
ter Horst, Ramon; Lolkema, Juke S.

Published in:
Biochimica et Biophysica Acta-Biomembranes

DOI:
10.1016/j.bbamem.2011.09.021

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:
2012

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

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.

Ramon ter Horst, Juke S. Lolkema *

Molecular Microbiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, The Netherlands

A R T I C L E   I N F O

Article history:
Received 10 August 2011
Received in revised form 22 September 2011
Accepted 23 September 2011
Available online 29 September 2011

Keywords:
Membrane topology
Structural classification
ST[3]
Topology prediction
MemGen
Reporter fusion

A B S T R A C T

The MemGen structural classification of membrane proteins groups families of proteins by hydropathy profile alignment. Class ST[3] of the MemGen classification contains 32 families of transporter proteins including the IT superfamily. Transporters from 19 different families in class ST[3] were evaluated by the TopScreen experimental topology screening method to verify the structural classification by MemGen. TopScreen involves the determination of the cellular disposition of three sites in the polypeptide chain of the proteins which allows for discrimination between different topology models. For nearly all transporters at least one of the predicted localizations is different in the models produced by MemGen and predictor TMHMM. Comparison to the experimental data showed that in all cases the prediction by MemGen was correct. It is concluded that the structural model available for transporters of the [st324]ESS and [st326]2HCT families is also valid for the other families in class ST[3]. The core structure of the model consists of two homologous domains, each containing 5 transmembrane segments, which have an opposite orientation in the membrane. A reentrant loop is present in between the 4th and 5th segments in each domain. Nearly all of the identified and experimentally confirmed structural variations involve additions of transmembrane segments at the boundaries of the core model, at the N- and C-termini or in between the two domains. Most remarkable is a domain swap in two subfamilies of the [st312] NHAC family that results in an inverted orientation of the proteins in the membrane.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Secondary transporters are integral membrane proteins commonly encoded by a single gene. They consist of a bundle of α-helices that are more or less perpendicular to the plane of the membrane. Despite this simple architecture, their phylogenetic diversity is enormous as evidenced by the approximately 100 gene (super)families in the electrochemical potential-driven transporters section of the Transport Classification system (TC system) [1]. Most likely, the genetic diversity is a consequence of divergent evolution and many different families may represent a similar fold and translocation mechanism. In recent years high resolution X-ray structures have been presented of transporters from different gene families that showed the same core structure while no significant sequence similarity could be identified between the transporters. The families include the Neurotransmitter:Sodium Symporter (2.A.22 NSS) family [2], the Solute:Sodium Symporter (2.A.21 SSS) family [3], the Nucleobase:Cation Symporter-1 (2.A.39 NCS1) family [4], the Betaine/Carnitine/Choline Transporter (2.A.15 BCFT) family [5] and the Amino Acid-Polyamine-Organocation (2.A.3 APC) family [6]. Structural similarity between transporters of the NSS, SSS, NCS1 and APC families was predicted by the MemGen classification system that identifies distant evolutionary relationships by hydropathy profile alignment [9,10,11]. MemGen groups families of transporters with the same global fold in structural classes. So far, 4 structural classes were defined termed ST[1], ST[2], ST[3] and ST[4]. The NSS, SSS, NCS1 and APC families are found in class ST[2]. In addition to ST[2], high resolution structures are available for class ST[1] (LacY, GlpT) [12,13] and class ST[4] (GltPh) [14]. Class ST[3] in the MemGen classification groups 32 protein families and includes the IT superfamily [15] in the TC system [1,16]. Functional characterized members of the 32 families are secondary transporters of inorganic and organic anions and Na⁺/H⁺ antiporters. Most families contain exclusively transporters from prokaryotic origin, but f. i. the Divalent Anion:Na⁺ Symporter (2.A.47 DASS) family (also known as SLCT) contains eukaryotic transporters for organic di- and tricarboxylic Krebs cycle intermediates as well as dicarboxylate amino acids, and inorganic sulfate and phosphate ions. No high resolution structure is available of any of the transporters in class ST[3]. Detailed studies of the Na⁺-citrate transporter CitS of Klebsiella pneumoniae and the Na⁺-glutamate transporter GltS of Escherichia coli, members of the [st326]2HCT and [st324]ESS families in class ST[3], respectively, have resulted in a model for the common core

