10702 *Biochemistry* **2010,** *49,* 10702–10713
DOI: 10.1021/bi101256x

BIOCHEMISTRY
including biophysical chemistry & molecular biology
Article

A Study of the Evolution of Inverted-Topology Repeats from LeuT-Fold Transporters Using AlignMe[†]

Kamil Khafizov,[‡] René Staritzbichler,[‡] Marcus Stamm, and Lucy R. Forrest*

Computational Structural Biology Group, Max Planck Institute of Biophysics, 60438 Frankfurt am Main, Germany.
[‡]*These authors contributed equally to this work*

Received August 6, 2010; Revised Manuscript Received November 3, 2010

ABSTRACT: X-ray crystal structures have revealed that numerous secondary transporter proteins originally categorized into different sequence families share similar structures, namely, the LeuT fold. The core of this fold consists of two units of five transmembrane helices, whose conformations have been proposed to exchange to form the two alternate states required for transport. That these two units are related implies that LeuT-like transporters evolved from gene-duplication and fusion events. Thus, the origins of this structural repeat may be relevant to the evolution of transport function. However, the lack of significant sequence similarity requires sensitive sequence search methods for analyzing their evolution. To this end, we developed a software application called AlignMe, which can use various types of input information, such as residue hydrophobicity, to perform pairwise alignments of sequences and/or of hydropathy profiles of (membrane) proteins. We used AlignMe to analyze the evolutionary relationships between repeats of the LeuT fold. In addition, we identified proteins from the so-called DedA family that potentially share a common ancestor with these repeats. DedA domains have been implicated in, e.g., selenite uptake; they are found widely distributed across all kingdoms of life; two or more DedA domains are typically found per genome, and some may adopt dual topologies. These results suggest that DedA proteins existed in ancient organisms and may function as dimers, as required for a would-be ancestor of the LeuT fold. In conclusion, we provide novel insights into the evolution of this important structural motif and thus potentially into the alternating-access mechanism of transport itself.

Secondary transporters, present in all biological cells, use the free energy stored in ion concentration gradients to drive the transport of substrates across membranes (*1*, *2*). In the past several years there have been significant strides made in the determination of X-ray crystallographic structures of secondary transporters (*3*, *4*). Interestingly, the structural data have revealed that some transporters previously assigned to different sequence families nevertheless share a similar 3D architecture. One such group of structures so far includes seven different proteins, originating from five different sequence families, namely, AdiC and ApcT from the amino acid/polyamine/organocation (APC)[1] transporter family (*5*, *6*); BetP and CaiT from the betaine/carnitine/choline transporter (BCCT) family (*7*, *8*); Mhp1 from the nucleobase:cation symporter-1 (NCS1) family (*9*); LeuT, a bacterial homologue of the neurotransmitter:sodium symporters (NSS) (*10*); and vSGLT from the solute:sodium symporter (SSS) family (*11*). These proteins share a common structural core consisting of two sets of five transmembrane (TM) helices, often called a 5TM repeat (Figure 1). Together, these repeated 5TM units are interwoven to form a 10TM helix subdomain (Figure 2A) whose structure we refer to as the LeuT fold, since this was the first of the structures to be determined. Alongside this core 10TM helix structural unit, most of the transporters contain additional peripheral TM helices, either at the carboxy- (C-) terminal end, at the amino- (N-) terminal end, or both (Figure 1).

The 5TM helix segments within these transporters are related by a pseudo-2-fold axis of symmetry that runs along the plane of the membrane, and their structures superimpose relatively well (the root-mean-squared difference in the positions of the Cα atoms is ~3−5 Å; Figure 2B). We have previously shown that the alternate conformations of LeuT-fold transporters required for transport can be modeled by threading the sequence of one of the repeat units onto the structure of the other, and vice versa (*12*, *13*). The conformational change thus predicted is consistent with that recently described for Mhp1 based on crystal structures of inward- and outward-facing states (*9*). Together, these findings demonstrate that the 5TM inverted-topology repeats play a central role in the alternating-access mechanism of transport, by allowing the protein to adopt two symmetry-related states. Furthermore, the conservation of this fold over a large number of transporters and transporter families suggests that they all share a common mechanism that involves these structural repeats.

Given the structural similarity of the 5TM repeat units (Figure 2B), it seems a reasonable assumption that the LeuT-like structural fold evolved from ancient internal gene-duplication and

[†]This work was supported by the DFG (German Research Foundation), Collaborative Research Center 807 "Transport and Communication across Biological Membranes".

*Corresponding author. Tel: +49 69 6303 1600. Fax: +49 69 6303 1502. E-mail: lucy.forrest@biophys.mpg.de.

[1]Abbreviations: TM, transmembrane; APC, amino acid/polyamine/organocation; BCCT, betaine/carnitine/choline transporter; NCS1, nucleobase:cation symporter-1; NSS, neurotransmitter:sodium symporters; SSS, solute:sodium symporter; WW, Wimley and White hydrophobicity scale; HWvH, Hessa, White, and von Heijne hydrophobicity scale.

 Published on Web 11/12/2010

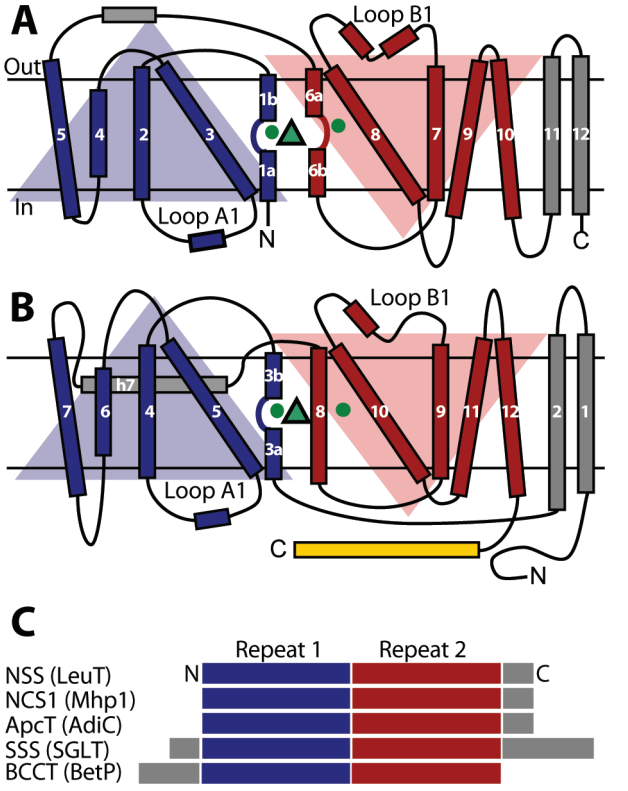

FIGURE 1: Schematics of the TM topologies of the LeuT fold in (A) LeuT, (B) BetP, and (C) comparing all known structures. These transporters share the core architecture of LeuT-like secondary transporters, consisting of a 5TM helix inverted-topology repeat (blue and red). Peripheral TM helices, the amphipathic helix h7 of BetP, and its analogue in LeuT are colored gray. The first loop of each repeat, i.e., loops A1 and B1, is labeled. In BetP, the long cytoplasmic C-terminal helix is colored yellow. Substrates and ions are indicated using green triangles and circles, respectively.

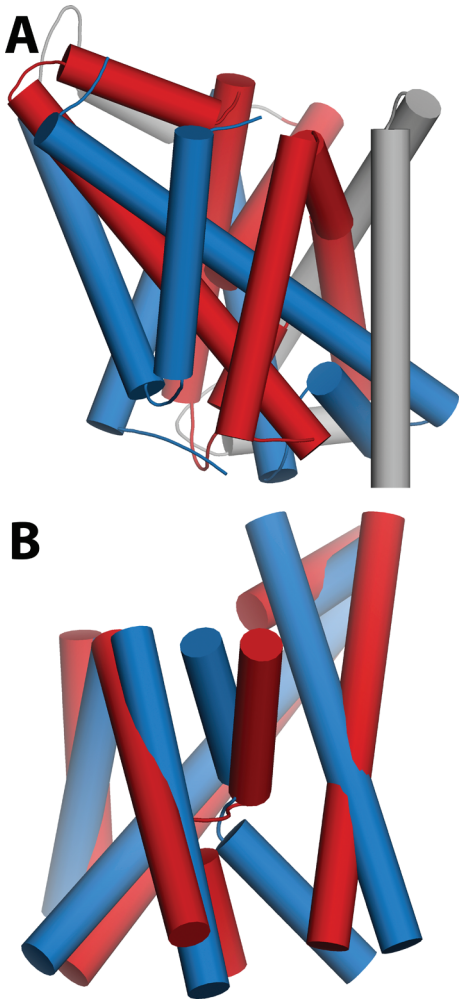

FIGURE 2: Atomic structure of the LeuT fold and its two 5TM helix structural repeats. (A) The X-ray structure of LeuT viewed from the membrane plane. Coloring is as described in the legend for Figure 1. (B) The two 5TM repeat units of LeuT superimposed using SKA (*57*). Loops have been omitted for clarity.

fusion events (*14*), probably involving a dual-topology intermediate (*15*−*19*). Numerous other membrane proteins, e.g., in the major facilitator superfamily (MFS) of transporters (*20*, *21*) or aquaporins (*22*), contain two such structurally related domains, which can even contain conserved sequence motifs, providing clear examples of evolutionary gene-duplication and fusion events (*23*, *24*). Furthermore, it has been shown that genes coding for proteins of the DUF606 family exist as singletons, as gene pairs, and as two-gene fusions, demonstrating that two homologous membrane protein domains may indeed fuse to form larger proteins (*15*, *18*).

