein families are taken as fingerprints of the 3D structure of the proteins and, therefore, are able to detect more distant evolutionary relationships than amino acid sequences. A procedure was developed in which hydropathy profile analysis is used initially as a filter in a BLAST search of the NCBI protein database. The strength of the procedure is demonstrated by the classification of 29 families of secondary transporters into a single structural class, termed ST[3]. An exhaustive search of the database revealed that the 29 families contain 568 unique sequences. The proteins are predominantly from prokaryotic origin and most of the characterized transporters in ST[3] transport organic and inorganic anions and a smaller number are Na^+/H^+ antiporters. All modes of energy coupling (symport, antiport, uniport) are found in structural class ST[3]. The relevance of the classification for structure/function prediction of uncharacterised transporters in the class is discussed.

© 2003 Elsevier Science Ltd. All rights reserved

*Corresponding author

Keywords: membrane protein; classification; structural class; secondary transporter; hydropathy profile

Secondary transporters are a diverse class of integral membrane proteins that are found in all kingdoms of life. They translocate solutes across biological membranes driven by *trans* membrane gradients and are involved in important cellular functions; besides the uptake of nutrients, secondary transporters operate in processes like the excretion of metabolic end products, metabolic energy generation, pH homeostasis, osmoregulation, defence mechanisms, neurotransmission and brain function.

Estimates based on the sequence of a number of bacterial genomes revealed that up to 7% of the proteome may be secondary transporters.¹ Although the number of genes coding for secondary transporters varies considerably between different organisms it is estimated that 20,000–30,000 of the roughly 750,000 entries that are currently (April 2002) in the protein database at the National Center for Biotechnology Information (NCBI)[†] are secondary transporters. Sequence analysis has revealed the existence of a few large superfamilies of secondary transporters, e.g. the major facilitator superfamily² (MFS) and the amino acid/polyamine/organocation superfamily³ (APC), and many more smaller families. The transport protein classification (TC) system developed by Saier and co-workers⁴ now lists 78 different families in the uniporters, symporters and antiporters category‡. Possibly, all these transporter proteins have a single common ancestor that has

Present address: D. -J. Slotboom, The MRC-Dunn Human Nutrition Unit, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, UK.

Abbreviations used: MFS, major facilitator superfamily; APC, amino acid/polyamine/organocation superfamily; TC, transport protein classification; ST[3], secondary transporters (class 3); IT, ion transporters; SDS, structure divergence score; PDS, profile difference score; TMS, *trans* membrane segment.

E-mail address of the corresponding author: j.s.lolkema@biol.rug.nl

[†] <http://www.ncbi.nlm.nih.gov/entrez/>

[‡] <http://www-biology.ucsd.edu/~msaier/transport/>

evolved to the large diversity of proteins known today. However, a common evolutionary pathway of secondary transporters from different (super)-families cannot be detected anymore from a comparison of the amino acid sequences. The sequence identity between members of different families is in the same order as expected for random sequences of the same composition. In general, evolutionary relationships between proteins may be detected at three levels: (i) a close relationship follows when a significant sequence identity is observed after optimal alignment of the two amino acid sequences; (ii) a more distant relationship is detected by local alignment procedures, like the BLAST algorithm,⁵ that search for significant identities between parts of the sequences; and (iii) even more distant relationships are revealed when the three-dimensional structures of the proteins are compared. The latter is a manifestation of the fact that structure is much better conserved during evolution than amino acid sequence. Detection of similarity at the structural level requires the availability of a large database of protein structures^{6–8} which is not available for membrane proteins. Despite considerable progress in the last few years the number of available structures of membrane proteins is still very low compared to soluble proteins. For secondary transporters, only a few low resolution projection structures and 3D reconstructions from electron microscopy are available.^{9–14} Previously, we have explored the possibility of using hydropathy profiles¹⁵ of the amino acid sequences as a substitution for 3D structures to demonstrate distant evolutionary relationships between membrane proteins.^{16,17} The use of hydropathy profiles to predict secondary structure of membrane proteins is widely spread.^{18,19} To use the profiles to compare the 3D fold of the proteins requires that they contain information on tertiary structure as well. Using family-averaged hydropathy profiles and statistical approaches, a number of families of secondary transporters were grouped in four different structural classes.¹⁶ The families in a structural class have the same fold and, therefore, are related. Here, we report on an exhaustive search of the NCBI protein database to identify all proteins that belong to one of the structural classes, termed ST[3] (secondary transporters (class 3)). ST[3] is part of a larger membrane protein classification scheme. The class is supplied as an electronic appendix (see Supplementary Material) to this paper and is stored in the MemGen relational database of membrane proteins that will be updated regularly†.