Abbreviations: LIC, ligation independent cloning; GFP, green fluorescent protein; FM, fluorescein-5-maleimide; TMS, transmembrane segment

* Corresponding author at: Molecular Microbiology, Center for Life Sciences, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands. Tel.: +31 50 3632155; fax: +31 50 3632154.
E-mail address: j.s.lolkema@rug.nl (J.S. Lolkema).

0005-2736/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
structure of the transporters. The core consists of two homologous domains with opposite orientation in the membrane and containing 5 TMSs and a re-entrant (or pore) loop each [17]. Members of the two families share no significant sequence similarity, but the data strongly suggested that CitS and GltS share the same fold which confirmed their assignment to the same structural class [18].

To verify the structural classification by MemGen, we have developed the TopScreen membrane topology screening method that experimentally discriminates between different topology models [19]. TopScreen involves determination of the cellular disposition of three positions of the polypeptide chain of a membrane protein using a combination of conventional techniques and was used before to confirm the distribution of a total of 16 secondary transporters over the 4 MemGen structural classes ST[1], ST[2], ST[3] and ST[4]. Here, the TopScreen approach was used to confirm the assignment of 19 transporter families to class ST[3] thereby strongly indicating that the [st326]2HCT/[st324]ESS structural core model applies to all the families in ST[3] including the families of the T1 superfamily. Structural variations adopted by different (sub)families in class ST[3] are discussed.

2. Materials and methods

2.1. Materials

Phusion DNA polymerase was obtained from Finnzymes (Espoo, Finland). T4 ligase was obtained from New England Biolabs (Frankfurt am Main, Germany). All other enzymes were obtained from Fermentas (Burlington, Canada). Mutagenic oligonucleotides were obtained from Biolegio (Nijmegen, The Netherlands), or from Operon (Ebersberg, Germany) for ligation independent cloning. p-Nitrophenyl phosphate (pNPP) was obtained from Sigma (Zwijndrecht, The Netherlands), and fluorescein maleimide (FM) was obtained from Invitrogen (Carlsbad, United States).

2.2. Bacterial strains and growth conditions

E. coli strain SF100 (recA Δlac ompT) [20] harboring the indicated plJC vector (see below), was routinely grown in Luria Broth medium at 37 °C, with ampicillin added at a final concentration of 50 μg/ml. Overnight cultures were diluted 30-fold in 3 ml of fresh medium and when the optical density measured at 660 nm (OD660) reached a value between 0.6 and 0.8, arabinose was added at a final concentration of 0.002–0.05% (w/v) to induce protein production from the plasmids. Following growth for another 1.5–2 h, cells were harvested by centrifugation in a table top centrifuge operated at 4 °C. Cells were resuspended in the indicated buffer and kept on ice until use.

2.3. Ligation independent cloning

Ligation independent cloning (LIC) was done as described [19,22]. Briefly, a synthetic double stranded piece of DNA (the LIC cassette) was inserted downstream of the arabinose promoter in the commercial plBAD24 vector (Invitrogen) [21]. The genes encoding alkaline phosphatase and green fluorescence protein (GFP) were inserted at the 5′-end of the LIC cassette in frame with the initiation codon yielding pLIC1 or pLIC2, respectively. The two vectors were used to produce PhoA and GFP fusion proteins, respectively. Vector pLIC3 which is used to produce His-tagged proteins contains a double stop codon inserted at the same site. The second codon of the cassette in pLIC3, GGT (Gly) was mutated into TGT (Cys) by site directed mutagenesis rendering vector pLIC4 which is used for N-terminal localizations. Transporter genes were amplified using forward and backward primers containing 5′ flanking regions corresponding to the nucleotide sequences upstream and downstream of the Swal site in the LIC cassette.