In the case of the LeuT fold, the gene-duplication event(s) would have required the presence of one or more proteins with five TM helices that presumably shared a similar three-dimensional structure with the individual repeat units in the LeuT fold. Moreover, these 5TM proteins may have functioned as homo- or heterodimers, either as transporters or in some other functional role. Potentially, the descendants of such single-unit proteins still exist, in which case they should provide important insights into the evolutionary origins of the inverted-topology repeat motif that appears to be so fundamental to transporter function in this family. Unfortunately, however, identification of putative 5TM proteins related to the LeuT-like transporters is very challenging, given the fact that sequence relationships between these proteins have been difficult to identify (*2*, *25*, *26*). Moreover, even within a given transporter no significant sequence similarity is found between the two inverted repeat units (*7*, *12*). Thus, standard sequence search tools such as PSI-BLAST (*27*) are unlikely to be able to identify such remote relationships.

In the case of membrane proteins it has been shown that hydropathy profiles (*28*), i.e., one-dimensional representations of the hydrophobicity of the underlying amino acid sequence, provide a simple fingerprint of the TM topology of the protein (*29*). Such profiles are relatively tolerant to mutations because different amino acids can have similar hydrophobicity values and because they are typically constructed by averaging over a sliding window along the sequence (*30*, *31*). Thus, in the case of low sequence similarity, comparison of α-helical membrane protein sequences by alignment of amino acid hydrophobicity values can be more accurate than by alignment based on sequence alone (*30*−*32*). In addition, the output profile alignments can be readily interpreted and compared in graphical form to identify analogous structural features.

Here, we use hydropathy profile alignments, implemented in a new program, called AlignMe, to investigate the evolution of the LeuT fold. After assessing the quality of the alignments, we use the methodology to compare the repeats of known structure within and between transporters and to search for a 5TM protein that might share a common ancestor of those repeats. We show how one candidate protein family, the DedA proteins, shares many similarities with the LeuT repeats and has a broad sequence

distribution, suggesting that DedA domains may have undergone gene-duplication and fusion events that resulted in the LeuT fold.

EXPERIMENTAL PROCEDURES

Hydropathy Profile Alignments. Alignments of hydropathy profiles were calculated using our new program AlignMe (Alignment of Membrane proteins), which is available for download and as a server at http://www.forrestlab.org/AlignMe. Hydropathy profiles were constructed using one of two well-known hydrophobicity scales, namely, from Wimley and White (WW (*33*)) and from Hessa, White, and von Heijne (HWvH (*34*)). The method for aligning profiles is based on that of Lolkema and Slotboom (*31*), i.e., a Needleman−Wunsch dynamic programming algorithm, but with several modifications from their implementation. First, when introducing gaps, we distinguish between different regions of the profile (see below) and carry out automated optimization of the relevant gap penalties using a test search procedure. Second, profiles are averaged using a triangular sliding window 15 residues in length, rather than a rectangular window, so as to retain more of the detail in the profiles, as inspired by ref *35*. And finally, scoring of the alignments during searches was designed to penalize the introduction of gaps, as described below.

When considering whether to introduce gaps in the alignments, we attempt to disfavor gaps in the hydrophobic regions of the sequences (typically corresponding to TM helices) and, analogously, to favor gaps in hydrophilic regions (which correspond to loops). Specifically, the penalties for opening or extending a gap at any given position are selected based on whether the hydrophobicity value of the preceding "nongap" position is above or below a given hydrophobicity threshold (0 or −0.5 for the WW or HWvH scales, respectively). In addition, to maintain the compactness of the hydropathy profiles, terminal gaps were treated as distinct from "internal" gaps. For sequence alignment based on similarity matrices, it is common practice to not penalize terminal gaps at all, i.e., using a zero gap insertion penalty (*36*−*38*). However, when aligning hydropathy profiles, using zero for the terminal gap penalty causes both profiles to be aligned against terminal gaps, because the optimal value of any cell in the dynamic programming matrix (i.e., the difference between two identical hydrophobicity values) is also zero. At the other extreme, if the penalty for introducing terminal gaps is exactly the same as for "internal" gaps, then it becomes difficult to align sequences of very different lengths. In such cases, the hydrophobic peaks of the shorter sequence become distributed along the length of the longer sequence, resulting in long insertions between, or even within, TM helices. Thus, we use smaller values for penalizing terminal gaps than internal gaps for such cases. In total, we used five different gap penalties: for terminal gaps (p^{ter}); for opening gaps in hydrophobic (p_o^{hp}) and hydrophilic regions (p_o^{hf}); and for extending gaps in hydrophobic (p_e^{hp}) and hydrophilic regions (p_e^{hf}).

Optimization of Gap Penalties Used in the Search for 5TM Proteins. In order to identify the best combination of gap penalties for our 5TM protein search, we first made the assumption that the ideal gap penalties are those that rank hydropathy profile alignments of 5TM repeats of known structure (e.g., repeats of BetP or LeuT) high among alignments with many "random" membrane proteins (decoys). Thus, ten sequences corresponding to 5TM repeats of known-structure transporters (two repeats of five transporters) were added to the data set of 210 sequences from a single bacterium (*Corynebacterium glutamicum*), which had been prefiltered on the basis of their hydrophobicity and length (see below). The final set contains 220 sequences.

All five gap penalties in our search (see above) were allowed to change in the range 0.5−2.5 with a step size of 0.5 (in units of hydropathy), except for p^{ter}, which had a maximum value of 1.5, and p_o^{hp}, which had a maximum of 3.0. In addition, we also imposed the following conditions: $p^{ter} \leq p_e^{hf} \leq p_o^{hf} \leq p_e^{hp} \leq p_o^{hp}$.

All ten 5TM repeat sequences were converted to hydropathy profiles, and each was then used as a query, i.e., aligned against the hydropathy profile of every sequence in the augmented *C. glutamicum* sequence data set, using all allowed combinations of gap penalties. Results for each query and each gap penalty combination were sorted on the basis of their gap-dependent profile difference scores, GPDS (see below). We identified the ranks of the ten known sequences within this list and summed these ten ranks in order to obtain a measure (or score) of the accuracy of each gap penalty combination for a given query sequence. Thus, a smaller "sum of ranks" reflects that a specific gap penalty combination gives good alignments of the ten known sequences to that query. To account for the fact that we used multiple queries, the different gap penalty combinations were then sorted according to those "sums of ranks", and the combination with the smallest total rank over all ten queries was selected as the optimal combination. This second step of reranking for all queries helps to minimize the influence of large differences between repeats and instead favors gap penalty combinations that are more generally useful. The gap penalties with the best ranking over all queries are $p^{ter} = p_e^{hf} = p_o^{hf} = p_e^{hp} = 1.0$, $p_o^{hp} = 2.5$.

Scoring of Hydropathy Profile Alignments. Two schemes were used for scoring the hydropathy profiles. The profile difference score (PDS) used by Lolkema and Slotboom (*31*) is defined as the root-mean-squared difference between the hydrophobicity values at aligned positions in the two hydropathy profiles, while the gap-dependent profile difference score (GPDS) is defined as the last element of the corresponding dynamic-programming matrix (which comprises the differences in aligned hydrophobicity values and the gap penalties used in that alignment) divided by the number of aligned (nongap) positions. Thus, GPDS explicitly takes into account the sum of the penalties due to all introduced gaps. One potential problem with the latter scheme is that it will penalize alignments in which the sequences differ significantly in length, even if the hydropathy profiles in the aligned region are very similar. On the other hand, GPDS is less likely than PDS to find false positives in which the hit contains more hydrophobic peaks than the query, because PDS matches peaks optimally and does not penalize insertion of long gaps. Thus, PDS is more suitable for assessing the absolute similarity between two profiles, while GPDS is more suitable for identifying relationships even in sequences of very different lengths.

Search for the Evolutionary Ancestors of 5TM Inverted Repeats in the LeuT Fold. Five full-length sequences of transporters of known structure (one per family) were selected as queries for the search for putative 5TM proteins: arginine: agmatine antiporter AdiC (PDB identifier 3HQK (*5*)), glycine betaine:sodium symporter BetP (2WIT (*7*)), amino acid:sodium symporter LeuT (2A65 (*10*)), benzylhydantoin:sodium symporter Mhp1 (2JLN (*39*)), and sodium:galactose symporter vSGLT (3DH4 (*11*)). The sequences of the ten repeat units (two per protein) were prepared in such a way that they include the characteristic 5TM helices plus a few (typically three to five) residues either end, ensuring that the corresponding hydropathy

profile starts and ends with values below the hydrophobicity threshold, where possible.