The MemGen classification

Structural class ST[3] contains 568 unique sequences classified in 29 families and 48 sub-

families (Table 1), many of which had not been shown to be related before. The 29 families include 15 families present in the TC system,⁴ 11 of which were assigned to the putative superfamily of ion transporters²⁰ (IT superfamily). Fourteen families in ST[3] are not present in the TC system. Subfamilies in the MemGen classification represent proteins that produce significant pair-wise sequence identities in a multiple sequence alignment (CLUSTAL.^{21,22}) Apart from grouping sequences that are clearly related throughout the whole sequence, the classification in subfamilies serves the purpose of producing the averaged hydropathy profiles that rely heavily on a reliable multiple sequence alignment. Pair-wise sequence identities within the subfamilies range between 19 and 63%. The structure divergence score (SDS) is a measure of the divergence of the individual hydropathy profiles of the members in a subfamily relative to the averaged hydropathy profile. As observed before, SDS values are typically between 0.10 and 0.12 hydrophobicity units.¹⁶

Subfamilies are grouped in families when local sequence alignment tools indicate that the members in the different subfamilies are related. The assignment is based on the Expect value produced by a BLAST search⁵ of the NCBI protein database‡. The family level represents the most distant relationship between proteins that can be detected based up on the amino acid sequences. The MemGen family level corresponds more or less to the family level in the TC classification system⁴ (Table 1). The structural relationship between the subfamilies of most families is evident from the alignment of the averaged hydropathy profiles that result in values for the *S*-test below 1 (Table 2, upper off-diagonal elements). The *S*-test compares the difference between two averaged profiles after optimal alignment (PDS; profile difference score) with the divergence of the individual profiles within each of the two averaged profiles (SDS).¹⁶ An *S*-test value of 1 indicates that the “distance” between two averaged profiles is the same as the divergence within the two averaged profiles and an *S*-test value of ≥ 1 indicates that the two averaged profiles are dissimilar. The low pair-wise *S*-test values for the subfamilies show that the profiles are very similar, even though sequence identities between members of the different subfamilies are well below 25%.

The averaged profile alignment of the [st307]DCTM subfamilies reveals an inherent weakness of the alignment algorithm when hydropathy profiles are complex, containing many hydrophobic segments and extensions/insertions that are not present in both profiles. Then, the calculated optimal alignment may not represent a physical reality and the alignment must be done completely by hand. Nevertheless the result of this exercise was an *S*-test value of 0.60 (Table 2(D)). The

† <http://www.biol.rug.nl/microbiologie/MemGen/main.htm>

‡ <http://www.ncbi.nlm.nih.gov/blast/>

Table 1. Overview of the transporter families in structural class ST[3]

Family ^a	TC system	Sub-families ^a	Unique sequences	Kingdom	Selected members	Known substrates	References ^b
[st301]MeCit	2.A.11 CitMHS	1	12	B	CITM, CITH	Me ²⁺ -citrate	39,40
[st302]ArsB	2.A.45 ArsB	4	64	B, A, E	ARSB	Arsenite	23
[st303]AIT	2.A.47 DASS	7	91	B, A, E	SOT1, CITT, NADC, SUT1, NDC1, NASU	Malate, fumarate, succinate, 2-oxo glutarate, citrate, tartrate, sulphate	24,41,42
[st304]GNT	2.A.8 GntP	3	61	B, A	GNTP, IDNT	Gluconate, L-idonate	43
[st305]DCUA	2.A.13 Dcu	1	24	B	DCUA, DCUB	C4-dicarboxylates	44,45
[st306]NHAB	2.A.34 NhaB	1	10	B	NHAB	Na ⁺ /H ⁺	29,46
[st307]DCTM	2.A.56 TRAP-T	3	80	B, A	DCTM	C4-dicarboxylates	20,47
[st308]DCTB		2	6	B			48
[st309]DCUC	2.A.61 DcuC	1	11	B	DCUC	C4-dicarboxylates	44
[st310]ATO	2.A.1.37 AtoE	1	11	B, A			
[st311]AITB		1	8	B			
[st312]NHAC	2.A.35 NhaC	5	48	B, A	NHAC, MLEN	Na ⁺ /H ⁺ , malate, lactate	30,49
[st313]AITC		1	19	B			
[st314]AITD	2.A.68 AbgT	1	16	B	YDAH	Aminobenzoyl-glutamate	50
[st315]AITE		1	16	B, A			
[st316]NHAD	2.A.62 NhaD	1	7	B, E	NHAD	Na ⁺ /H ⁺	51
[st317]AITF		1	5	B, A			
[st318]AITG		1	2	B			
[st319]AITH		1	3	A			
[st320]AITI		1	2	B			
[st321]AITJ		1	3	B			
[st322]AITK		1	1	B			
[st323]AITL		1	1	B			
[st324]GltS	2.A.27 ESS	1	18	B	GLTS	Glutamate	52,53
[st325]2HMCT	2.A.14 LctP	2	29	B, A	LLDP	Lactate, glycolate	54,55
[st326]2HCT	2.A.24 CCS	1	17	B	CITS, CITP, MLEP, CIMH, MALP, MAEN	Citrate, malate, lactate	25,30,56,57
[st327]AITM		1	1	B			
[st328]AITN		1	1	B			
[st329]AITO		1	1	A			
29		48	568			Anions and Na ⁺ /H ⁺	