2.4. GFP and PhoA assays

Alkaline phosphatase activity of E. coli SF100 cells harboring a plJC1 vector carrying the indicated insert was measured in Miller units [23] as described before [19]. GFP fluorescence emission intensity at 508 nm was measured at an excitation wavelength of 470 nm using an AMINCO Bowman Series 2 Luminescence Spectrometer. Cells of E. coli SF100 harboring a plJC2 vector carrying the indicated insert were transferred into a precision cell (Hellma, Quartz SUPR SIL). For further details see [19,24].

2.5. Evaluation of data

Mean values and standard deviations were calculated from at least three independent measurements. PhoA and GFP activities of the cells were normalized by the mean PhoA activity of all positive PhoA fusions (>100 Miller units) and all positive GFP fusions (>0.5 emission units), respectively. The logarithm of the ratio of the normalized PhoA and GFP activities was calculated for each full-length and half protein to obtain a measure for the cellular localization of the fusion point. A positive value corresponds to a periplasmic localization, a negative value to a cytoplasmic localization. The length of the bar indicates the significance of the localization and would be independent of the expression levels [19]. Values >2.5 or <−2.5 were arbitrarily set to 2.5 and −2.5, respectively.

2.6. Labeling studies

E. coli SF100 cells harboring plJC4 or pLIC3 vectors carrying the indicated insert were washed once and, subsequently, resuspended in ice-cold 50 mM potassium phosphate buffer pH 7.0. Following treatment of the cells with fluorescein maleimide and, subsequently, Ni2+-NTA affinity purification of the His-tagged transporters from the membranes, the 25 μl samples were run on a 12% sodium dodecyl sulfate-polyacrylamide gel (SDS-PAGE). In-gel fluorescence was recorded using a FujiFilm LAS-4000 luminescent image analyzer, and the gel was stained with Coomassie Brilliant Blue (CBB). Experimental details were as described before [19].

3. Results


The TopScreen approach involves the determination of the cellular disposition of three positions of a membrane protein, the N- and C-termini and a position in the central loop that is measured by the localization of the C-terminus of the half-protein [19]. The locations of the C-termini of full-length proteins and half-proteins are determined using complementary reporter fusion techniques. Alkaline phosphatase fused at the C-terminus results in an active enzyme when located in the periplasm, while fusion of GFP results in high fluorescence only when located in the cytoplasm. The N-termini are determined by the accessibility of an introduced cysteine residue by the fluorescent sulphydryl reagent fluorescein maleimide.

Table 1 lists a total of 38 proteins from structural class ST[3] of the MemGen classification (http://molmic35.biol.rug.nl/memgen/mgweb.dll) that have been evaluated by the TopScreen method. Most of the proteins have been functionally characterized as secondary transporters. All of them are from bacterial origin, originating from the phyla proteobacteria and firmicutes. The proteins are distributed over 19 different families. In general no significant sequence similarities can be detected between transporters from different families. Transporters in the [st301]CITMH5, [st324]ESS and [st326]HCT families were evaluated before [19] and were included for completeness throughout this study. Detailed topology models are available for the [st324]GLTS and [st326]HCT families. The T1 superfamily [15] from

### 3.2. C-terminal location of full-length and half-proteins

Genes encoding the full-length and half-proteins of each of the transporters in Table 1 were cloned into the appropriate vectors resulting in PhoA and GFP fusion proteins. The C-terminal residue of each half-protein was indicated in Table 1. Alkaline phosphatase activity of cells expressing PhoA fusions was plotted against the GFP fluorescence of cells expressing the corresponding GFP fusion (Fig. 1A and B, left). The data points formed two well separated groups falling along the two axes. High alkaline phosphatase activity and low GFP activity of cells expressing PhoA fusions was plotted against the sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 371 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a misannotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with cytoplasmic C-termini are found in the family of Na+/H+ antiporters [st312] NHAC, suggesting structural diversity in this family (Fig. 1A, right). Finally, YUIFbsub of Bacillus subtilis is found in the protein databases as a protein consisting of 37 residues of which the shorter is most likely a miss annotated sequence DcuAecol* was located in the cytoplasm (Fig. 1A) indicating that three of the full-length proteins with }
the protein. Half proteins of the glycolate transporter GLCAecol [38] of E. coli and YVFHbsub of B. subtilis were constructed by truncating the sequence just before and after the hydrophobic region (see Table 1). For both proteins, both versions of the half proteins ended in the cytoplasm (Fig. 1B), suggesting two TMSs in the central hydrophobic regions.