We searched for a 5TM helix homologue of the 5TM repeat units in the genomes of 20 bacteria from a range of phyla, using one representative per phylum, which were obtained from the NCBI in February 2010. These bacteria are *Acetobacter pasteurianus*, *Acholeplasma laidlawii*, *Acidimicrobium ferrooxidans*, *Acidobacterium capsulatum*, *Acinetobacter baumannii*, *Anabaena variabilis*, *Anaeromyxobacter dehalognans*, *Arcobacter butzleri*, *Aromatoleum aromaticum*, *Bacteroides fragilis*, *Borrelia afzelii*, *Candidatus protochlamydia*, *Chloroflexus aggregans*, *Corynebacterium glutamicum*, *Deinococcus deserti*, *Escherichia coli*, *Fervidobacterium nodosum*, *Hydrogenobaculum* sp. Y04AAS1, *Leptotrichia buccalis*, and *Rhodopirellula baltica*. We chose not to scan all available bacterial genomes in order to reduce the redundancy in the results and to minimize the computational cost. Since sequences of 5TM proteins are expected to fall within a certain length range, we excluded sequences shorter than 140 or longer than 320 amino acid residues, reducing the data set from 63420 to 24077 sequences. In addition, protein sequences were only included in the search if at least 20 positions in the window-averaged hydropathy profile were located in the hydrophobic region. In total there were 11987 protein sequences in the final data set.

Hydropathy profiles based on the HWvH hydrophobicity scale were generated for every sequence in the data set, and each was then aligned against the HWvH profiles of the ten query sequences using AlignMe. Thus, we calculated 119870 alignments in total (ten repeat units $\times$11987 sequences), which were then ranked according to their GPDS values.

Constructing Family-Averaged Hydropathy Profiles. Homologues of sequences of interest were collected from PSI-BLAST searches (*27*, *40*) of the National Center for Biotechnology Information (NCBI) nonredundant database from January 2010. Five PSI-BLAST iterations were run, with a next-round cutoff of 0.005, a final *E*-value cutoff of 10^{-4}, and retaining 2500 hits. Sequences containing nonstandard amino acid types were removed. To exclude incomplete sequences (fragments) and sequences possessing atypically long terminal domains or insertions, the PSI-BLAST hits were filtered so that their length was in the range $\mu \pm \sigma$, where μ is the mean length of all the hits and σ is the standard deviation thereof.

Multiple-sequence alignments of homologues were constructed from the filtered set of PSI-BLAST results using MUSCLE with default settings (*41*). MUSCLE was previously shown to compare favorably with other commonly used multiple-sequence alignment algorithms for membrane proteins of $>40\%$ sequence similarity (*42*). Positions (i.e., columns) in the resultant multiple-sequence alignment in which $>50\%$ of the sequences contained a gap were excluded from the family-averaged profiles.

Gap penalty values for aligning family-averaged profiles were based on those optimized for the search procedure, but adjusted depending on the specific goal. For aligning family-averaged profiles of entire LeuT-fold sequences, we needed to allow for terminal gaps, which occur frequently in alignments of sequences with significant variations in length, or with different organization of the TM helices, as in LeuT transporters (Figure 1). The gap penalty values used in that case were $p^{ter} = 0.25$, $p_o^{hf} = p_e^{hp} = 1.0$, $p_e^{hf} = 0.85$, and $p_o^{hp} = 2.5$. For aligning family-averaged profiles of individual 5TM repeats with one another, where the lengths are similar, alignments in good agreement with known structural relationships were obtained using the following penalties: $p^{ter} = p_o^{hf} = p_e^{hp} = 1.0$, $p_e^{hf} = 0.9$, and $p_o^{hp} = 2.5$.

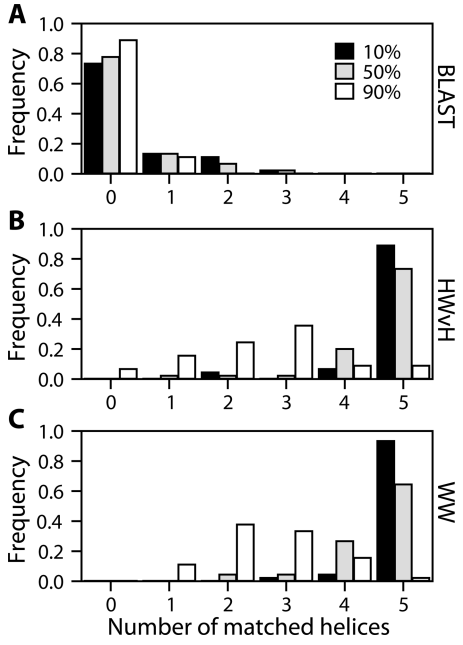

FIGURE 3: Number of TM helices correctly matched in 45 alignments of 5TM repeat units from transporters with the LeuT fold. Three different cutoffs are shown as measures of a "correct" alignment, namely, 10% (black), 50% (gray), or 90% (white) of the TM residues matched with residues in the equivalent TM domain. (A) Sequence alignments generated using BLAST. (B, C) Hydropathy profile alignments generated using AlignMe with either (B) HWvH or (C) WW hydrophobicity scales.

Other Analysis. TM helix predictions were performed using the TMHMM (*43*), HMMTOP (*44*) and OCTOPUS (*45*) web servers. Alignments with HMAP were created using all default parameters (*46*). Ends of TM domains of known structures were defined using the "Orientation of Proteins in Membranes" database (*47*), with slight modifications.

RESULTS

Comparison of AlignMe Hydropathy Profile Alignments with BLAST. We used AlignMe to align hydropathy profiles of all ten known 5TM repeat sequences against one another (see Experimental Procedures) and compared the extent to which individual TM helices were matched in each of those 45 alignments. A "correct" match was defined at three levels; i.e., 10%, 50%, or 90% of the residues in the membrane-spanning segment align against residues in the equivalent TM segment. Using BLAST, very few of the TM helices were correctly matched in those 45 alignments (Figure 3A), consistent with the fact that the expectation values of the alignments were high (mean = 1.58 and best = 2×10^{-4}). By contrast, hydropathy alignments generated using the HWvH scale correctly match all five helices in 73% of the alignments, and either four or five helices in 93% of the alignments, when defining a match as 50% of the residues being aligned (Figure 3B). Similar results were found with the WW scale (Figure 3C). These results suggest that the hydropathy-based alignments are suitable for identifying equivalent helical regions in a five TM helix protein sharing a common ancestor.

Comparison of AlignMe Family-Averaged Hydropathy Profiles with HMAP. Alignments using family-averaged profiles require the construction of a multiple-sequence alignment of

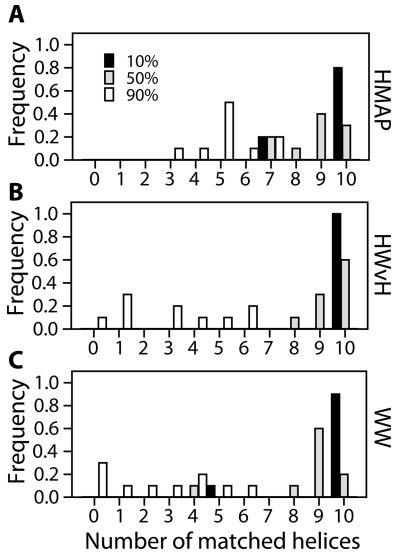

FIGURE 4: Number of core TM helices correctly matched in 10 alignments of LeuT-fold transporter protein sequences. Three different cutoffs are shown as measures of a "correct" alignment, namely, 10% (black), 50% (gray), or 90% (white) of the TM residues matched with residues in the equivalent TM domain. (A) Sequence alignments generated using HMAP. (B, C) Family-averaged hydropathy profile alignments generated using AlignMe with either (B) HWvH or (C) WW hydrophobicity scales.

homologues and averaging of the hydrophobicity values at each position. Such alignments are generally more accurate than alignments using profiles of individual proteins and are particularly useful for comparing membrane protein sequences believed to have similar structures. To test the accuracy of family-averaged profile alignments generated using AlignMe, we created alignments of all five full-length LeuT-fold sequences with one another (Supporting Information Figure S1). These alignments were compared with alignments generated with HMAP, which combines information from several sources, including secondary structure, homologous sequences, and, where possible, related structures; HMAP was previously found to be more accurate for aligning membrane protein sequences than all other methods tested (*42*), similar to its performance on water-soluble proteins (*46*). As shown in Figure 4, the AlignMe family-averaged hydropathy alignments are as good as, or even slightly better than, those generated with HMAP, depending on the hydrophobicity scale used. Therefore, family-based hydropathy profile alignment is a simple but accurate method for comparing membrane proteins of similar structure that lack significant sequence similarity.