Building the class. Step 1: a random sequence from one of the families was used as the query in a BLAST search of the NCBI protein database. All hits up to a *pE* value of 10 (*pE* is defined as the absolute value of the exponent of the Expect value) were stored in the local MemGen database. The BLAST search was performed without setting any filters and “composition based statistics” was not applied. The imported sequences were sorted in the categories trash (truncates and reading frame errors, etc), database duplicates (>98% identical and from the same organism), similars (>60% identical to a typical) and typicals (represents a group of linked proteins that consists of one typical and a variable number of similar groups and database duplicates). The typical groups were aligned with the CLUSTAL X multiple sequence alignment program²¹ and the first subfamily was created by selecting the group of proteins, including the query sequence, with pair-wise sequence identities of at least 20–25%. The procedure was then repeated with a BLAST search of the typical sequence with the highest *pE* value that was not included in the subfamily. If the BLAST search produced new hits with *pE* values higher than 10, these were imported, sorted and aligned with the remaining typical groups from the previous BLAST search. Subsequently, the next subfamily was defined containing the new query. The procedure for constructing subfamilies was repeated until no new hits with *pE* > 10 were found and all proteins in the multiple sequence alignments had been assigned to a subfamily. Step 2: the BLAST results of all the typical sequences found in step 1 were collected and stored in a local database. All hits with *pE* values of greater than 2 that were not yet in the MemGen database were imported and a decision was made whether or not they belonged to the class for which a number of techniques were used: the hits with high *pE* values were aligned with all the typicals in the subfamilies, which usually resulted in the assignment of the hit to one of the subfamilies as duplicate, similar or typical. For the hits with low *pE* values the sequence identities obtained in the alignment procedure was usually below 30%. Then the decision to assign the protein to the class was made based on the results of a BLAST search of the hit using different settings of the parameters (“low complexity region filter”, “composition based statistics”, etc.), visual inspection of the hydropathy profile, and/or hydropathy profile alignment with all the proteins in the subfamilies.¹⁶ With low *pE* values, a hit assigned to the class usually resulted in a new subfamily. The procedure was repeated to account for hits produced by the newly assigned typical groups until all hits with a *pE* of 2 or higher were evaluated.

^a For the procedure to group subfamilies into families see [Table 2](#).

^b Whenever possible the reader is referred to a recent review.

Table 2. Sequence and structural homology of the subfamilies in ST3 families

A.	[st302]ArsB1	[st302]ArsB2	[st302]ArsB3	[st302]ArsB4		
[st302]ArsB1	–	0.76	0.95	0.77		
[st302]ArsB2	57	–	0.99	0.56		
[st302]ArsB3	16	65	–	0.57		
[st302]ArsB4	59	91	85	–		
B.	[st303]AIT1	[st303]AIT2	[st303]AIT3	[st303]AIT4	[st303]AIT5	[st303]AIT6
[st303]AIT1	–	0.93	0.64			
[st303]AIT2	48	–	0.50			
[st303]AIT3	6	52	–			
[st303]AIT4	56	22	13	–		
[st303]AIT5	62	48	26	100	–	
[st303]AIT6	0	0	100	0	0	–
[st303]AIT7	0	55	0	0	0	0
C.	[st304]GNT1	[st304]GNT2	[st304]GNT3	[st308]DCTB1	[st308]DCTB2	
[st304]GNT1	–	1.35				
[st304]GNT2	4	–				
[st304]GNT3	80	100	–			
[st308]DCTB1	–			–		
[st308]DCTB2				100	–	
D.	[st307] DCTM1	[st307] DCTM2	[st307] DCTM3	[st325] 2HMCT1	[st325] 2HMCT2	
[st307]DCTM1	–	0.60				
[st307]DCTM2	82	–				
[st307]DCTM3	91	76	–			
[st325]2HMCT1				–	0.79	
[st325]2HMCT2				100	–	
E.	[st307] NHAC1	[st307] NHAC2	[st307] NHAC3	[st325] NHAC4	[st325] NHAC5	
[st312]NHAC1	–	2.71	0.87			
[st312]NHAC2	21	–	1.65			
[st312]NHAC3	94	50	–			
[st312]NHAC4	63	46	100	–		
[st312]NHAC5	83	46	100	100	–	

In the MemGen classification the following operational definition of a family is used. Two subfamilies belong to the same family when, in a BLAST search, the typicals in one subfamily “hit” 50% of the typicals in the other subfamily with a *pE* value of 8 or more. A second way by which two subfamilies are grouped in one family is when the 50% criterion given above is made *via* a third subfamily. So, when subfamily A produces 50% hits with a *pE* value of 8 or more with subfamilies B and C, but B and C produce less than 50% hits, B and C are still in the same family.^{58,59} The lower off-diagonal elements indicate the number of observed links between the subfamilies as a percentage of the total number of possible links using a *pE* threshold of 8. The upper off-diagonal elements indicate the *S*-test value obtained from the optimal alignment of the averaged hydropathy profiles of the subfamilies. No value is given when a subfamily contains too few proteins to produce a reliable averaged hydropathy profile. For a definition of the *S*-test see Lolkema & Slotboom.¹⁶

resulting alignment extends over 1000 positions and contains over 20 hydrophobic regions. The complex structure of the proteins will be discussed elsewhere.

