Article No. jmbi.1999.2630 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 287, 695–715

JMB

Alignment and Structure Prediction of Divergent Protein Families: Periplasmic and Outer Membrane Proteins of Bacterial Efflux Pumps

Jason M. Johnson and George M. Church*

Graduate Program in Biophysics and Department of Genetics, Harvard Medical School, 200 Longwood Ave Boston, MA 02115, USA

**Corresponding author*

Broad-specificity efflux pumps have been implicated in multidrug-resistant strains of *Pseudomonas aeruginosa* and other Gram-negative bacteria. Most Gram-negative pumps of clinical relevance have three components, an inner membrane transporter, an outer membrane channel protein, and a periplasmic protein, which together coordinate efflux from the cytoplasmic membrane across the outer membrane through an unknown mechanism. The periplasmic efflux proteins (PEPs) and outer membrane efflux proteins (OEPs) are not obviously related to proteins of known structure, and understanding the structure and function of these proteins has been hindered by the difficulty of obtaining reasonable multiple alignments. We present a general strategy for the alignment and structure prediction of protein families with low mutual sequence similarity using the PEP and OEP families as detailed examples. Gibbs sampling, hidden Markov models, and other analysis techniques were used to locate motifs, generate multiple alignments, and assign PEP or OEP function to hypothetical proteins in several species. We also developed an automated procedure which combines multiple alignments with structure prediction algorithms in order to identify conserved structural features in protein families. This process was used to identify a probable α-helical hairpin in the PEP family and was applied to the detection of transmembrane β-strands in OEPs. We also show that all OEPs contain a large tandem duplication, and demonstrate that the OEP family is unlikely to adopt a porin fold, in contrast to previous predictions.

Keywords: motif detection; multidrug resistance; structure prediction; circular dichroism; multiple alignment

Introduction

The emergence of drug-resistant strains of bacteria is a significant and growing human health problem. Active efflux pumps with broad specificity are involved in the intrinsic and acquired resistance of *Pseudomonas aeruginosa*, *Escherichia coli*, and other pathogens (reviewed by Nikaido, 1998). The efflux pumps known to contribute to clinically relevant resistance in Gram-negative bacteria share a common, three-component organization. The inner membrane components of these pumps are translocases of three families: resistance-nodulation-division (RND; Saier *et al*., 1994), major facilitator (Griffith *et al*., 1992; Marger & Saier, 1993), and the family of ATP-binding cassette (ABC) transporters (reviewed by Binet *et al*., 1997). The inner membrane transporters function with periplasmic accessory proteins, sometimes called ''membrane fusion proteins'' (Dinh *et al*., 1994), and outer membrane proteins to facilitate *sec*-independent, direct passage from the cytoplasm into the external medium of a variety of substrates, including proteins, oligosaccharides, small molecules, and large cations (Binet & Wandersman, 1995; Dinh *et al*., 1994; Létoffé *et al*., 1996). Here, we shall refer to the components of these tripartite efflux pumps simply as inner membrane efflux proteins (IEPs), periplasmic efflux proteins (PEPs),

Abbreviations used: PEP, periplasmic efflux protein; OEP, outer membrane efflux protein; IEP, inner membrane efflux protein; MAPF, multiple alignment of protein features; TFE, trifluoroethanol; TM, transmembrane; NR, non-redundant; HMM, hidden Markov model.

E-mail address of the corresponding author: church@salt2.med.harvard.edu

0022-2836/99/130695–21 $30.00/0

and outer membrane efflux proteins (OEPs). Understanding the mechanism of these export complexes is the first step towards identifying structural targets for drug design and eventually overcoming this form of multidrug resistance.

The periplasmic efflux protein provides the connection between the inner and outer membrane components of the efflux pump. Many PEPs, such as the hemolysin translocator accessory protein HlyD (Felmlee *et al*., 1985), have a stretch of hydrophobic residues near the N terminus believed to span the cytoplasmic membrane (Schülein *et al*., 1992). Others, like AcrA and AcrE of *E. coli* (Ma *et al*., 1993) and MexA of *P. aeruginosa* (Poole *et al*., 1993) have a lipoprotein modification signal and are likely to insert in the cytoplasmic membrane *via* an attached lipid moiety (Dinh *et al*., 1994; Seiffer *et al*., 1993). PEPs appear to form trimers (Thanabalu *et al*., 1998), and have been shown to interact specifically with OEPs (Akatsuka *et al*., 1997; Binet & Wandersman, 1995; Hwang *et al*., 1997; Létoffé *et al*., 1996).