Comparison of 5TM Repeats to One Another. We used family-averaged profile alignments to study the evolution of the 5TM repeats by comparing them within and across transporter families (Table 1). The mean PDS of alignments between the first repeat units of a given transporter family with other first repeat units (R1−R1), or alignments of second repeat units with other second repeat units (R2−R2), was higher for R2−R2 than for R1−R1 (Table 1a), suggesting that there is less variation in the first half of the fold than in the second half, over all the families. These scores were slightly lower than for alignments between first and second repeat units (R1−R2, Table 1b), particularly for alignments between R1 and R2 of the same family (Table 1b, diagonal). This suggests that the evolutionary distance between the duplicated domains is greater than between repeats in the same region of the sequence of different families, implying that all

Table 1: Scores of Family-Averaged 5TM Repeat Profile−Profile Alignments[a]

(a) Repeat Units in Same Position (R1−R1 or R2−R2)

R2/R1	APC	BCCT	NCS1	NSS	SSS	mean R1−R1
APC		0.178	0.139	0.164	0.194	
BCCT	0.251		0.181	0.229	0.193	
NCS1	0.088	0.221		0.147	0.206	0.184 ± 0.027
NSS	0.207	0.187	0.268		0.213	
SSS	0.185	0.251	0.200	0.235		
mean R2−R2			0.209 ± 0.048			**0.197 ± 0.042**

(b) Repeat Units of Different Positions (R1−R2)

R2/R1	APC	BCCT	NCS1	NSS	SSS	
APC	**0.176**	0.247	0.168	0.191	0.253	
BCCT	0.193	**0.254**	0.173	0.251	0.218	
NCS1	0.204	0.215	**0.280**	0.226	0.306	
NSS	0.191	0.211	0.201	**0.243**	0.243	
SSS	0.179	0.245	0.247	0.237	**0.250**	
mean (**diagonal**)						0.224 ± 0.035 (**0.241 ± 0.039**)

[a]PDS values are reported for family-averaged profile−profile alignments between 5TM repeat units of all transporters with the LeuT fold. (a) Comparison of repeats in the same position in the sequence, i.e., repeat 1 of one family (R1) against R1 of a different family in the top right quadrant and repeat 2 (R2) against R2 of a different family in the lower left quadrant. Mean values over each quadrant are shown, as is the mean over all values (bold). (b) Comparison between repeats in different positions in the sequence, i.e., R1 (columns) against R2 (rows). The mean includes all R1−R2 alignments. The diagonal contains scores of alignments of R1 against R2 of the same family (bold) and the average thereof (in parentheses); the top right quadrant contains alignments of, e.g., R1 of the BCCT family against R2 in the APC family, and the lower left quadrant contains alignments of, e.g., R1 of the APC family against R2 of the BCCT family.

the transporter families evolved from the same ancestral fused transporter and that there was a single duplication event.

Interestingly, the PDS of the R1−R2 alignment for the APC superfamily is lower than for the other transporter families (Table 1b). In addition, the first 5TM repeat unit in the APC transporters is more similar to R1 of other families (mean PDS = 0.168 ± 0.023) than the BCCT, NCS, NSS, and SSS family repeat units 1 are to one another (mean PDS = 0.195 ± 0.029). This is also true for the second repeat (mean PDS of 0.183 ± 0.069 and 0.227 ± 0.031, respectively). However, we note that NCS1 transporters also have a small evolutionary distance to other transporters, at least for repeat unit 1 (mean PDS of 0.168 ± 0.031), although less so for repeat unit 2 (0.194 ± 0.076). One possible interpretation of these data is that the APC repeat units have evolved the least from one another and from other families, and therefore APC proteins may be the most ancient of the transporters; i.e., they arose soonest after the gene-duplication event.

Taken together, the alignments of the LeuT-fold repeats with one another suggest that gene-duplication and fusion events occurred only once to form a larger transporter protein, possibly of the APC superfamily, before the specific mutations that led to the other LeuT-fold transporter families.

Properties of 5TM Repeat Profiles. Alignments of family-averaged profiles of the 5TM repeats reveal the defining (constant) and variable features of the LeuT fold (Figure 5). The most conserved features are the well-defined hydrophobic peaks for TMs 2, 4, and 5 and the similar lengths of the TM4−5

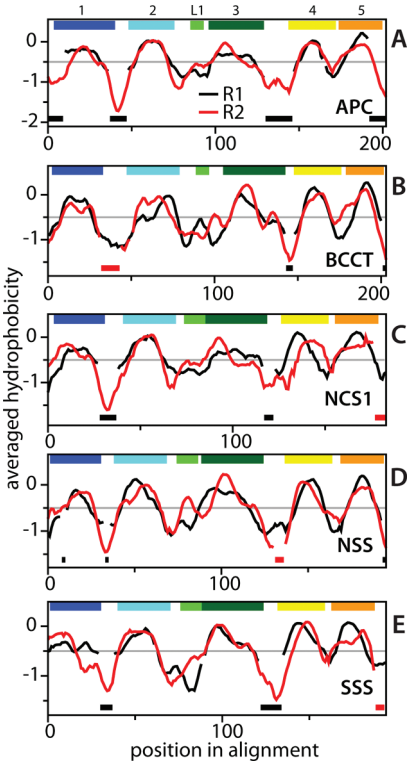

FIGURE 5: Family-averaged hydropathy profile alignments of the individual repeats of each representative LeuT-fold transporter against one another, i.e., repeat unit 1, R1 (black), against repeat unit 2, R2 (red), for the APC (A), BCCT (B), NCS1 (C), NSS (D), and SSS (E) families. The HWvH hydrophobicity scale was used here. The approximate positions of the five TM helices and loop A1/B1 are marked above the plots. Gray horizontal lines in this and all other hydropathy profile plots represent the threshold above which a position is defined as membrane-spanning for assignment of gap penalties (see Experimental Procedures). Gaps in the alignments are shown as bars underneath the profiles, in corresponding colors.

loops. The first and third TM helices of each repeat tend to have less well defined peaks, reflecting the presence of polar residues involved in substrate binding (*5*−*11*). The variability in these helices across families reflects the differences in specificity at those binding sites. The third hydrophobic peak is typically very broad, due to the long, highly tilted structure of TM3 (*10*). Preceding TM3 is a loop, L1, whose length is conserved but whose hydrophobicity varies (e.g., Figure 5D); this region typically contains a short amphipathic helix (Figure 1), and the loop is known to be critical for function in, e.g., the NSS family (*48*, *49*). Variability is found in the lengths of the loops between TMs 1 and 2 and TMs 3 and 4, although eight out of those ten insertions occur in R2 rather than in R1 (Figure 5), consistent with our observation that the first half of the fold is more conserved than the second (Table 1a).

Screening 5TM Repeat Units from LeuT-Fold Transporters against Bacterial Membrane Proteomes. In an attempt to find 5TM proteins that may share a common ancestor with the individual LeuT-fold repeat units, we screened the hydropathy profile of each of the ten 5TM repeat units against the profiles of a filtered set of protein sequences from the genomes of 20 different bacteria (see Experimental Procedures). All sequences in the top five hits for any one of the ten queries were sorted according to the sum of their ten rankings (Table 2). We then analyzed the ten highest ranked of those hits in detail, considering their profile alignments as well as the number of predicted TM helices. For reference, 28 out of 30 TM predictions for the ten known LeuT repeat sequences (i.e., using three predictors on ten sequences) say they would contain five TM helices. (The exception is the first helix of vSGLT R1, which is not identified by TMHMM or HMMTOP.) Thus, we expect a protein related to the LeuT repeat ancestor should also be predicted to contain five TM helices.

Hits (1) and (4) are both electron transport subunits of NADH:ferredoxin oxidoreductase from the RNF-NQR superfamily. TM helix predictions indicate that these proteins contain six TM helices (Table 2). Analysis of the corresponding hydropathy profile alignments reveals why 6TM helix proteins rank so high in a search for 5TM proteins. Specifically, a small amphipathic helix in loop A1 or B1 of the LeuT repeats can sometimes align to one of the membrane-spanning helices of the search hit. For example, in the alignment of hit (1) to the first repeat of BetP (Figure 6A), against which this hit was ranked highest, the third hydrophobic peak of the hit, which is large enough to span the membrane, is aligned against the small hydrophobic peak in BetP corresponding to loop A1 (Figure 1). Similar results are found for hit (6), a CDP-diacylglycerol−glycerol-3-phosphate 3-phosphatidyltransferase, for which a distinct (likely membrane-spanning) hydrophobic peak is aligned against the cytoplasmic loop B1 in BetP (Figure 6E). Thus, hits (1), (4), and (6) probably correspond to false positives.

Other hits that can be ruled out are as follows: hit (3), a putative BioY protein from the BioY biotin transporter superfamily; hit (5), an intracellular cell-division protein from the IspA superfamily; and hit (9), a multiple-antibiotic resistance protein MarC. For these three hits, the third of the six predicted TM segments has no equivalent in the aligned repeat profile (Figure 6C,D,H). Hit (7), a chromate transporter, belongs the CHR family, which is known to contain six TM segments (*50*); the fourth peak in its profile is aligned to a gap in the corresponding repeat of Mhp1 (Figure 6F). Finally, hit (8) is a LytS histidine kinase that can be excluded since it also likely contains six TM segments (Table 2) and since its profile lacks several of the conserved features of the 5TM repeat (Figure 6G; cf. Figure 5).