The alignment of the family hydropathy profiles of the [st312]NHAC2 subfamily with the [st312]NHAC1 and [st312]NHAC3 subfamilies results in *S*-test values larger than 1, indicative of significant differences between the profiles (Table 2(E)). Therefore, the overall folding of the proteins is unlikely to be the same even though seemingly significant sequence homology exists between parts of the proteins in subfamily [st312]NHAC2 and the other [st312]NHAC subfamilies. Hydropathy profile alignments provide a more stringent criterion for classification than the BLAST expect value, because it is a “full sequence” property, i.e. the profiles of the query and the true positive picked up in this way are similar over the complete sequence as should be the case for proteins with a similar fold. In contrast, the local alignments of

the sequences in [st312]NHAC1 and members in [st312]NHAC2 provided by the BLAST algorithm amount to only half of the length of the proteins and, importantly, align different parts of the proteins. Sequence homology is observed between the N-terminal half of [st312]NHAC1 and the C-terminal half of [st312]NHAC2. This relation is consistently observed between all proteins in [st312]NHAC2 and the proteins in the other four subfamilies.

Structural classes group families of membrane proteins with similar averaged hydropathy profiles. The criterion for structural similarity is based on a statistical analysis of the difference between the averaged profiles after optimal alignment and the divergence observed in the individual profiles of the members of the families.¹⁶ Fourteen of the 29 families in ST[3] contain at least one subfamily with six or more sequences which allows the construction of an averaged hydropathy profile and the calculation of a meaningful SDS value based

Table 3. Averaged hydropathy profile alignments

pE ^a	Subfamily	Subfamily	Observed links ^b (%)	S-value	PDS ^c
8	[st302]ArsB2	[st303]AIT3	29	0.73	0.115
6	[st304]GNT1	[st311]AITB1	15	1.05	0.116
	[st313]AITC1	[st314]AITD1	12	1.09	0.130
5	[st304]GNT2	[st312]NHAC2	17	0.87	0.101
4	[st301]MeCit1	[st302]ARSB2	14	0.81	0.109
	[st304]GNT2	[st312]NHAC3	23	1.59	0.136
	[st310]ATO1	[st313]AITC1	15	1.06	0.119
	[st310]ATO1	[st314]AITD1	15	1.07	0.109
3	[st304]GNT2	[st312]NHAC1	14	2.21	0.155
	[st304]GNT2	[st313]AITC1	14	1.06	0.113
2	[st301]MeCit1	[st313]AITC1	13	1.00	0.116
	[st303]AIT3	[st307]DCTM1	11	0.71	0.108
	[st304]GNT2	[st307]DCTM1	11	1.14	0.116
1	[st301]MeCit1	[st304]GNT1	14	0.82	0.102
	[st301]MeCit1	[st307]DCTM1	15	1.16	0.124
	[st302]ArsB2	[st307]DCTM1	15	0.99	0.124
	[st302]ArsB2	[st313]AITC1	11	0.70	0.105
	[st303]AIT3	[st304]GNT2	12	0.81	0.109
	[st304]GNT2	[st314]AITD1	16	0.81	0.109
	[st307]DCTM1	[st313]AITC1	14	0.99	0.118
0	[st301]MeCit1	[st303]AIT3	17	0.62	0.100
	[st301]MeCit1	[st305]DCUA1	24	1.25	0.116
	[st303]AIT3	[st305]DCUA1	17	0.98	0.118
	[st304]GNT1	[st325]2HMCT1	11	0.92	0.107
	[st324]GLTS	[st326]2HCT1	3	0.96	0.116
	[st326]2HCT	[st301]MeCit1	5	1.07	0.111

The averaged hydropathy profiles of subfamilies in different families were aligned. Subfamilies with at least six sequences were included in the analysis.

^a Highest pE value observed between proteins in the two subfamilies.

^b Percentage of links with the highest pE value between proteins in the two subfamilies observed by the BLAST algorithm.

^c Profile difference score.¹⁶

on the multiple sequence alignment. The averaged hydropathy profiles of subfamilies from different families were aligned to confirm the structural relationship between the families. Table 3 lists the results of the alignments of pairs of subfamilies. With few exceptions (see below), the S-values of the profile alignments were close to 1, justifying their classification in one structural class. The pairs of subfamilies given in Table 1 relate each of the families, either directly or indirectly, to all other families in the group of 14.

Within the [st312]NHAC family, [st312]NHAC2 and the other [st312]NHAC subfamilies seemed to form different clusters (see above). The averaged profiles of subfamilies [st312]NHAC1, [st312]NHAC2, and [st312]NHAC3 were aligned with the profile of [st304]GNT2 yielding S-values of 2.21, 0.87, and 1.59, respectively (Table 3). Since [st304]GNT2 aligns with other subfamilies with S-values around 1, it follows that, most likely, [st312]NHAC2 has the global folding common to the proteins in ST[3], while the other [st312]NHAC subfamilies have not.

Fifteen of the 29 families in ST[3] did not contain subfamilies with a sufficient number of proteins to construct a reliable averaged profile. From each of the subfamilies in these 15 families, a protein was selected, of which the hydropathy profile was aligned with the averaged hydropathy profiles of the subfamilies for which an SDS value was available. The difference between the profiles (PDS)

was compared to the SDS value of the averaged profile. The PDS values were similar to the SDS values, indicating that the difference between the individual profile and the family profile was similar to the average difference of the individual profiles within the family profile (results not shown). In other words, the query profile could well have been a member of the family profile. It should be noted that this criterion for structural similarity is less stringent than the comparison of family profiles, which allows for a statistical test in the form of the S-value.