The outer membrane efflux proteins, reviewed by Paulsen *et al*. (1997), are often found in the same gene clusters as the IEPs and PEPs. Some OEP genes, however, such as *tolC* of *E. coli* (Morona *et al*., 1983), are not in the same operon as PEPs and IEPs and may associate with several different IEP-PEP complexes (Nikaido, 1998). Perhaps the most widely studied OEP, TolC is a multifunctional protein involved in organic solvent tolerance, export of α-hemolysin, and uptake of colicin (Aono *et al*., 1998; Fath *et al*., 1991; Wandersman & Delepelaire, 1990; Webster, 1991). TolC appears to interact directly with the periplasmic efflux protein HlyD, but only in the presence of the IEP (Schlör *et al*., 1997). The assembled structure of TolC is also believed to be trimeric (Koronakis *et al*., 1997), and the protein forms oligomeric ion-permeable pores in reconstituted bilayers (Benz *et al*., 1993). The TolC family of OEPs has recently been predicted to adopt a β-barrel structure similar to that of outer membrane porins (Koronakis *et al*., 1997; Paulsen *et al*., 1997).

The poor sequence conservation of the PEP and OEP families has largely prevented prior functional interpretation of their sequence data. For instance, in the only previous systematic analysis of OEPs, no common sequence motif could be found for the family (Paulsen *et al*., 1997). Several different algorithms designed to locate subtle sequence signals (Lawrence *et al*., 1993; Neuwald & Green, 1994; Neuwald *et al*., 1995) were used to discover motifs common to each family, which were used in turn to guide the creation of larger protein alignments using hidden Markov models (HMMs; Eddy, 1996) and PSI-BLAST (Altschul *et al*., 1997). An accurate sequence alignment for a diverse family of proteins is a powerful tool for functional and structural analysis, since conserved residues are most likely to be important to the common structure and function. To maximize the utility of the PEP and OEP alignments, we developed an automated procedure which identifies conserved structural or other sequence properties given an input multiple alignment. We used the MAPF (multiple alignment of protein features) program in coordination with a variety of algorithms to predict the locations of transmembrane β-strands in the OEPs and in families of porins as controls. MAPF was also employed in three-dimensional fold prediction of a 150-residue PEP domain which may include an antiparallel coiled coil. We suggest that this helical hairpin may be involved in the formation of a multimeric assembly of PEP and OEP helices in the periplasm. HMM representations of PEP and OEP alignments were used to identify new efflux pumps in *Aquifex aeolicus* and *Synechocystis*, and a potential PEP homologue in *Bacillus subtilis*, demonstrating that this efflux mechanism is not limited to Gram-negative bacteria. The strategy and analysis techniques presented here are applicable to other protein families whose sequences have diverged into the twilight zone of sequence identity or for which no homologue with known structure is available.

Results and Discussion

Periplasmic efflux proteins

Identification of a conserved coiled-coil probability pattern and a tandemly repeated motif in PEPs

Although relatively few coiled-coil proteins have been discovered in prokaryotes, coiled coils are believed to be present in some periplasmic protein domains (Engel *et al*., 1992; McLachlan, 1978; Scott *et al*., 1993). Thus, the occurrence of coiled coils in the PEP family (previously predicted for several PEPs; Pimenta *et al*., 1996) would not be surprising. We systematically examined all known PEPs for regions likely to form coiled coils with the COILS algorithm (Lupas *et al*., 1991) and found that almost all have a high probability of containing coiled coils. The typical PEP coiled-coil signature is centrally located in the primary sequence and has 2-fold symmetry: two regions of high coiled-coil probability of approximately equal length (four to five heptad repeats) separated by a gap of five to ten residues (Figure 1).

To find other sequence patterns common to PEPs, we used an iterative motif-detection and refinement procedure employing Gibbs sampling (Lawrence *et al*., 1993; Neuwald *et al*., 1995) and depth-first search (Neuwald & Green, 1994) strategies (see Materials and Methods). This procedure identified a single, high-scoring motif occurring in two copies (M_N and M_C) in all PEPs, separated by 60-300 residues. Intriguingly, M_N and M_C are typically positioned at each end of the coiled-coil pattern discussed above, within a few residues of the region of high coiled-coil probability.