Hits (2), gi 258542235, and (10), gi 62389739, are sequences belonging to the DedA family of proteins, which are often also annotated as SNARE-associated proteins. TM helix predictions indicate that these hits are very likely to contain five TM helices. In addition, an alignment of gi 258542235 with the second repeat unit of BetP shows a strong similarity between their hydropathy profiles (Figure 6B). The DedA proteins are analyzed in more detail below.

Screening Family-Averaged Hydropathy Profiles of 5TM Repeat Units against Bacterial Membrane Proteomes. In the previous search the queries were hydropathy profiles of 5TM repeats. Here, we repeated the search but employing family-averaged profiles as queries to provide additional confidence in the results. As before, all sequences in the top five hits for any one of the ten queries are shown (Supporting Information Table S1), and the top ten of these were analyzed in detail.

The top hit in this search (1′) is a nucleoside recognition domain belonging to the Gate superfamily. The predictions for this protein favor five TM helices, and the family-averaged profile alignment shows reasonable matching between hydrophobic peaks, except for the first (Figure 6I). Similar results were found for hit (9′), a CapC capsule biosynthesis protein (Figure 6J). These proteins will be analyzed further below.

Table 2: Ranking of Bacterial Proteins Homologous to the LeuT-Fold Transporter Repeats, Identified Using Hydropathy Profile Alignments[a]

	AdiC		BetP		LeuT		Mhp1		vSGLT		
gi	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	TM
56476885	70	7	3	62	105	13	182	50	23	86	6
258542235	**43**	**49**	**87**	**5**	**452**	**75**	**54**	**96**	**13**	**122**	**5**
226356960	148	4	79	161	268	31	148	111	34	215	6
56478216	105	46	33	56	264	1	218	87	244	301	6
56479150	495	3	37	7	8	7	256	376	353	77	6
213155757	420	2	113	1	65	205	405	523	276	1	5-4-6
257126086	164	112	20	432	883	24	167	3	69	169	4-5-5
154249937	341	31	72	81	414	2	289	108	97	639	4-6-6
256371651	306	1	6	628	259	23	401	76	258	119	6
62389739	**440**	**44**	**2**	**74**	**248**	**106**	**352**	**606**	**196**	**54**	**4-5-5**
157737779	**16**	**76**	**141**	**2**	**1198**	**155**	**99**	**225**	**17**	**413**	**5**
219847375	365	155	1	369	682	3	582	77	88	87	6
220918164	74	324	202	188	892	223	77	2	141	286	5
32471233	10	41	198	64	1395	46	1	12	3	998	5
213157931	720	21	29	18	3	85	604	482	626	263	6
219847566	919	110	209	48	5	138	498	300	755	8	6
225872752	614	24	4	278	518	100	680	192	609	62	4-4-6
32471617	754	17	78	43	26	4	727	1052	468	3	6
154249537	549	605	41	160	152	183	581	710	974	2	6
46447431	3	1351	642	169	1546	132	96	446	81	405	4
32473095	**586**	**270**	**5**	**46**	**1387**	**200**	**551**	**1336**	**416**	**255**	**3-3-5**
75908489	5	127	731	110	2460	275	10	91	19	1695	4
53714603	552	1042	636	873	753	283	132	4	270	1139	5
225873264	1097	50	470	4	845	568	1001	457	1092	106	1
226354834	482	726	175	790	1939	41	767	5	642	146	4-5-7
46445933	559	107	390	15	2	276	252	2342	1078	791	6
62391494	769	5	138	345	1506	609	1741	243	501	201	5-6-6
226355900	408	1197	656	358	1452	191	107	1	278	1539	3-4-5
32473229	911	379	1104	3	486	1065	453	1036	575	190	5
219850286	54	287	191	164	3363	1049	305	133	5	1315	4-4-5
56477184	4	1076	569	714	1606	184	42	246	39	2926	4
213155834	19	1127	281	384	2547	452	85	185	2	2499	5
56477783	1014	787	584	1041	4	300	783	1014	953	1392	7
258542195	2	1295	454	121	3452	1173	149	545	7	1704	4-4-5
75908265	93	1208	299	577	2558	833	12	206	4	3148	5
219849097	35	2315	1308	532	2311	458	3	119	63	1939	4-5-4
226357011	1607	1511	156	1729	354	177	1141	654	2206	5	4-5-5
56475758	932	2529	629	393	2376	5	750	436	2226	382	3
53715382	31	2161	1781	316	3079	631	2	128	131	2808	4-4-5
213155583	1	859	2673	755	3143	1039	5	62	1	2989	3-5-4
75910419	2238	1350	567	1820	176	1136	1273	1061	2793	4	8
213158575	1820	1266	1696	2001	1	1521	1582	3605	1984	909	5-5-6
154250204	145	3869	3035	390	5256	2190	4	17	51	488	4-5-6

[a]Reference gi numbers (first column) of protein sequences identified using the hydropathy profile search. Each sequence was found within the top five hits for at least one of the ten 5TM repeat queries. R1 and R2 refer to repeat units 1 and 2, respectively, of the transporter whose name is indicated above. Numbers in columns 2−11 correspond to the rank, out of 11987, of the GPDS alignment score for that sequence when aligned against the indicated transporter repeat using the HWvH hydrophobicity scale. Results are sorted according to the sum of all ten rankings for each sequence hit (i.e., all values in that row). The last column contains the predicted number of TM helices, according to TMHMM, HMMTOP, and OCTOPUS (*X*-*Y*-*Z*, respectively; when all programs predicted the same number of TM helices, only one number is shown). Entries in bold font belong to the DedA family.

Among the top ten hits there were three sequences (hits 2′, 3′, and 4′) that correspond to CDP-diacylglycerol−glycerol-3-phosphate 3-phosphatidyltransferases, which were excluded from the previous search (see above). Several of the remaining hits from this second search have a low probability of containing five TM helices (Supporting Information Table S1); these include hit (6′), a protein of unknown function, as well as two proteins from the RNF-NQR superfamily (hits 5′ and 8′), which are related to hits that were excluded from the previous search results ((1) and (4); see above).

Finally, two of the top ten hits, i.e., (7′) gi 62389739 and (10′) gi 258542235 are DedA proteins, which also ranked high in the previous search.

Number and Distribution of Best Hits. Three hits from the two searches were predicted to contain five TM helices and had profiles that matched well with those of known repeats. For these DedA, Gate, and CapC proteins, we explored the diversity and distribution of their homologues using PSI-BLAST searches.

For the CapC protein (gi 32471233) only 34 homologous proteins were identified after five iterations of PSI-BLAST, and these are limited to bacteria. Similarly, for the Gate nucleoside recognition domain (gi 220919305), PSI-BLAST identified < 300 related sequences, around half of which are fused to SpmA domains. By contrast, a PSI-BLAST search with the DedA sequence gi 258542235 returned 3225 sequences of DedA domain-containing proteins. The existence of such a large number of