The transporters of structural class ST[3]

A remarkable fraction of the proteins in class ST[3] is of bacterial origin. They are found in all but two families, [st319]AITH and [st329]AITO, which contain only proteins of archeal origin. Seventeen of the 29 families contain proteins exclusively from the bacterial kingdom, while 26 families are represented by bacterial and/or archeal proteins only. Since relatively few archeal genomes have been sequenced to date, it may be anticipated that ST[3] contains typically prokaryotic proteins. Proteins from eukaryotic origin are found in three families, [st302]ArsB, [st303]AIT and [st316]NHAD, subdivided over seven subfamilies. They amount to less than 10% of all the proteins in ST[3] (42/568). The highest fraction of eukaryotic proteins is found in subfamily

[st303]AIT2 where they constitute roughly half of the proteins. Remarkably, proteins from eukaryotic microorganisms are poorly represented in ST[3] with only a single protein from *Dictyostelium discoideum* in [st304]ArsB4.

As expected, few proteins in class ST[3] have been functionally characterised. Best studied examples are ARSB, the subunit of the arsenical resistance pump of *Escherichia coli*²³ ([st302]ArsB1), the renal and intestinal dicarboxylate transporters NaDC²⁴ ([st303]AIT2), and the citrate and malate transporters CITS of *Klebsiella pneumoniae*, CITP of *Leuconostoc mesenteroides* and MLEP of *Lactobacillus lactis* in the 2-hydroxycarboxylate transporter family ([st326]HCT1).^{25–28} All characterized proteins represent secondary transporters even though the ArsB proteins in [st302]ArsB are a special case.²³ Many transporters in different families of ST[3] transport organic anions, like the C₄-carboxylates fumarate and succinate, the hydroxycarboxylates glycolate, lactate, malate, tartrate and citrate, the oxocarboxylate α -ketoglutarate, the amino acids glutamate and aminobenzoyl-glutamate, and the sugar acids gluconate and idonate. Two families contain transporters for inorganic anions, arsenite in [st302]ArsB and sulphate in [st303]AIT. A distinct group of transporters are found in the three families [st306]NHAB, [st312]NHAC and [st316]NHAD. Members of these families are Na⁺/H⁺ antiporters.²⁹ Summarizing, the substrates of the transporters in class ST[3] seem to be of either of two types, anions and Na/H ions. In this respect, the activity of the protein YQKI of *Bacillus subtilis* in subfamily [st312]NHAC1 is interesting as it is believed to combine the activities of Na⁺/H⁺ antiport and malate/lactate antiport.³⁰

Secondary transporters are driven by *trans* membrane gradients of solutes and ions, and they couple fluxes in a number of different ways. All these different modes of energy coupling seem to be represented by the transporters in class ST[3]. Na⁺/H⁺ antiporters obligatorily translocate Na ions and protons in opposite directions. H⁺/solute symporters couple proton and solute flux in the same direction and CITM of *B. subtilis* in [st301]MeCit and GNT of *B. subtilis* in [st304]GNT are examples of this mode of transport. Examples of Na⁺/solute symporters are NDC2 of rat in [st303]AIT2, GLTS of *E. coli* in [st324]GLTS, and CITS of *K. pneumoniae* in [st326]2HCT. Transporters that translocate solutes in opposite directions (exchangers) are represented by CITP of *E. coli* in [st302]AIT1, and CITP of *L. mesenteroides* and MLEP of *L. lactis* in [st326]2HCT1. Finally, the *E. coli* ARSB subunit in [st302]ArsB1 may be an example of a uniporter. Clearly, mode of energy coupling is not class or even family-specific.

Comparative hydropathy profile analysis

Predicting the membrane topology by computational methods is usually the first step in the

study of a membrane protein, and the computational methods are becoming more and more important now that the number of sequences representing membrane proteins in the databases increase rapidly as the result of genome sequencing projects. The classification procedure we propose here, in principle, brings together membrane proteins with the same structure, without knowing the structure. Nevertheless, the classification may make strong predictions about certain aspects of families in a structural class by a comparative analysis as will be demonstrated for the [st324]GLTS and [st326]2HCT families.

Both the [st324]GLTS and [st326]2HCT families are very weakly linked to the other families in ST[3] based on sequence alone. Between members of the two families, only a few links with a *pE* value of 0 were observed (Table 3). Nevertheless, the hydropathy profiles of the two families are very similar (*S*-value of 0.96; Figure 1(A)) confirming their classification in one and the same structural class. While no structural information is available on members of the [st324]GLTS family, the structure of several members of the [st326]2HCT family has been studied using mutagenesis and gene fusion techniques. Membrane topology studies of the Na⁺-dependent citrate transporter CitS of *K. pneumoniae* have revealed the presence of 11 transmembrane segments (TMSs; Figure 1(B)). The N terminus is located in the cytoplasm and the C terminus in the periplasm. A twelfth hydrophobic segment around position 210 (Figure 1), predicted by most secondary structure prediction programs to be transmembrane, is located in the periplasm between TMSs V and VI and termed VB.^{28,31–33} TMS XI and the cytoplasmic loop preceding this segment are part of the substrate binding site.^{27,34,35} A conserved arginine residue at the interface of TMS XI and the loop was shown to directly interact with the substrate. The family hydropathy profile alignment of the GLTS and 2HCT families reveals that (i) the [st324]GLTS family misses the first transmembrane segment present in the [st326]2HCT family (positions 1–60); (ii) both families have a large hydrophilic loop in the middle of the sequence (positions 260–290); (iii) a moderately hydrophobic region in the [st324]GLTS family is missing in the [st326]2HCT family (positions 420–450); and (iv) the C termini of both families are hydrophobic. One additional remarkable feature of the transporters in the [st326]2HCT family is the presence of a putative amphipathic α -helix in the loop between TMSs VIII and IX around position 350. The structural feature has been implicated in the co-insertion of TMSs VIII and IX during the biogenesis of the Na⁺-citrate symporter of *K. pneumoniae*.^{32,36} Similar putative amphipathic helices are observed in the corresponding loops of the members of the [st324]GLTS family.