The procedure also identified two copies of the M_N/M_C motif in biotin carboxyl carrier proteins and in the lipoyl domains of 2-oxo acid dehydro-

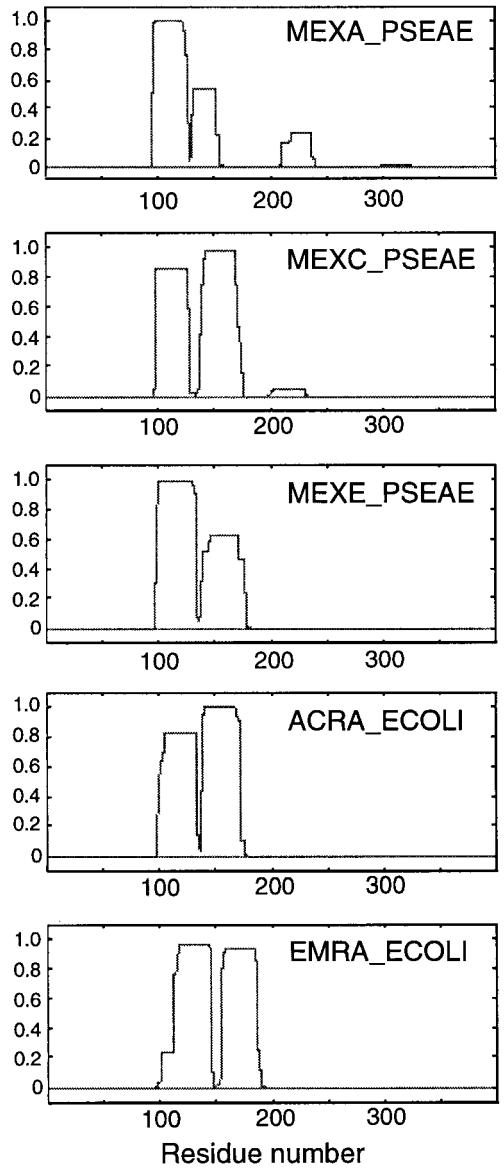

Figure 1. The probability of forming coiled coils as a function of residue number is shown for five PEPs implicated in multidrug resistance in *P. aeruginosa* and *E. coli*. The graphs were generated by the COILS program with a window of 28 residues (Lupas *et al*., 1991). A 21-residue window is used for MexA, which more clearly shows the central gap, concurrent with a break in the heptad register.

genases (reviewed by Berg & de Kok, 1997). These protein domains, in eukaryotes and bacteria, transfer a covalently attached lipoyl or biotinyl moiety between enzymatic components. Blocks of similarity between the lipoyl/biotinyl domains and some PEPs have been reported previously (Neuwald *et al*., 1997). The method presented here confirms the homology between the families, but also reveals that the homologous region represents two copies of a single motif, present in both the PEP and lipoyl/biotinyl families. This is consistent with the previous observation of weak sequence similarity between the N and C-terminal halves of the lipoyl domain (Spencer *et al*., 1984). An alignment of these repeated segments including gaps is shown in Figure 2.

Structure of the PEP motif

The structures of several lipoyl and related domains have been solved by NMR and X-ray crystallography, including the lipoyl domains of the dihydrolipoyl acetyltransferase components of pyruvate dehydrogenase complexes in *Bacillus stearothermophilus* (Dardel *et al*., 1993), *Azotobacter vinelandii* (Berg *et al*., 1997), and *E. coli* (Green *et al*., 1995), and the biotinyl domain of acetyl-CoA carboxylase in *E. coli* (Athappilly & Hendrickson, 1995). The fold common to these domains is a flattened β-barrel, or barrel-sandwich hybrid (Chothia & Murzin, 1993). Each copy of the PEP motif in the domain represents a three-β-strand ''hammerhead''-shaped structure (Athappilly & Hendrickson, 1995) plus an N-terminal fourth strand. The two motif-containing peptides interlock to form a small, globular domain (Figure 3) with 2-fold quasi-symmetry (Dardel *et al*., 1993). The conserved lipoyl-lysine residue lies in a turn between the two motifs and is not present in PEPs, signifying a different function for the lipoyl domain in PEPs.

PEP multiple alignment

Using the motif-detection results and previous structure-based alignments of lipoyl/biotinyl domains (Athappilly & Hendrickson, 1995; Berg *et al*., 1996), a combined motif alignment of the lipoyl and PEP families including gaps was created (Figure 2). The same residues are conserved in the combined alignment as in separate alignments of each family. Moreover, these residues correspond to structurally important features of the lipoyl domain: the conserved glycine residues are found in tight bends in the peptide backbone, the proline residue marks the first turn of the motif, and the conserved hydrophobic residues pack into a well-defined core (Berg *et al*., 1996; Dardel *et al*., 1993). Further, the consistency of the conserved residue pattern and the uniformity of solved lipoyl domain structures has led to general agreement that all lipoyl domains, despite their poor overall residue conservation, have a similar fold. Since each PEP motif contains a complete half-set of lipoyl domain structural features, each motif segment should adopt half of the lipoyl domain fold.