### 3.3. N-terminal location of full-length proteins

For each of the transporters in Table 1 two constructs were made to determine the location of the N-terminus, one variant with a Cys and one with a Gly residue at position 2, immediately upstream of the His$_2$-tag. Treatment of cells expressing the two variants of the transporters with 0.1 mM of the fluorescent sulphydryl reagent fluorescein maleimide (FM) for 20 min and at room temperature was used to discriminate between a cytoplasmic and periplasmic location of the N-terminus of the proteins. Under these conditions, the thiol reagent is membrane impermeable and only labels Cys residues exposed to the external medium [19]. Determination of the localization of the N-termini is the more elaborate and cumbersome of the three positions in the TopScreen approach. Main problems are the labeling of endogenous Cys residues and lack of or low expression of either of the variants. The goal was to obtain one N-terminal localization in each of the nineteen families, which succeeded for all but one family.

The N-terminus of the transporters from nine different families was clearly located in the periplasm, as was observed before for CitM of B. subtilis in family [st301]CitMHS and GltS of E. coli in [st324]ESS [Fig. 2A] [19]. Following treatment of the cells with fluorescein maleimide and Ni$^{2+}$-NTA affinity purification of the His-tagged transporters from the membranes, fluorescence imaging after SDS-PAGE showed clear labeling of the Cys variant, while the Gly variant was not labeled (Fig. 2A). Coomassie Brilliant Blue staining of the gel revealed more or less the same protein levels in the membranes. It follows that these transporters expose their N-termini at the exterior of the cell surface. Transporters from three additional families did not result in labeling of either the Gly or Cys variant as was observed before for CitS of K. pneumoniae in the [st326]2HCT family (Fig. 2B). Both variants were clearly produced in the membrane and it was concluded that the N-termini of YBHIecol in [st301]DASS, the Na$^+$/H$^+$ antiporter NhaBecol [32] in [st306]NhaB and YCGAbsub in [st313]AitB reside in the cytoplasm where they are inaccessible to FM. The latter protein, YCGAbsub of B. subtilis is found in the protein databases containing 396 and 477 residues of which the shorter sequence is, most likely, a misannotated sequence truncated at the N-terminus. The N-terminus of the truncated sequence YCGAbsub was exposed to the periplasm (Fig. 2A) suggesting that the N-terminal 81 residue region of YCGAbsub contains an odd number of TMSs.

Several of the transporters from different families showed labeling in the Gly variant, indicating the accessibility of endogenous Cys residues from the periplasmic side of the membrane (Fig. 2C). In case of the C4-dicarboxylate transporter DCUCecol [33] in [st309]DcuC, the aminobenzoyl-glutamate transporter AbgTecol [36] in [st314]AbgT and the aspartate/alanine exchanger AspTstrt [37] of Streptococcus thermophilus in [st315]AAE the labeling intensity of the Gly variants was modest and the ratio of fluorescence intensity over protein stain of the bands was considerably higher for the Cys than the Gly variants indicating labeling of the N-terminal Cys residue and a periplasmic location (Fig. 2C). It is concluded that the N-termini of the Cys variant of these transporters are exposed at the exterior of the cell. The Gly and Cys variants of the tartrate/succinate exchanger Td Tecol [27] in [st303]DASS and YP350496flu in [st310]ATO showed similar ratios of
fluorescence intensities over protein stain, indicating no additional labeling in the latter variant and a cytoplasmic location of the introduced Cys residue (Fig. 2D). The Cys variant of YUIFbsub in \([st312]\)NHaC showed high fluorescence intensity over protein stain, which generally indicates a periplasmic localization (see panel A). However, a particular strong labeling was observed for the Gly variant of YUIFbsub which in combination with the low expression level of the Cys variant makes it difficult to evaluate the ratios of fluorescence intensity over protein stain rendering the localization inconclusive.