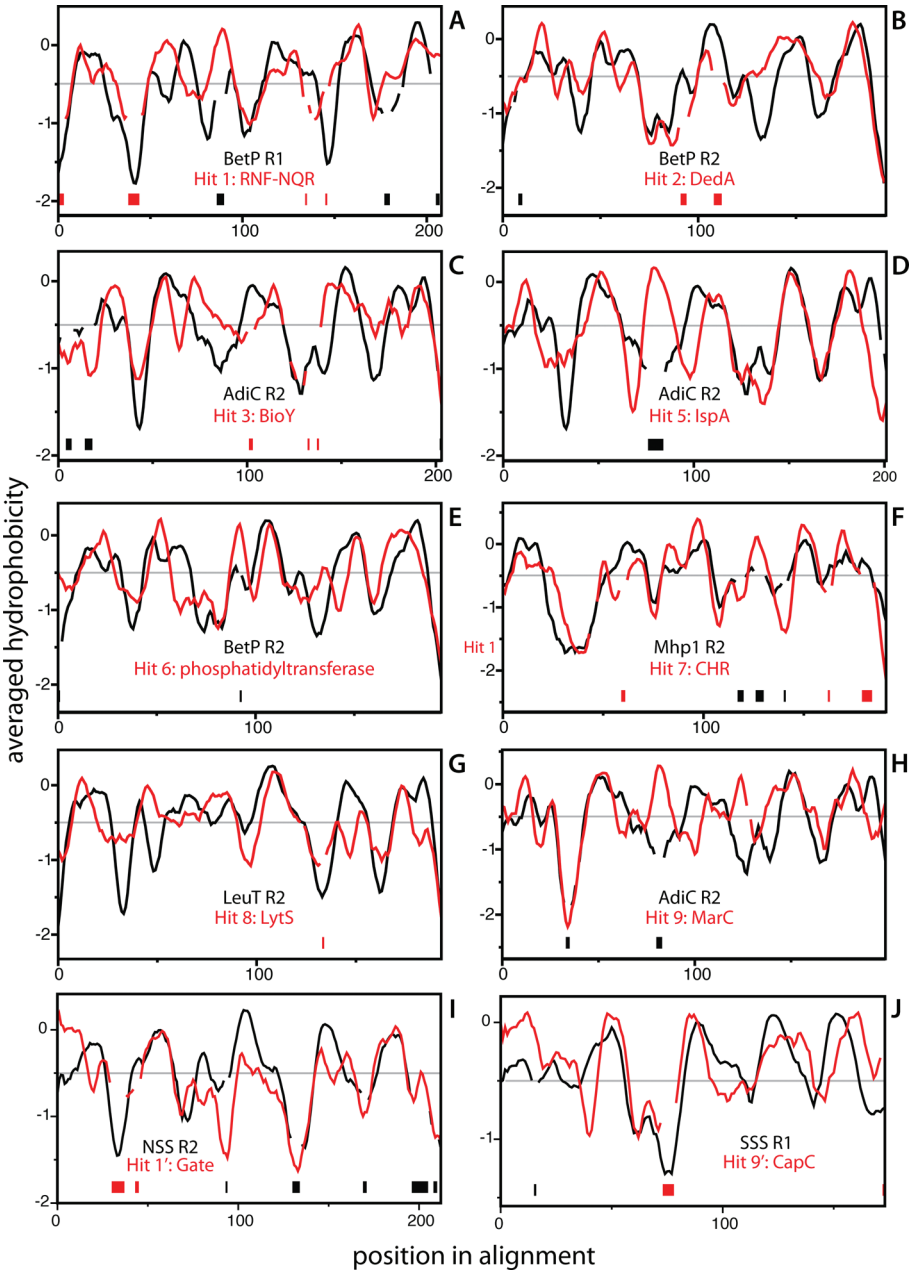

FIGURE 6: Hydropathy profile alignments of search hits against 5TM repeats of LeuT fold, generated using the HWvH hydrophobicity scale. (A−H) Results obtained using individual hydropathy profiles as queries. Aligned profiles in this case are window-averaged. (A) R1 of BetP and gi 56476885. (B) R2 of BetP and gi 258542235, a DedA domain. (C) R2 of AdiC and gi 226356960. (D) R2 of AdiC and gi 56479150. (E) R2 of BetP and gi 213155757. (F) R2 of Mhp1 and gi 257126086. (G) R2 of LeuT and gi 154249937. (H) R2 of AdiC and gi 256371651. (I, J) Results obtained using family-averaged (and window-averaged) hydropathy profiles as queries. (I) R2 of the NSS family and gi 220919305. (J) R1 of the SSS family and gi 32471233. Profiles are colored black for the query and red for the search hit.

related proteins suggests that DedA domains belong to an ancient family. Moreover, sequences of DedA domains are found in genomes from all kingdoms of life, from bacteria to humans, and are often found in duplicate or triplicate (see below). Thus, there is a strong possibility that larger proteins evolved from a DedA-like protein, making this the most promising candidate for sharing a common ancestor to the LeuT-fold repeat.

Ranking of DedA Domains Compared to the Ranking of LeuT Repeats. To further assess the significance of the similarity between DedA domains and the LeuT repeats, we compared the ranks of the alignments between the hydropathy profiles of DedA and the 5TM queries against the ranks of alignments between pairs of 5TM repeat units. Specifically, we doped the filtered *C. glutamicum* genome with the sequences of the ten 5TM repeat units (as used for gap penalty optimization; see Experimental Procedures). We then screened each of the ten 5TM repeat hydropathy profiles against this set, using both WW and HWvH hydrophobicity scales. The results were sorted according to the sum of the ranks for all ten searches (Table 3). The ten sequences of known structure have high rankings: they are distributed among the first 17 hits (out of 220) and occupy the first five positions. This is expected since the gap-penalty optimization was designed to minimize the sum of their ranks. The lowest ranking of the ten repeats (at position 17) is LeuT R1, which contains a relatively long insertion between TMs 3 and 4, emphasizing how specific differences between the repeats can make their similarities difficult to detect, even with this sensitive approach. Importantly, however, the sequence gi 19552076, a

Table 3: Highest Ranking Sequences in a Search of 5TM Repeats against the Membrane Proteome of *C. glutamicum* Doped with the Same 5TM Sequences[a]

final rank	sequence	final rank	sequence
1	AdiC R1	13	Mhp1 R1
2	BetP R1	14	Mhp1 R2
3	BetP R2	15	gi 19552684
4	LeuT R2	16	gi 19552036
5	vSGLT R2	17	LeuT R1
6	**gi 19552076**	18	gi 19551526
7	AdiC R2	19	gi 19551396
8	gi 19553210	20	gi 19554088
9	vSGLT R1	21	gi 19553655
10	gi 19553853	22	gi 19551396
11	gi 19551455	23	gi 23308908
12	gi 19551257		

[a]Results of hydropathy profile alignments of 5TM repeats of known structure transporters against a set that includes 210 sequences from *C. glutamicum* plus 10 sequences of 5TM repeats themselves. Thus, final rankings are out of 220 positions. R1 and R2 refer to repeat units 1 and 2, respectively. Entries in bold font belong to the DedA family.

DedA protein that was not optimized against, was ranked sixth. These results show that the profile of the DedA proteins shares a higher similarity with the 5TM repeats than some of the 5TM repeats share with one another.

Comparison of DedA Domains to LeuT-Fold Transporters. We next examined the similarities between the DedA and LeuT repeat structures by aligning a DedA family-averaged profile with those for all ten individual 5TM repeats (Supporting Information Figure S2). The resultant alignments show conservation and variability in the same regions as observed in LeuT-fold repeats (Figure 5). For example, the canonical hydrophobic peaks for TMs 2, 4, and 5 are also found in the DedA profile, as are the unusual features of TM3, reflecting the presence of conserved polar residues in this region of the DedA domains. The variable TM1−2 loop in the repeats aligns to a small hydrophobic peak in DedA.

We also attempted to align the DedA domain family-averaged profiles with those of the full-length transporters. For the BCCT transporters, against which the DedA sequences ranked particularly high (Table 2 and Supporting Information Table S1), the DedA family profile was perfectly aligned against R1 of the BCCT proteins, suggesting a very strong similarity between them (Figure 7A). When we then manually removed a region corresponding to the first three TM helices from the BCCT multiple-sequence alignment, i.e., to deliberately truncate the profile of the first repeat unit, the DedA profile "jumps" and aligns to the second repeat instead (Figure 7B), indicating similarity also between the DedA profile and the second 5TM repeat unit, as opposed to a random matching to any five sequential TM domains of the transporter.

However, in family-averaged alignments against the other LeuT-fold transporter families, the DedA helices do not align against the 5TM repeats in the full-length sequences (data not shown).

Orientation of DedA Proteins in the Membrane. We selected 32 DedA proteins from our data set of bacterial genomes whose TM predictions were very consistent, i.e., those for which all three methods predicted five TM helices (15 sequences), or that did not contain significant hydrophobic insertions compared to those 15. The average sequence identity of these 32 sequences to one another is 20.4 ± 7.0%. We then calculated the number of positively charged residues in both termini and in the loops, guided by the TM predictions, but making manual adjustments to ensure all positive residues at the interfacial regions were included.

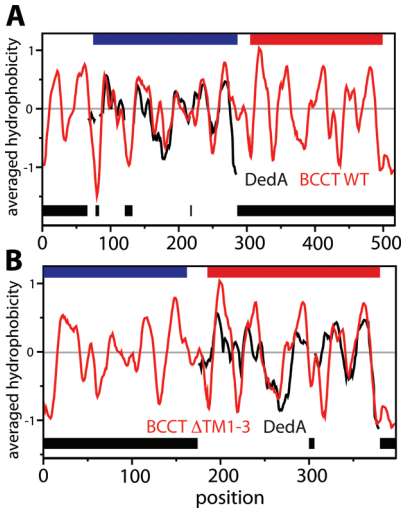

FIGURE 7: Hydropathy profile alignments of DedA (black) and BCCT (red) family proteins generated using the WW hydrophobicity profile. Either full-length sequences were aligned (A) or a region corresponding to the first three TM helices was manually removed from the multiple-sequence alignment of the BCCT proteins and a truncated hydropathy profile was aligned (B). The locations of the first and second repeat units of BCCT are indicated above the alignment using blue and red bars, respectively.

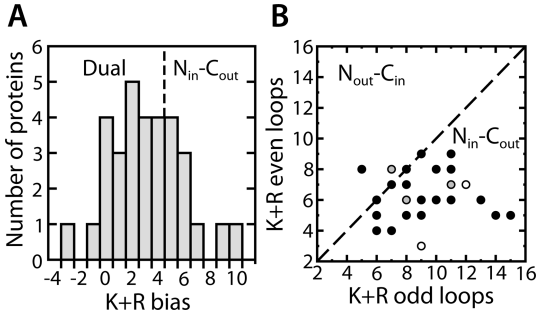

FIGURE 8: Positive charge (K+R) bias in 32 DedA proteins. (A) Distribution of positive charge bias. Positive charge differences (above ~4) correspond to probable $N_{in}-C_{out}$ orientations, while those below ~4 are more likely to have dual topologies. (B) Number of positive charges in the odd-numbered loops plotted against the number in the even-numbered loops. Points representing more than one protein are in gray (two proteins) or white (three proteins). The dashed line indicates where the charge bias is zero.