A phylogenetic distance analysis (Galtier *et al*., 1996) was performed on the M_N and M_C motifs of PEPs and lipoyl domains. Although the N-terminal motifs from PEPs and lipoyl domains tend to cluster together, the trend is weak, making the evolutionary relationship between these half-domains uncertain. For example, we cannot say with certainty that M_N of the PEP family is more closely related to M_N of the lipoyl family than it is to M_C

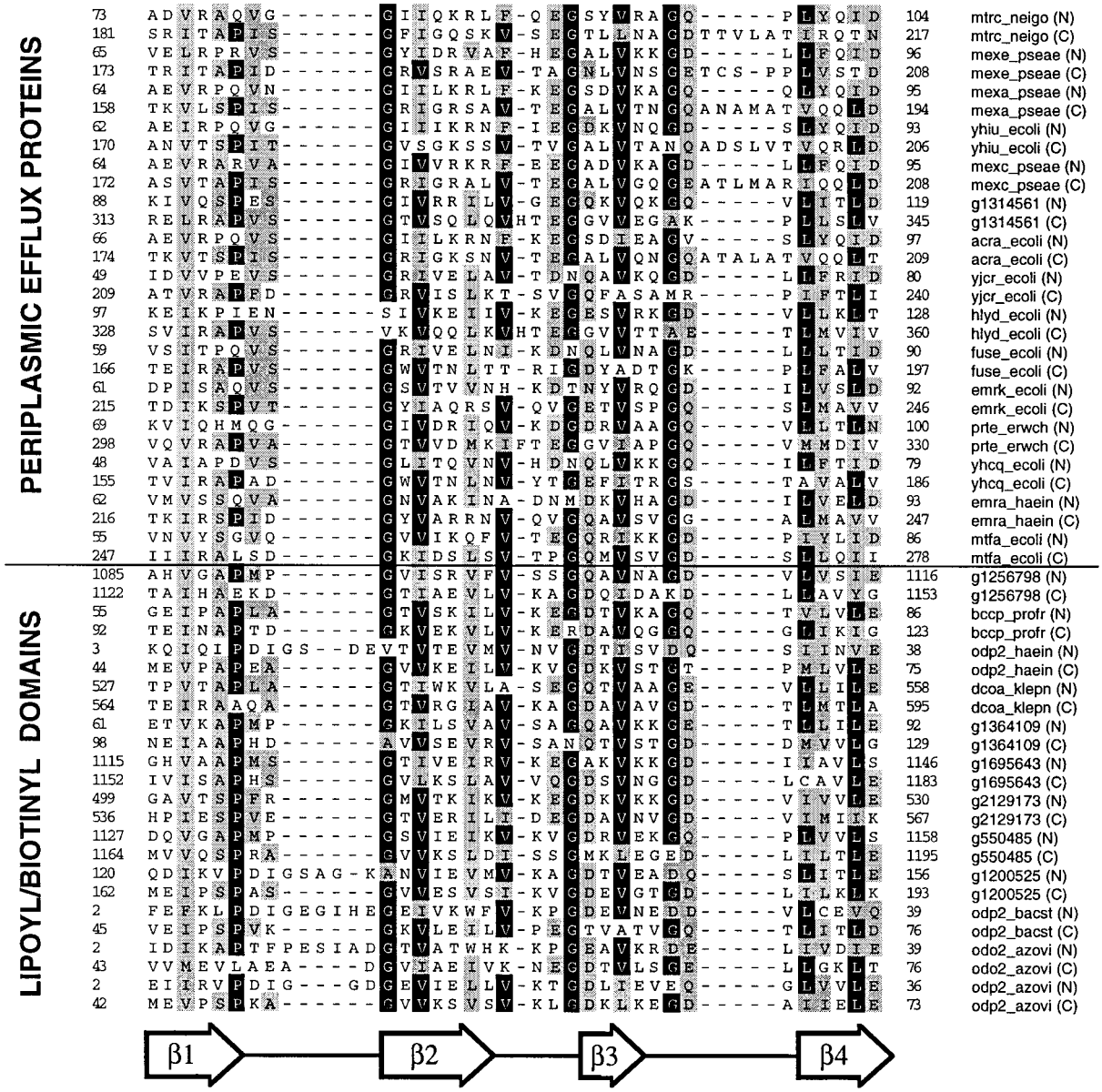

Figure 2. Multiple alignment of sequences with highest scoring matches to the PEP motif, after close homologues have been removed. The 15 PEP and ten lipoyl-biotinyl sequences with the highest scoring matches to the PEP motif (*P*-value $\leqslant 0.0002$) are shown with SWISS-PROT (Bairoch & Boeckmann, 1991) or GenBank (Benson *et al.*, 1993) identifiers. Sequence positions are indicated at each end of the motif. M_N and M_C are shown for each sequence (indicated as N and C), and two lipoyl domain sequences from *Azotobacter vinelandii* with known three-dimensional structures are also included. Black-shaded residues are identical in >50 % of the sequences; gray-shaded residues are similar in >50 % of the sequences. The secondary structure of the motif (from odp2_azovi) is shown below the alignment and labeled as in Figure 3. CLUSTAL W and Megalign (DNASTAR Inc.) were used to make the final alignment including gaps, which are not explicitly identified by the motif-detection algorithms. A few adjustments were made by hand after inspection and superposition of solved three-dimensional structures using the InsightII software package (Biosym/MSI), in order to align residues with the same structural role.

of the lipoyl family. Thus, a combined alignment of the $M_N + M_C$ regions with a full lipoyl domain is speculative, since a full lipoyl domain could also be constructed from two half domains in an intermolecular fold (e.g. $M_N + M_N$), as discussed below.