3.4. Comparison TopScreen results and MemGen predictions

Together with the results from a previous study [19], the TopScreen topology screen resulted in a complete set of three localizations for 20 proteins covering 17 families in structural class ST[3] in the MemGen classification (Table 2). In addition, combining the results from different proteins in the \([st325]\)LctP family resulted in a complete set for this family which leaves the \([st312]\)NHaC family as the only family for which only the location of the central loop and C-terminus could be determined. The proteins in thirteen families follow the out–in–out (o/i/o) order for the position of the N-terminus, central loop, and C-terminus, respectively, which is according to the putative core structure of class ST[3] consisting of two domains with 5 TMSs each with the connecting central loop in the cytoplasm. Included in this group is the \([st325]\)LctP family, even though the hydrophobic properties of the central loop suggest additional TMSs in this region (see below). Transporters from five families have a cytoplasmic N-terminus (i/i/o) which would be in line with the membrane topology of CitS of \(K.\) pneumoniae (\([st326]\)2HCT) that has an additional TMS N-terminal to the core structure. Transporters of the \([st312]\)NHaC family have more complicated topologies that will be discussed below.

Table 2 gives the predicted localizations of the three positions of the proteins by MemGen (see below) and by TMHMM, a widely used membrane topology predictor [40]. For nearly all transporters at least one of the predicted localizations is different in the models produced by MemGen and TMHMM. Comparison to the experimental data shows that in all cases the prediction by MemGen is correct. While TMHMM predicts the location of the central loop mostly in
the cytoplasm in line with the experimental results, the number of TMSs in the two halves of the proteins is often at variance with the experiments which, as noted before [17,18], would be due to the presence of the reentrant loop structures in the two domains.

### 3.5. Diversity of membrane topology in class ST[3]

A schematic representation of the predicted structural models for the different families in ST[3] is shown in Fig. 3. The core structure, based on extensive studies of the membrane topology of GltS of E. coli ([st324]ESS family) and CitS of K. pneumoniae ([st326]2HCT), is represented by structure A. It consists of two domains with 5 TMSs each (5 + 5 topology) that are oppositely oriented in the membrane (inverted topology). In between the 4th and 5th segments in each domain a reentrant loop is found. In 10 families of class ST[3] the core structure is the topology adopted by most, if not all, transporters (Fig. 3, structure A). Deviations from the core structure were identified when the alignment of the family hydropathy profile of a family with the profile of one of these families showed additional hydrophobic regions or, less frequently, lacked hydrophobic regions [10]. Alignment of the family profiles of the [st324]ESS family representing the core structure and the [st326]2HCT was shown before and clearly revealed an additional TMS at the N-terminus of the proteins in the latter family (1 + 5 + 5 topology; structure C) [18]. Fig. 4A shows the additional N-terminal TMS of the [st326]2HCT family by alignment with the [st305]Dcu family. The additional TMS was also identified in the family profiles of the [st303]DASS and [st306]NhaB families and is in agreement with the TopScreen results (Table 2). The alignment in Fig. 4A also shows that the locations of the C-terminus of the C-terminally truncated and full-length DCUAecol, cytoplasmic and periplasmic, respectively, are in agreement with the model even though the hydropathy profile clearly shows two hydrophobic regions in between the two positions.

The family profiles of the [st310]Ato and [st313]AIB families indicated that the first TMS of the core structure was missing yielding a 4 + 5 topology (structure B). The N-terminal localization of the N-terminally truncated YCGAbub protein in the periplasm would be in agreement with this model (Fig. 4B). The same topology is found in the [st302] ArsB family, but only in two subfamilies. The remaining three subfamilies fold according to core structure A (Fig. 3). Therefore, different transporters in one family have different numbers of TMSs.