The lysine plus arginine (K+R) positive-charge bias of the DedA proteins varies from −3 to +10 (Figure 8). More than half, i.e., 22 of the proteins, possess a very low positive-charge bias, i.e., between −4 and +4. These proteins may be dual topology, i.e., insert in both orientations in the membrane (*18*, *51*, *52*). The remaining 10 sequences have a strong K+R bias (≥4) corresponding to an orientation with the N-terminal domain inside the cell and the C-terminal domain outside, i.e., $N_{in}-C_{out}$. No proteins were strongly predicted to have a $N_{out}-C_{in}$ orientation.

Within a given bacterial genome, we often find two or more DedA proteins, although none are immediately adjacent on the chromosome. In some cases, such as in *E. coli*, all copies may have dual topologies, as suggested by K+R biases of 0, +2, +3, and +3. More commonly, as in, e.g., *A. pasteurianus*, the copies have different predicted orientations (with a K+R bias of −1, +2, and +6), whereas for *A. capsulatum*, for example, all

three are predicted to have a $N_{in}-C_{out}$ topology, with K+R biases of +5, +7, and +9.

Thus, DedA proteins might still function as homodimers (in the case of the dual-topology sequences), but if they did transition through opposite-topology heterodimers, they apparently did not retain that functionality.

DISCUSSION

In this study, we have shown that hydropathy profile alignments generated with AlignMe are suitable for searching for sequence-unrelated membrane proteins of similar structure (Figure 3). Alignments generated using family-averaged hydropathy profiles are even more accurate and are useful for detailed comparison of related proteins (Figures 4, 5, and 7 Supporting Information Figures S1 and S2). These findings build on previous work by Lolkema and Slotboom, who were able to correctly predict structural similarities between families of secondary transporters using hydropathy profile alignments, including a relationship between the APC, SSS, and NSS families (*30*−*32*).

We used the hydropathy profile alignment methodology to study the evolution of the LeuT fold. First, comparison of the repeats in the known structures suggested that the first half of the core 10TM domain is more conserved than the second (Table 1) and that each half is generally more conserved over evolution (in different families) than the two halves are to one another, even in the same protein. These results imply that gene-duplication and fusion events occurred only once to form a larger transporter protein, from which the current families evolved. In addition, comparison of the different families to one another suggests that the first repeat of APC transporters has evolved the least from that ancient individual domain.

We also used AlignMe to screen 5TM repeat profiles against membrane proteomes of bacteria. These searches identified DedA domains as being likely to share a common ancestor to the individual repeats. We note that in those search results, the ranking of a given hit was sometimes very sensitive to the query sequence used (Table 2 and Supporting Information Table S1). The most extreme example is sequence gi 154250204, which ranks fourth against the first repeat unit of Mhp1 but only 5256th against the first repeat unit of LeuT. The use of ten different queries minimizes the effect of these variations, which may be due to the limitations of our scoring scheme but certainly also reflect true differences between hydropathy profiles of repeats from different transporters, such as the presence of a relatively long loop between the third and fourth TM helices of LeuT (in R1; see Figure 5). Such differences also explain the 17th place ranking of LeuT R1 in the test ranking of all the repeats (Table 3). In the light of such variability between known structures, it is also clear that having a fairly high number of false positives (as found in our search) is preferable to the alternative, i.e., missing an interesting sequence, especially since such false positives could be screened out using other criteria. Additional support for the proposal that DedA proteins are good candidates comes from predictions that they contain five TM helices (Table 2). In addition, their hydropathy profiles are sometimes more similar to those of the LeuT-fold repeats than the profiles of 5TM repeat units are to one another (Table 3). Moreover, alignments of family-averaged hydropathy profiles of DedA and 5TM repeats reveal that they share similar conserved features (Supporting Information Figure S2 and Figure 7). Based on the rankings of the search hits (Table 2 and Supporting Information Table S1) and on alignments to the full transporter profiles (Figure 7), the DedA domains appear to be most similar to the BCCT family, although this observation apparently contradicts the analysis of the repeats, which suggested that the APC repeat 1 is the more ancient domain.

Unfortunately, the functional role of DedA proteins is poorly understood. In one study, DedA-related proteins expressed in *E. coli* appear to facilitate 5-hydroxybutyrate utilization; the mechanism of this effect is not known, although it is not unreasonable that it requires an uptake step (*53*). Interestingly, *dedA* gene knockout mutants of *Ralstonia metallidurans* CH34 are more resistant to selenite toxicity and less efficient at removing selenite from the environment, indicating that DedA proteins might be involved in selenite uptake, either directly as transporters or as helpers thereof (*54*). In other studies, simultaneous knockout of *yghB* and *yqjA*, two *dedA* genes encoding proteins with 61% identity, resulted in a temperature-sensitive phenotype of *E. coli* with altered levels of phospholipids, which affects cell division (*55*). This phenotype appears to derive from a deficiency in the export of periplasmic amidases through the twin arginine transport (Tat) pathway (*56*), although the role of the DedA domains therein remains unclear. Importantly, the fact that both proteins are required for the phenotype suggests that they are either redundant or that they function interchangeably as hetero- or homodimers (*55*). Indeed, the presence of two DedA proteins of the same subtype (so-called group B) in several different organisms, including in *E. coli*, is indicative that *dedA* underwent gene duplication (*54*). In fact, we find that multiple copies of *dedA* genes often exist in the same genome, with 14 of the 20 bacterial genomes used in our search containing at least two DedA proteins. Interestingly, many DedA proteins have a weak positive-charge bias, suggesting that they may adopt dual topologies, although the distribution of charge bias is broad and others are likely to insert in a $N_{in}-C_{out}$ orientation (Figure 8).

The search procedure used here may be used for α-helical membrane proteins other than LeuT-fold transporters. It is likely to be particularly useful in determining whether a protein of interest is related to a protein of known structure or for identification of structurally similar domains within a protein chain, such as repeat units that share no detectable sequence similarity. An advantage of hydropathy alignments is that they provide a simple visual representation of the relationship between the two proteins that can be interpreted rapidly and intuitively.

Finally, in spite of the obvious value of hydropathy profile alignments, it should be noted that, as with any alignment procedure, the selection of parameters such as gap penalties, weights for scoring function, and averaging-window size and shape can be arbitrary. In addition, the results may be sensitive to the choice of hydrophobicity scale (Figures 3 and 4). As we have shown here, it is possible to optimize some of those parameters using a systematic scanning of gap penalty values. However, further improvements to alignment accuracy, or more general solutions, may be obtained using optimization over a wider range of structural families, which we will address in future studies.

The bioinformatic analysis presented here suggests that there are similarities in the sequence origins of DedA domains and transporters of the LeuT fold. Furthermore, previous studies hint at similarities in function and oligomeric state, as well as the existence of a gene-duplication event to connect them. Clearly, experimental validation in the form of either structural or biochemical data will be required to confirm or disprove the proposed relationship. If such a relationship can be demonstrated, however, DedA domains can be expected to provide important

clues as to the origins of symmetry-based alternating-access mechanisms of secondary transport.

ACKNOWLEDGMENT

We thank Gary Rudnick and Barry Honig for discussions about the idea for this project and José Faraldo-Gómez for constructive criticism of the manuscript.