The low degree of sequence similarity of PEPs (often <25 % identical with one another), has meant that previous alignments contained only a few closely related family members (e.g. Saier *et al.*, 1994), contained incorrectly aligned sequences (Dinh *et al.*, 1994), or were limited to the lipoyl region (Neuwald *et al.*, 1997). The repetitive nature and variable length of the central coiled-coil region also contributes to the difficulty of whole-protein PEP alignment. Fragmentary alignments of smaller PEP subfamilies can also be found in the Pfam 2.1 database (Sonnhammer *et al.*, 1997; B_6656, B_1036, and B_144). A motif representing HlyD and its

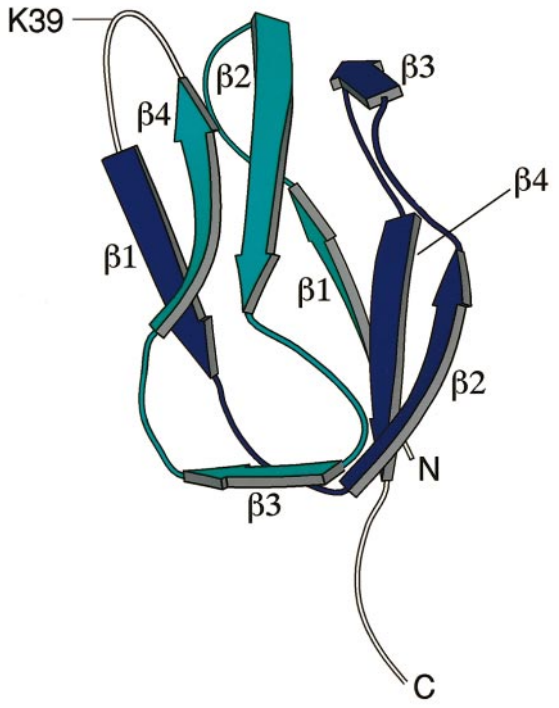

Figure 3. Lipoyl domain structure from *A. vinelandii* pyruvate dehydrogenase (Berg *et al*., 1997), showing that the domain fold comprises two, interlocking PEP-motif sequences. β-Strands from N-terminal (cyan) and C-terminal (blue) motif-containing regions are indicated with numbered arrows. The conserved lysine residue (K39) lies in a loop between motif regions. Structural coordinates were obtained from the Brookhaven database (Bernstein *et al*., 1977). The Figure was created with MOLSCRIPT (Kraulis, 1991).

close relatives is present in the PROSITE database (Bairoch *et al*., 1997; PS00543); it is not common to all PEPs and lies C-terminal to M_C. Preliminary alignments were made for this work using CLUSTAL W (Thompson *et al*., 1994), but these were sensitive to input parameters and did not consistently maintain the motif alignments found by Gibbs sampling. Most PEPs could be aligned together using PSI-BLAST, but the alignments contained an extremely large number of gaps and were poor in the coiled-coil region. Using the Gibbs motifs as a guide, reasonable PSI-BLAST alignments were eventually obtained by separating the PEPs into two subfamilies, one containing MexC with a shorter coiled-coil region and one containing HlyD with a longer coiled-coil region. Final, separate alignments of these two subfamilies were constructed from the PSI-BLAST output using hidden Markov model representations. The aligned coiled-coil regions of several proteins from the MexC alignment are shown in Figure 4.