The transporters of the [st307]TRAP-T family are exceptional as they are believed to be part of a complex that forms a transport system that appears to be a hybrid of a secondary and a ABC type of transport system [41]. The complex is believed to consist of the transporter subunit M, a smaller integral membrane protein subunit Q containing 4 TMSs and a periplasmic binding protein Q. Subunits M are found in the IT superfamily and in the MemGen model they fold into the core structure of ST[3] (Fig. 3A). In a small fraction of the transporters (~6%), the small subunit Q is fused to the N-terminus of the M subunit via a linker TMS resulting in topology D (4 + 1 + 5 + 5 topology). Even smaller fractions of the M subunits contain insertions of a pair of TMSs in between the two domains or between the second and third TMSs of the C-domain (not indicated).

Alignment of the family profile of the [st325]LctP family and the [st326]2HCT family (Fig. 4C) shows that the hydrophobic region in the central loop is not part of the core structure but, in line with the experimental results, represents two additional TMSs. The pair of TMSs would be inserted in between the two domains of the core structure (5 + 2 + 5; structure G), a feature that is present in all transporters in all three subfamilies. A minority of about 15% of the transporters in subfamily [st325]LctP has an additional TMS at the C-terminal end of the core structure (5 + 2 + 5 + 1; structure H).

The most diverse membrane topologies are found for the transporters in family [st312]NhaC. Hydropathy profile alignments show that the transporters in subfamily [st312]NhaC4 and a significant fraction of the transporters in [st312]NhaC2 fold according to the core structure A in line with the experimental evidence for YUGBub and NP149303ace, respectively. The predominant topology (~70%) in the [st312]NhaC2 subfamily is the core structure plus an additional TMS at the C-terminus (structure F) which was expermentally confirmed by the C-terminal localization of full length NP207738hpyl of H. pylori (Fig. 1A). The [st312]NhaC family is the only family with transporters that have the central loop in the periplasm rather than the cytoplasm (Fig. 1). This is caused by a domain swap that was reported before [16]. The domain swap causes a topology inversion and leaves the N- and C-termmi in the cytoplasm (structure E). The results of the TopScreen analyses of the C-terminal localizations of transporters YUGBub and MLENub of B. subtilis in subfamily [st312]NhaC1 are in agreement with the domain swap. Sequence analysis showed the same domain swap in subfamily [st312]NhaC3, while subfamilies [st312]NhaC2 and [st312]NhaC4 follow the ‘normal’ order of the two domains.

### 4. Discussion

#### 4.1. Validation of structural classification

Structural class ST[3] of the MemGen classification groups 32 families containing 59 subfamilies that are believed to share the same fold.
[16]. Class ST[3] includes in addition to the IT superfamily [15] of the TC classification system [1] several other families indicating that hydropathy profile analysis detects more distant evolutionary relationship between (membrane) proteins than sequence analysis. Before, detailed studies of CitS of K. pneumoniae, a member of the [st326]2HCT family, and GltS of E. coli, a member of the [st324]ESS family, provided strong support for the same fold of the transporters in these two ST[3] families that showed no sequence identity whatsoever [17,18]. TopScreen was developed to allow for discrimination between different membrane topology models of many more families by strategically selecting three positions in the polypeptide chain for experimental determination of their localization inside or outside the cell. Here, TopScreen was used to validate the structural classification of 19 secondary transporter protein families in class ST[3]. In the case of ST[3] families, the presence of putative reentrant loops in the MemGen structural model was particularly helpful as secondary structure prediction methods often mistake reentrant loops for TMSs which results in inversion of the topology of adjacent segments. For most proteins, predictor TMHMM placed the central loop in the cytoplasm which is in agreement with the experimental data (Table 2). However, the presence of the putative reentrant loops in the two halves of the proteins makes a strong case for the MemGen topology since the location of either or both of the N- and C-termini predicted by TMHMM is at variance with the TopScreen results. In conclusion, the present study suggests that the ‘core’ structure identified in the [st324]GltS and [st326]2HCT families is also valid for the IT superfamily transporters and other families in class ST[3].