SUPPORTING INFORMATION AVAILABLE

The tabulated results for the family-averaged profile search and family-averaged profile alignments of LeuT-fold transporters and of DedA proteins against individual repeats. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Sobczak, M., and Lolkema, J. S. (2005) Structural and mechanistic diversity of secondary transporters. *Curr. Opin. Microbiol. 8*, 161–167.
2. Saier, M. H., Tran, C. V., and Barabote, R. D. (2006) TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res. 34*, D181–D186.
3. White, S. H. (2009) Biophysical dissection of membrane proteins. *Nature 459*, 344–346.
4. Krishnamurthy, H., Piscitelli, C. L., and Gouaux, E. (2009) Unlocking the molecular secrets of sodium-coupled transporters. *Nature 459*, 347–355.
5. Fang, Y. L., Jayaram, H., Shane, T., Kolmakova-Partensky, L., Wu, F., Williams, C., Xiong, Y., and Miller, C. (2009) Structure of a prokaryotic virtual proton pump at 3.2 Å resolution. *Nature 460*, 1040–1043.
6. Shaffer, P. L., Goehring, A., Shankaranarayanan, A., and Gouaux, E. (2009) Structure and mechanism of a Na^+-independent amino acid transporter. *Science 325*, 1010–1014.
7. Ressl, S., Terwisscha van Scheltinga, A. C., Vonrhein, C., Ott, V., and Ziegler, C. (2009) Molecular basis of transport and regulation in the Na^+/betaine symporter BetP. *Nature 458*, 47–52.
8. Tang, L., Bai, L., Wang, W.-h., and Jiang, T. (2010) Crystal structure of the carnitine transporter and insights into the antiport mechanism. *Nat. Struct. Mol. Biol. 17*, 492–496.
9. Shimamura, T., Weyand, S., Beckstein, O., Rutherford, N. G., Hadden, J. M., Sharples, D., Sansom, M. S. P., Iwata, S., Henderson, P. J. F., and Cameron, A. D. (2010) Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1. *Science 328*, 470–473.
10. Yamashita, A., Singh, S. K., Kawate, T., Jin, Y., and Gouaux, E. (2005) Crystal structure of a bacterial homologue of Na^+/Cl^--dependent neurotransmitter transporters. *Nature 437*, 215–223.
11. Faham, S., Watanabe, A., Besserer, G. M., Cascio, D., Specht, A., Hirayama, B. A., Wright, E. M., and Abramson, J. (2008) The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na^+/sugar symport. *Science 321*, 810–814.
12. Forrest, L. R., Zhang, Y. W., Jacobs, M. T., Gesmonde, J., Xie, L., Honig, B. H., and Rudnick, G. (2008) Mechanism for alternating access in neurotransmitter transporters. *Proc. Natl. Acad. Sci. U.S.A. 105*, 10338–10343.
13. Forrest, L. R., and Rudnick, G. (2009) The rocking bundle: A mechanism for ion-coupled solute flux by symmetrical transporters. *Physiology 24*, 377–386.
14. Saier, M. H. (2003) Tracing pathways of transport protein evolution. *Mol. Microbiol. 48*, 1145–1156.
15. Lolkema, J. S., Dobrowolski, A., and Slotboom, D. J. (2008) Evolution of antiparallel two-domain membrane proteins: tracing multiple gene duplication events in the DUF606 family. *J. Mol. Biol. 378*, 596–606.
16. Rapp, M., Seppala, S., Granseth, E., and von Heijne, G. (2007) Emulating membrane protein evolution by rational design. *Nat. Struct. Mol. Biol. 315*, 1282–1284.
17. Poolman, B., Geertsma, E. R., and Slotboom, D.-J. (2007) A missing link in membrane protein evolution. *Science 315*, 1229–1231.
18. Rapp, M., Granseth, E., Seppala, S., and von Heijne, G. (2006) Identification and evolution of dual-topology membrane proteins. *Nat. Struct. Mol. Biol. 13*, 112–116.
19. Bowie, J. U. (2006) Flip-flopping membrane proteins. *Nat. Struct. Mol. Biol. 13*, 94–96.
20. Huang, Y. F., Lemieux, M. J., Song, J. M., Auer, M., and Wang, D. N. (2003) Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science 301*, 616–620.
21. Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R., and Iwata, S. (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science 301*, 610–615.
22. Fu, D. X., Libson, A., Miercke, L. J. W., Weitzman, C., Nollert, P., Krucinski, J., and Stroud, R. M. (2000) Structure of a glycerol-conducting channel and the basis for its selectivity. *Science 290*, 481–486.
23. Pao, S. S., Paulsen, I. T., and Saier, M. H., Jr. (1998) Major facilitator superfamily. *Microbiol. Mol. Biol. Rev. 62*, 1–34.
24. Park, J. H., and Saier, M. H., Jr. (1996) Phylogenetic characterization of the MIP family of transmembrane channel proteins. *J. Membr. Biol. 153*, 171–180.
25. Hediger, M., Romero, M., Peng, J.-B., Rolfs, A., Takanaga, H., and Bruford, E. (2004) The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins. *Pflueger's Arch. Eur. J. Physiol. 447*, 465–468.
26. Schlessinger, A., Matsson, P., Shima, J. E., Pieper, U., Yee, S. W., Kelly, L., Apeltsin, L., Stroud, R. M., Ferrin, T. E., Giacomini, K. M., and Sali, A. (2010) Comparison of human solute carriers. *Protein Sci. 19*, 412–428.
27. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res. 25*, 3389–3402.
28. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157*, 105–132.
29. Clements, J. D., and Martin, R. E. (2002) Identification of novel membrane proteins by searching for patterns in hydropathy profiles. *Eur. J. Biochem. 269*, 2101–2107.
30. Lolkema, J. S., and Slotboom, D. J. (1998) Hydropathy profile alignment: a tool to search for structural homologues of membrane proteins. *FEMS Microbiol. Rev. 22*, 305–322.
31. Lolkema, J. S., and Slotboom, D. J. (1998) Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Mol. Membr. Biol. 15*, 33–42.
32. Lolkema, J. S., and Slotboom, D. J. (2008) The major amino acid transporter superfamily has a similar core structure as Na^+-galactose and Na^+-leucine transporters. *Mol. Membr. Biol. 25*, 567–570.
33. Wimley, W. C., and White, S. H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol. 3*, 842–848.
34. Hessa, T., Meindl-Beinker, N. M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S. H., and von Heijne, G. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature 450*, 1026–1030.
35. Koehler, J., Woetzel, N., Staritzbichler, R., Sanders, C. R., and Meiler, J. (2009) A unified hydrophobicity scale for multiscan membrane proteins. *Proteins 76*, 13–29.
36. Thompson, J. D., Higgins, D. G., and 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. *Nucleic Acids Res. 22*, 4673–4680.
37. Huang, X. Q. (1994) On global sequence alignment. *Comput. Appl. Biosci. 10*, 227–235.
38. Huang, X. Q., and Zhang, J. H. (1996) Methods for comparing a DNA sequence with a protein sequence. *Comput. Appl. Biosci. 12*, 497–506.
39. Weyand, S., Shimamura, T., Yajima, S., Suzuki, S., Mirza, O., Krusong, K., Carpenter, E. P., Rutherford, N. G., Hadden, J. M., O'Reilly, J., Ma, P., Saidijam, M., Patching, S. G., Hope, R. J., Norbertczak, H. T., Roach, P. C. J., Iwata, S., Henderson, P. J. F., and Cameron, A. D. (2008) Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science 322*, 709–713.
40. Altschul, S. F., and Koonin, E. V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci. 23*, 444–447.
41. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res. 32*, 1792–1797.
42. Forrest, L. R., Tang, C. L., and Honig, B. (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J. 91*, 508–517.
43. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol. 305*, 567–580.
44. Tusnady, G. E., and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics 17*, 849–850.

45. Viklund, H., and Elofsson, A. (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics 24*, 1662–1668.
46. Tang, C. L., Xie, L., Koh, I. Y. Y., Posy, S., Alexov, E., and Honig, B. (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. *J. Mol. Biol. 334*, 1043–1062.
47. Lomize, M. A., Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I. (2006) OPM: Orientations of Proteins in Membranes database. *Bioinformatics 22*, 623–625.
48. Norregaard, L., Loland, C. J., and Gether, U. (2003) Evidence for distinct sodium-, dopamine-, and cocaine-dependent conformational changes in transmembrane segments 7 and 8 of the dopamine transporter. *J. Biol. Chem. 278*, 30587–30596.
49. Mitchell, S. M., Lee, E., Garcia, M. L., and Stephan, M. M. (2004) Structure and function of extracellular loop 4 of the serotonin transporter as revealed by cysteine-scanning mutagenesis. *J. Biol. Chem. 279*, 24089–24099.
50. Nies, D. H., Koch, S., Wachi, S., Peitzsch, N., and Saier, M. H. (1998) CHR, a novel family of prokaryotic proton motive force-driven transporters probably containing chromate/sulfate antiporters. *J. Bacteriol. 180*, 5799–5802.
51. Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S. H., and von Heijne, G. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature 433*, 377–381.
52. Kolbusz, M. A., ter Horst, R., Slotboom, D. J., and Lolkema, J. S. (2010) Orientation of small multidrug resistance transporter subunits in the membrane: correlation with the positive-inside rule. *J. Mol. Biol. 402*, 127–138.
53. Henne, A., Daniel, R., Schmitz, R. A., and Gottschalk, G. (1999) Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl. Environ. Microbiol. 65*, 3901–3907.
54. Ledgham, F., Quest, B., Vallaeys, T., Mergeay, M., and Coves, J. (2005) A probable link between the DedA protein and resistance to selenite. *Res. Microbiol. 156*, 367–374.
55. Thompkins, K., Chattopadhyay, B., Xiao, Y., Henk, M. C., and Doerrler, W. T. (2008) Temperature sensitivity and cell division defects in an *Escherichia coli* strain with mutations in yghB and yqj4, encoding related and conserved inner membrane proteins. *J. Bacteriol. 190*, 4489–4500.
56. Sikdar, R., and Doerrler, W. T. (2010) Inefficient Tat-dependent export of periplasmic amidases in an *Escherichia coli* strain with mutations in two DedA family genes. *J. Bacteriol. 192*, 807–818.
57. Petrey, D., Xiang, Z. X., Tang, C. L., Xie, L., Gimpelev, M., Mitros, T., Soto, C. S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I. Y. Y., Alexov, E., and Honig, B. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins 53*, 430–435.