Circular dichroism spectroscopy of PEP peptides

M_N and M_C peptides from a representative PEP, MexC of *P. aeruginosa,* were selected for synthesis and circular dichroism (CD) spectroscopy in order to determine whether they adopt secondary structure consistent with the predicted fold. MexC associates with the inner membrane transporter MexD and the outer membrane efflux protein OprJ to form a multidrug export complex, which is overexpressed in strains resistant to chloramphenicol, erythromycin, tetracycline, quinolones, and cefpirom (Michéa-Hamzehpour *et al*., 1995; Poole *et al*., 1996). The MexC peptides, residues 64-95 and 172-208, completely span the conserved motif (Figure 2). At 0.1-0.7 mg/ml in 20 mM sodium phosphate buffer, these peptides showed little ordered structure. At concentrations higher than 1 mg/ml the peptides tended to cluster, preventing interpretation of spectra. However, in the presence of 20 % trifluoroethanol (TFE), both peptides exhibited concentration-independent structure with high β-content (β-sheet + β-turn = 84 % and 71 % for M_N and M_C, respectively, at 0.3 mg/ml; see Figure 5). This indicates that the peptides preferentially adopt a β-like fold, in agreement with their role in the lipoyl domain. Low concentrations of TFE (⩽30 %) have been shown to promote peptide conformations consistent with the structure they

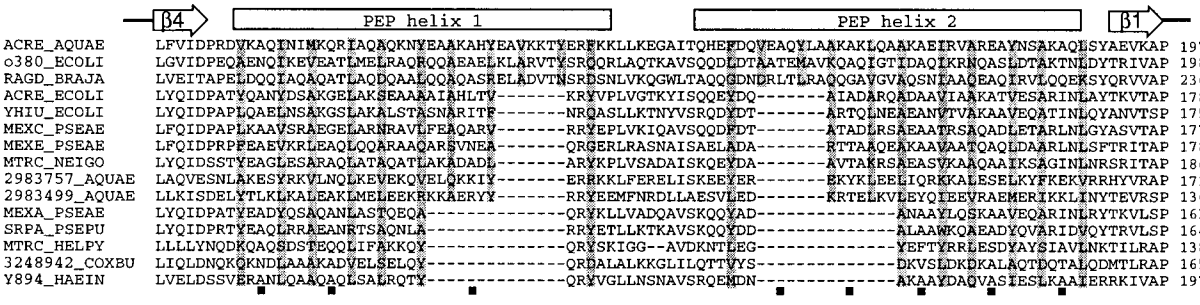

Figure 4. HMM alignment of the coiled-coil region between M_N and M_C for several PEPs of the RND family, where the predicted *a* and *d* positions of the heptad repeat have been shaded. A preference for small, hydrophobic residues is also observed at several *f* positions (black squares). The lengths of the PEP coiled-coil regions differ by integer multiples of seven residues, and the length differences are symmetric across the central aligned region. The exact alignment positions of the seven-residue gaps on each side of the central block are uncertain, given the repetitive similarity of the heptad register. The last β-strand of M_N and the first β-strand of M_C are also indicated.

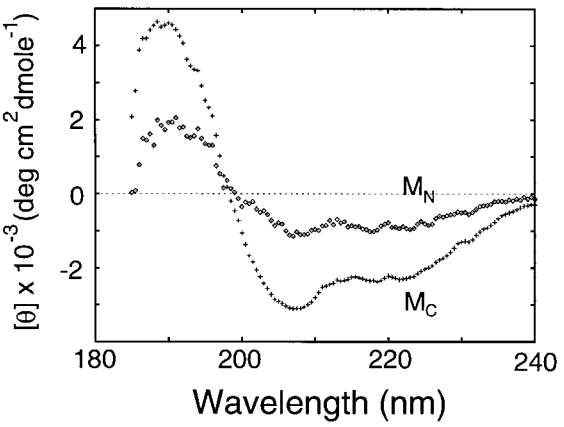

Figure 5. Mean residue ellipticity of M_N and M_C peptides from MexC of *P. aeruginosa*, at a concentration of 0.3 mg/ml in 20 mM sodium phosphate buffer and 20 % TFE. Data were collected in 0.5 nm steps over ten separate scans and averaged over bins of 1 nm width.

adopt in full proteins (Behrends *et al*., 1997), and while β-sheet regions do not always display β-structure as free peptides, a β-signature for a peptide in TFE generally corresponds to β-structure in the native protein. At high TFE concentrations (⩾50 %) both peptides showed increased α-helix content, particularly the M_C peptide, which became 51 % α-helical in 60 % TFE.