### 4.2. Comparison validated and previously reported topology models

A number of topology models of ST[3] transporters have been reported previously. ArsB of E. coli is a member of the [st302]ArsB family and belongs to the IT superfamily. A variety of lacZ, phoA and blmM gene fusions resulted in a topological model consisting of 12 membrane-spanning regions, with the central loop and the N- and C-termini localized to the cytoplasm [42]. The model is at variance with the TopScreen results and the MemGen model (Fig. 3, Model A). Models reported for the C4-dicarboxylate transporter DcuA of E. coli, ([st305]Dcu) [43] and the aspartate:alanine antiporter AspT of Tetragenococcus halophilus ([st315]AAE) [44] reveal two times 5 TMSs with the N- and C-termini outside the cell which is in agreement with the TopScreen results. In both models the reentrant loops in the two halves were absent (Fig. 3, Model A). Similarly, the cellular localization of the N- and C-termini and central loop of the topology model reported for the Na+/H+

### Table 2. ST[3] (sub)family membrane topology

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Membrane Topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>[st302]ArsB3, [st302]ArsB5</td>
<td>B</td>
</tr>
<tr>
<td>[st307]TRAP-T(58%)</td>
<td>D</td>
</tr>
<tr>
<td>[st307]TRAP-T(88%)</td>
<td>E</td>
</tr>
<tr>
<td>[st310]Ato, [st313]AIT-B</td>
<td>F</td>
</tr>
<tr>
<td>[st312]NHAC2(52%), [st312]NHAC4</td>
<td>G</td>
</tr>
<tr>
<td>[st312]NHAC1, [st312]NHAC3</td>
<td>H</td>
</tr>
<tr>
<td>[st312]NHAC2(88%)</td>
<td>I</td>
</tr>
<tr>
<td>[st325]LctP1(14%)</td>
<td>K</td>
</tr>
</tbody>
</table>

Fig. 3. Membrane topology models of structural class ST[3]. See text for explanation. Boxed parts represent domains.
antiporter NhaB of *Vibrio alginolyticus* ([st306]NhaB) [45] was in agreement with the TopScreen results and the hydrophobic regions that correspond to the reentrant loops in the MemGen model were identified as periplasmic/cytoplasmic loops. However, the NhaB model contains 4 TMSs in the N-terminal half, while the MemGen model contains 6 TMSs (Fig. 3, Model C). Moreover, a reentrant loop was proposed between the 1st and 2nd TMSs. This example shows a limitation of the TopScreen method when different models give the same cellular disposition for the three sites tested. The [st307]TRAP-T family is another example. While no experimental data is available, a consensus structure based on TMHMM [40] predictions of a number of sialic acid transporters ([SiaQM]) was proposed that shows 7 TMSs in the transporter domain [41]. The structure was proposed to have evolved from a protein with two homologous domains containing 5 TMSs each in which 2 TMSs were inserted in between the two domains. TopScreen does not discriminate between this model and the MemGen model that predicts a topology according to the core structure of class ST[3]. Finally, the [st303]DASS family contains many transporters of eukaryotic origin among which the well-studied human Na⁺–dependent dicarboxylate transporters NaDC1 and NaDC3 found in the apical and basolateral membranes of cells of renal proximal tubule, respectively. Topology models of these transporters are based on hydropathy profile analyses and

---

**Fig. 4.** Family averaged hydropathy profile alignments. Alignment of the family averaged hydropathy profiles of [st326]2HCT (blue) and (A) [st305]Dcu (red), (B) [st313]AIT-B (red) and (C) [st325]LctP (red). Red and blue arrows indicate the positions and outcome of the experimental TopScreen localizations (in/out) of one or more members of the corresponding families (see also Figs. 1B, C and 2A, B). Topology models are depicted above ([st305]Dcu, [st313]AIT-B and [st325]LctP), or below ([st326]2HCT) the corresponding alignments. Red squares, TMSs; green, external loops; blue, cytoplasmic loops; thickened loop regions, reentrant loops. The numbers of typical sequences [9] included in the family hydropathy profiles were 45, 31, 67 and 62 for [st326]2HCT, [st305]Dcu, [st313]AIT-B and [st325]LctP, respectively. The Structure Divergence Score (SDS) which is a measure of the similarity of the profiles after alignment [9] was 1.23 (A), 0.793 (B), and 0.751 (C).
secondary structure prediction algorithms and reveal 11 to 14 TMSs [46]. The models were never tested thoroughly even though a significant body of SCAM (substituted-cysteine accessibility method) data is available. The topology model in use for the best studied transporter in the family NaDC1 shows 11 TMSs with the N- and C-termini intra and extracellular, respectively, which is in line with the TopScreen data. The MemGen model adds the two reentrant loops to this model (Fig. 3, Model C). An alternative model based on studies of human NaDC3 contains a reentrant loop between TMSs VIII and IX [47].