The comparative analysis of the [st326]2HCT and [st324]GLTS families results in the following predictions for the [st324]GLTS family. (i) The N

A.

B.

Figure 1. Structural similarity between st[324]GLTS and [st326]2HCT families. (A) Averaged hydropathy profile alignment of st[324]GLTS (blue) and [st326]2HCT (red). The S-test value equals 0.96. The red and blue bars in the top part of the Figure indicate the positions of gaps introduced in the two profiles, respectively. The red bars in the lower part of the Figure indicate the positions of the TMSs in the Na⁺/citrate symporter CITS of *Klebsiella pneumoniae* in the 2HCT family. (B) Membrane topology model of the members of the 2HCT family based on studies of CITS. The position of Arg425 in the citrate/lactate exchanger CITP of *Leuconostoc mesenteroides* that is believed to interact with one of the carboxylate groups of the substrates citrate and malate is indicated. The green arrows point at (a) the hydrophobic segment VB in CITS that was shown to be periplasmic, (b) the loop that forms a putative amphipathic helix and is believed to play a role in the insertion of TMS VIII and IX, and (c) the region in [st324]GLTS that contains a hydrophobic peak not present in [st326]2HCT.

terminus is located in the periplasm. (ii) The hydrophobic region around position 200 is not transmembrane. The region corresponds to region VB in [st326]2HCT, which is located at the periplasmic side of the membrane. Remarkably, prediction of the membrane topology of GltS of *E. coli* by the TMHMM program³⁷ agrees with the periplasmic location of the segment. (iii) The moderately hydrophobic peak in the profile around position 440 does not represent a transmembrane segment but is cytoplasmic. This region is likely to contain the glutamate binding site. Recent experiments on the citrate/malate transporter CimH of *B. subtilis* in the [st326]2HCT family suggest that the conserved region between TMS X and XI may fold back between the helices, forming a re-entrant loop.³⁵ Such a structure would explain the hydrophobic character of the loop in the hydrophobicity plot.³⁸ (iv) The N and C termini are located in the periplasm.

Most importantly, the analysis shows that the structural classification presented here can be veri-

fied experimentally by relatively simple experiments. If the structural and functional features of the members of the [st326]2HCT and [st324]GLTS families that are very distant in ST[3], would correctly coincide as predicted above, this would strongly support the classification procedure and demonstrate the predicting power of hydropathy profile alignments. Currently, we are expanding the analysis to the other families in ST[3] and are verifying the predictions made for the [st324]GLTS family by studying the structural and functional features of GltS protein of *E. coli* experimentally.

References

1. Paulsen, I. T., Nguyen, L., Sliwinski, M. K., Rabus, R. & Saier, M. H., Jr (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.* **301**, 75–100.
2. Saier, M. H., Jr, Beatty, J. T., Goffeau, A., Harley, K. T., Heijne, W. H., Huang, S. C. *et al.* (1999). The major