Prediction of an α-helical hairpin

The PEP multiple alignment was employed in conjunction with coiled-coil prediction by the MAPF program (discussed in detail below) to reveal a 2-fold symmetry of coiled-coil and PEP-lipoyl motifs (Figures 4 and 6(a)). This symmetry suggests that the PEP molecule could simply fold back on itself at the gap between helical regions, forming an α-helical hairpin and a lipoyl domain fold from the M_N and M_C segments (Figure 6(b), top). Sequences from the α-hairpins of solved structures, such as those of bacterial seryl tRNA synthetases (Cusack *et al*., 1990; Fujinaga *et al*., 1993), also have a predicted region of high coiled-coil probability bisected by a short stretch of low probability. This antiparallel fold model is further supported by the observation that the paired coiled-coil regions of most PEPs are of similar length, and that among different PEPs the lengths of the regions frequently differ by integer multiples of seven. Moreover, these length differences are 2-fold symmetric across the center of the coiled-coil region (Figure 4), suggesting that the two helices interact. Also supporting a hairpin model is the observation that a mutation in the HlyB inner membrane pump can be suppressed by a C-terminal mutation of HlyD (Schlör *et al*., 1997). This places both ends of the proposed PEP hairpin near the inner membrane. Some PEPs with longer helical regions, like HlyD of *E. coli*, do not have clearly symmetrical coiled-coil probabilities and could have additional helices between M_N and M_C.

Alternative PEP fold models

As mentioned above, it is also possible that lipoyl domains are formed by intermolecular association of PEPs, and that the intervening helical region forms a longer coiled coil instead of a hairpin. These possibilities were further investigated by constructing homology models for this PEP region. As seen in Figure 6(b), the topological connection between the coiled-coil and lipoyl domains differs for models with parallel and antiparallel coiled-coil orientations. The reason for the difference is that the symmetry of the lipoyl domain fold is antiparallel, i.e. M_N and M_C enter from opposite sides of the domain. Thus, a parallel, intermolecular fold requires a linker between the coiled-coil and lipoyl modules if they are to be formed simultaneously (Figure 6(b), bottom). This is in conflict with the predicted extent of the coiled coil in most PEPs to within a few residues of M_N and M_C (e.g. two to three residues on each side for MexC), and strengthens the case for an antiparallel fold, either an intramolecular hairpin or an extended, antiparallel homodimer. These are essentially the same fold, since the latter would adopt the same coiled-coil-to-lipoyl configuration as shown in Figure 6(b) (top). If it is assumed that all PEP molecules involved in channel assembly are anchored to the cytoplasmic membrane, then a hairpin is the most plausible antiparallel fold, since both M_N and M_C must be near the membrane.

Models of PEP function

It is known for some efflux pumps that IEP-PEP-OEP complexes are formed only in the presence of substrate, and that the assembly appears to be ordered: substrate binds IEP, which binds PEP, which binds OEP (Létoffé *et al*., 1996). Some IEP-PEP complexes can also form in the absence of substrate, but do not interact with the OEP without bound substrate (Thanabalu *et al*., 1998). Since no exported substrate escapes into the periplasm (Gray *et al*., 1989; Koronakis *et al*., 1989), two very general models of complex formation and function have been proposed: either the inner and outer membranes are brought into close apposition for substrate transfer, or the substrate crosses the periplasm through a closed channel (Holland *et al*., 1990). It is also known that the PEP region which we predict to form a coiled coil is required for hemolysin export in *E. coli*, since deletion of a portion of this region or substitution with a similar sequence abolishes export (Schlör *et al*., 1997). Below, we discuss these two hypothetical models of efflux pump assembly, where the proposed structure of the PEP is taken into account.

In the first model, by association with the OEP or outer membrane, the PEP brings the inner membrane pump into close apposition with the OEP for