4.3. Structural diversity in transporter class ST[3]

The core structure common to the transporters in class ST[3], and also the complete structure in many of the families (Fig. 3), consists of two homologous domains with inverted topology and each containing 5 TMSs with a reentrant loop in between the 4th and 5th TMSs. Variations found in different families include an additional TMS at the N-terminus (1 + 5 + 5), two additional TMSs in between the two domains (5 + 2 + 5), and an additional TMS at the C-terminus (5 + 5 + 1). Combinations of these additions were observed as well (i.e. [st325] LctP1). Some of the additional TMSs are characteristic for a particular family (structures C and G/H), while in other families their presence appears to be optional. In many families, a minority of members have additional TMSs at the N- or C-termini. These were not included in Fig. 3 because they were not considered to be typical. Similar variations in the number of TMSs are observed in transporters of other structural classes. X-ray structures of transporters from 5 different families in class ST[2] show a core structure formed by an inverted structural repeat of 2 times 5 TMSs (different from the proposed ST[3] core structure) [2–8]. In addition, two TMSs are found at the N-terminus of BetP of the betaine/choline/carnitine transporters (BCCT) family [5] and two at the C-termini of LeuT of the neurotransmitter sodium symporter (NSS) family [2], Mhp1 of the nucleobase-cation-syntporter-1 (NCS) family [4] and AdIC and ApIC of the amino acid/polyamine/organocation (APC) superfamily [6–8]. In the same class, vSGLT in the sodium solute symporter (SSS) family has one additional N-terminal TMS and three additional C-terminal TMSs [3]. In class ST[1] containing the Major Facilitator Superfamily (MFS) transporters, the core structure is formed by two homologous domains containing 6 TMSs each. Transporters in the drug/H+ antiporter family (TC 2.A.1.3 DHA) consist of 14 TMSs with the two extra TMSs inserted in between the two domains, similar to the [st325] LctP transporters in class ST[3] [48]. The TMSs that are in addition to the core structures are found at the boundaries of the core, i.e. at the N- and C-termini and in between the two repeats which leaves the structural integrity of the core intact. An insertion in the central loop is not exceptional when the core is considered to be built from two domains with the same fold that fold upon one another [17,18]. The two domains are connected by the central loop, but, from a structural point of view, might as well be connected by the N- and C-terminal ends of the core, while having the central loop disconnected. The latter structure would correspond to a swapping of the two domains as is actually observed in the [st324]ESS family in class ST[3]. The two domains of GltS of E. coli from the [st324]ESS family in class ST[3] were swapped by genetic engineering [51]. GltSswap was equally active as the wild type protein demonstrating that from a structural or functional perspective the order of the two domains is irrelevant.

One structural variation in class ST[3] was identified that affected the 5 + 5 core structure. Hydrodraphy profile alignment showed that secondary transporters from the families [st310]Ato, [st313]ATIB and two subfamilies of [st312]ArsB are missing the first TMS of the core structure, rendering a 4 + 5 topology (Fig. 3). In case of the [st312]ArsB subfamilies, deletion of the N-terminal TMS was confirmed by sequence alignment with the other subfamilies. The data suggests that the core structure as a whole is not essential for functional and structural integrity of the class ST[3] transporters.

Acknowledgements

This work was supported by a grant from The Netherlands Organization for Scientific Research (NWO-CW).

References