- facilitator superfamily. *J. Mol. Microbiol. Biotechnol.* **1**, 257–279.
3. Jack, D. L., Paulsen, I. T. & Saier, M. H. (2000). The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology*, **146**, 1797–1814.
 4. Saier, M. H., Jr (2000). A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.* **64**, 354–411.
 5. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
 6. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modelling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* **301**, 665–678.
 7. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301**, 679–689.
 8. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.* **301**, 691–711.
 9. Heymann, J. A., Sarker, R., Hirai, T., Shi, D., Milne, J. L., Maloney, P. C. & Subramaniam, S. (2001). Projection structure and molecular architecture of OxlT, a bacterial membrane transporter. *EMBO J.* **20**, 4408–4413.
 10. Williams, K. A. (2000). Three-dimensional structure of the ion-coupled transport protein NhaA. *Nature*, **403**, 112–115.
 11. Williams, K. A., Geldmacher-Kaufer, U., Padan, E., Schuldiner, S. & Kuhlbrandt, W. (1999). Projection structure of NhaA, a secondary transporter from *Escherichia coli*, at 4.0 Å resolution. *EMBO J.* **18**, 3558–3563.
 12. Tate, C. G., Kunji, E. R., Lebendiker, M. & Schuldiner, S. (2001). The projection structure of EmrE, a proton-linked multidrug transporter from *Escherichia coli*, at 7 Å resolution. *EMBO J.* **20**, 77–81.
 13. Hacksell, I., Rigaud, J. L., Purhonen, P., Pourcher, T., Hebert, H. & Leblanc, G. (2002). Projection structure at 8 Å resolution of the melibiose permease, an Na-sugar co-transporter from *Escherichia coli*. *EMBO J.* **21**, 3569–3574.
 14. Hirai, T., Heymann, J. A., Shi, D., Sarker, R., Maloney, P. C. & Subramaniam, S. (2002). Three-dimensional structure of a bacterial oxalate transporter. *Nature Struct. Biol.* **9**, 597–600.
 15. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
 16. Lolkema, J. S. & Slotboom, D. J. (1998). Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Mol. Membr. Biol.* **15**, 33–42.
 17. Lolkema, J. S. & Slotboom, D. J. (1998). Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins. *FEMS Microbiol. Rev.* **22**, 305–322.
 18. Claros, M. G. & Von Heijne, G. (1994). TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **10**, 685–686.
 19. Von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494.
 20. Rabus, R., Jack, D. L., Kelly, D. J. & Saier, M. H., Jr (1999). TRAP transporters: an ancient family of extra-cytoplasmic solute-receptor-dependent secondary active transporters. *Microbiology*, **145**, 3431–3445.
 21. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403–405.
 22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
 23. Gatti, D., Mitra, B. & Rosen, B. P. (2000). *Escherichia coli* soft metal ion-translocating ATPases. *J. Biol. Chem.* **275**, 34009–34012.
 24. Pajor, A. M. (2000). Molecular properties of sodium/dicarboxylate cotransporters. *J. Membr. Biol.* **175**, 1–8.
 25. Marty-Teyssset, C., Posthuma, C., Lolkema, J. S., Schmitt, P., Davies, C. & Konings, W. N. (1996). Proton motive force generation by citrolactic fermentation in *Leuconostoc mesenteroides*. *J. Bacteriol.* **178**, 2178–2185.
 26. Bandell, M., Ansanay, V., Rachidi, N., Dequin, S. & Lolkema, J. S. (1997). Membrane potential-generating malate (MleP) and citrate (CitP) transporters of lactic acid bacteria are homologous proteins. Substrate specificity of the 2-hydroxycarboxylate transporter family. *J. Biol. Chem.* **272**, 18140–18146.
 27. Bandell, M. & Lolkema, J. S. (2000). Arg425 of the citrate transporter CitP is responsible for high affinity binding of di- and tricarboxylates. *J. Biol. Chem.* **275**, 39130–39136.
 28. Van Geest, M., Nilsson, I., Von Heijne, G. & Lolkema, J. S. (1999). Insertion of a bacterial secondary transport protein in the ER membrane. *J. Biol. Chem.* **274**, 2816–2823.
 29. Padan, E., Venturi, M., Gerchman, Y. & Dover, N. (2001). Na⁺/H⁺ antiporters. *Biochim. Biophys. Acta*, **1505**, 144–157.
 30. Wei, Y., Guffanti, A. A., Ito, M. & Krulwich, T. A. (2000). *Bacillus subtilis* YqkI is a novel malic/Na⁺-lactate antiporter that enhances growth on malate at low protonmotive force. *J. Biol. Chem.* **275**, 30287–30292.
 31. Van Geest, M. & Lolkema, J. S. (1996). Membrane topology of the Na⁺-dependent citrate transporter of *Klebsiella pneumoniae*. Evidence for a new structural class of secondary transporters. *J. Biol. Chem.* **271**, 25582–25589.
 32. Van Geest, M. & Lolkema, J. S. (2000). Membrane topology and insertion of membrane proteins: Search for topogenic signals. *Microbiol. Mol. Biol. Rev.* **64**, 13–33.
 33. Van Geest, M. & Lolkema, J. S. (2000). Membrane topology of the Na⁺-dependent citrate transporter of *Klebsiella pneumoniae* by insertion mutagenesis. *Biochim. Biophys. Acta*, **1466**, 328–338.
 34. Bandell, M. & Lolkema, J. S. (2000). The conserved C terminus of the citrate (CitP) and malate (MleP) transporters of lactic acid bacteria is involved in substrate recognition. *Biochemistry*, **39**, 13059–13067.

35. Krom, B. P. & Lolkema, J. S. (2003). Conserved residues R430 and Q428 in a cytoplasmic loop of the citrate/malate transporter CimH of *Bacillus subtilis* are accessible from the external face of the membrane. *Biochemistry*, **42**, 467–474.
36. Van Geest, M. & Lolkema, J. S. (1999). TMS VIII of the Na⁺/citrate transporter CitS requires downstream TMS IX for insertion in the *Escherichia coli* membrane. *J. Biol. Chem.* **274**, 29705–29711.
37. Sonnhhammer, E. L., Von Heijne, G. & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–182.
38. Slotboom, D.-J., Konings, W. N. & Lolkema, J. S. (2001). Glutamate transporters combine transporter- and channel-like features. *Trends Biochem. Sci.* **26**, 534–539.
39. Krom, B. P., Warner, J. B., Konings, W. N. & Lolkema, J. S. (2000). Complementary metal ion specificity of the metal-citrate transporters CitM and CitH of *Bacillus subtilis*. *J. Bacteriol.* **182**, 6374–6381.
40. Warner, J. B. & Lolkema, J. S. (2002). Growth of *Bacillus subtilis* on citrate and isocitrate is supported by the Mg²⁺-citrate transporter CitM. *Microbiology*, **148**, 3413–3421.
41. Weber, A., Menzlaff, E., Arbinger, B., Gutensohn, M., Eckerskorn, C. & Flugge, U. I. (1995). The 2-oxoglutarate/malate translocator of chloroplast envelope membranes: molecular cloning of a transporter containing a 12-helix motif and expression of the functional protein in yeast cells. *Biochemistry*, **34**, 2621–2627.
42. Pos, K. M., Dimroth, P. & Bott, M. (1998). The *Escherichia coli* citrate carrier CitT: a member of a novel eubacterial transporter family related to the 2-oxoglutarate/malate translocator from spinach chloroplasts. *J. Bacteriol.* **180**, 4160–4165.
43. Peekhaus, N., Tong, S., Reizer, J., Saier, M. H., Jr, Murray, E. & Conway, T. (1997). Characterization of a novel transporter family that includes multiple *Escherichia coli* gluconate transporters and their homologues. *FEMS Microbiol. Letters*, **147**, 233–238.
44. Janausch, I. G., Zientz, E., Tran, Q. H., Kroger, A. & Uden, G. (2002). C₄-Dicarboxylate carriers and sensors in bacteria. *Biochim. Biophys. Acta*, **1553**, 39–56.
45. Six, S., Andrews, S. C., Uden, G. & Guest, J. R. (1994). *Escherichia coli* possesses two homologous anaerobic C₄-dicarboxylate membrane transporters (DcuA and DcuB) distinct from the aerobic dicarboxylate transport system (Dct). *J. Bacteriol.* **176**, 6470–6478.
46. Enomoto, H., Unemoto, T., Nishibuchi, M., Padan, E. & Nakamura, T. (1998). Topological study of *Vibrio alginolyticus* NhaB Na⁺/H⁺ antiporter using gene fusions in *Escherichia coli* cells. *Biochim. Biophys. Acta*, **1370**, 77–86.
47. Kelly, D. J. & Thomas, G. H. (2001). The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol. Rev.* **25**, 405–424.
48. An, J. H. & Kim, Y. S. (1998). A gene cluster encoding malonyl-CoA decarboxylase (MatA), malonyl-CoA synthetase (MatB) and a putative dicarboxylate carrier protein (MatC) in *Rhizobium trifolii*—cloning, sequencing, and expression of the enzymes in *Escherichia coli*. *Eur. J. Biochem.* **257**, 395–402.
49. Ito, M., Guffanti, A. A., Zemsy, J., Ivey, D. M. & Krulwich, T. A. (1997). Role of the *nhaC*-encoded Na⁺/H⁺ antiporter of alkaliphilic *Bacillus firmus* OF4. *J. Bacteriol.* **179**, 3851–3857.
50. Hussein, M. J., Green, J. M. & Nichols, B. P. (1998). Characterization of mutations that allow *p*-aminobenzoyl-glutamate utilization by *Escherichia coli*. *J. Bacteriol.* **180**, 6260–6268.
51. Nozaki, K., Kuroda, T., Mizushima, T. & Tsuchiya, T. (1998). A new Na⁺/H⁺ antiporter, NhaD, of *Vibrio parahaemolyticus*. *Biochim. Biophys. Acta*, **1369**, 213–220.
52. Deguchi, Y., Yamato, I. & Anraku, Y. (1990). Nucleotide sequence of *gltS*, the Na⁺/glutamate symport carrier gene of *Escherichia coli* B. *J. Biol. Chem.* **265**, 21704–21708.
53. Tolner, B., Ubbink-Kok, T., Poolman, B. & Konings, W. N. (1995). Cation-selectivity of the L-glutamate transporters of *Escherichia coli*, *Bacillus stearothermophilus* and *Bacillus caldoteanax*: dependence on the environment in which the proteins are expressed. *Mol. Microbiol.* **18**, 123–133.
54. Dong, J. M., Taylor, J. S., Latour, D. J., Iuchi, S. & Lin, E. C. (1993). Three overlapping *lct* genes involved in L-lactate utilization by *Escherichia coli*. *J. Bacteriol.* **175**, 6671–6678.
55. Nunez, M. F., Teresa, P. M., Badia, J., Aguilar, J. & Baldoma, L. (2001). The gene *yghK* linked to the *glc* operon of *Escherichia coli* encodes a permease for glycolate that is structurally and functionally similar to L-lactate permease. *Microbiology*, **147**, 1069–1077.
56. Kawai, S., Suzuki, H., Yamamoto, K. & Kumagai, H. (1997). Characterization of the L-malate permease gene (*maeP*) of *Streptococcus bovis* ATCC 15352. *J. Bacteriol.* **179**, 4056–4060.
57. Kastner, C. N., Schneider, K., Dimroth, P. & Pos, K. M. (2002). Characterization of the citrate/acetate antiporter CitW of *Klebsiella pneumoniae*. *Arch. Microbiol.* **177**, 500–506.
58. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210.
59. Park, J., Teichmann, S. A., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**, 349–354.

Edited by G. von Heijne

(Received 10 October 2002; received in revised form 6 February 2003; accepted 7 February 2003)

SCIENCE @ DIRECT®
www.sciencedirect.com

Supplementary Material for this paper comprising one Table is available on Science Direct