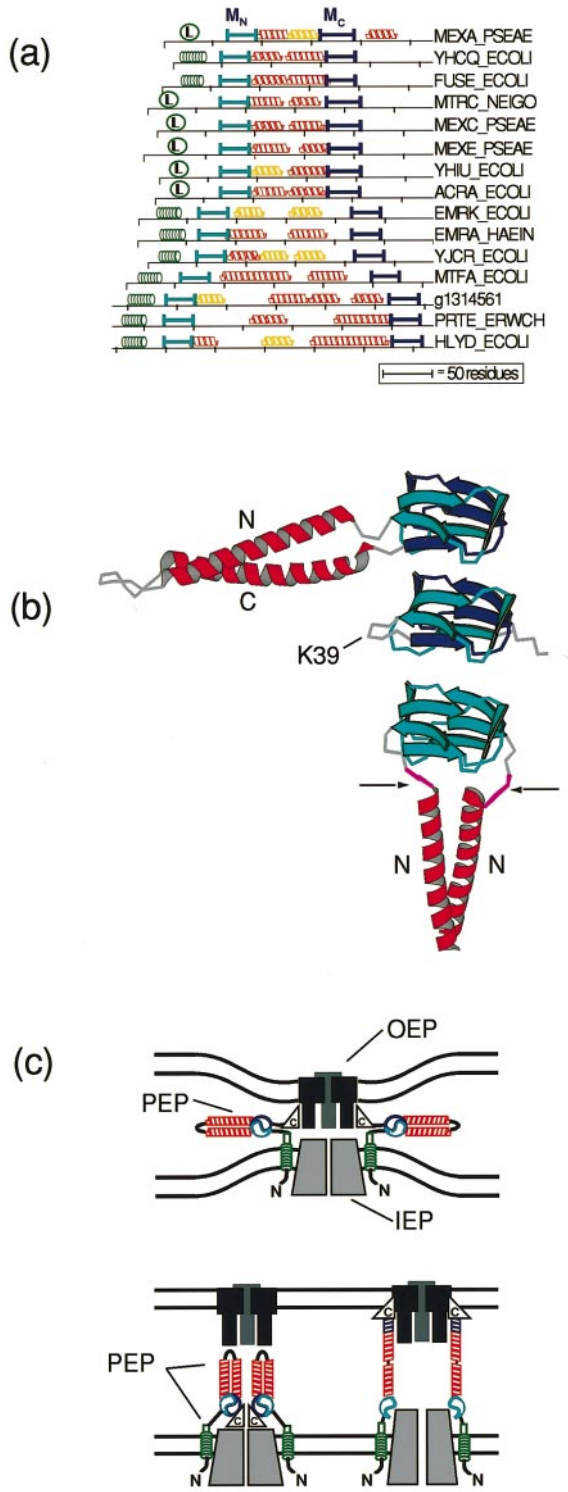

Figure 6. (a) Primary structure map of sequence analysis results for the 15 PEP sequences of Figure 2, showing the predicted locations of PEP motifs (cyan and blue), transmembrane helices (Hofmann & Stoffel, 1993; green coils), and lipoprotein attachment sites (green circles). Red helices represent regions with continuous coiled-coil probability >0.50 (window = 28), and yellow helices correspond to continuous regions of 14 residues or more with probability 0.25-0.50 for a 28-residue window, or >0.50 for a 21-residue window. The sequences are centered between M_N and M_C and grouped by motif spacing; some are truncated at the ends. (b) Intramolecular (top) and intermolecular (bottom) models of lipoyl domain and coiled-coil structure, shown next to the lipoyl domain of *A. vinelandii* pyruvate dehydrogenase in the same lipoyl domain orientation (center). The M_N (cyan) and M_C (blue) regions of MexC, as well as the two MexC segments with coiled-coil probability >0.6 (red, labeled N and C), were aligned (as in Figure 2 and by heptad register) and fit to the conformation of appropriate lipoyl (Berg *et al*., 1997), antiparallel coiled-coil (Fujinaga *et al*., 1993), and parallel coiled-coil (O'Shea *et al*., 1991) structures using the Insight II-Homology package (Biosym/MSI). Residues between the coiled-coil regions and lipoyl motifs (gray) were assigned a loop conformation and energy-minimized (200 steps steepest-descent). Three residues from the MexC coiled-coil region were added to each of the connecting loops of the intermolecular model (magenta, indicated by arrows), to bridge the distance across the domain without unnatural bond lengths. The intermolecular model shows the M_N to M_N pairing only. The Figure was created with MOLSCRIPT. (c) Two hypothetical, schematic models of PEP function, assuming intramolecular formation of the lipoyl domain. In one model (top) two coiled-coil regions (red) and two lipoyl half-domains (M_N in cyan, M_C in blue) associate in each PEP to clamp together the IEP and OEP. The PEP molecule bends at the gap between the coiled-coil regions, contacts the trimeric OEP *via* a C-terminal domain, and is anchored to the inner membrane by a hydrophobic, N-terminal helix (green). A second model (bottom) shows that the PEP helices could instead form part of a *trans*-periplasmic channel, either in a hairpin conformation or in an extended conformation, perhaps in response to conformational changes in the IEP which disrupt the lipoyl domain. Only two PEP molecules are shown for clarity, although HlyD appears to be trimeric in crosslinking experiments (Thanabalu *et al*., 1998).

substrate transfer. The formation of the α-hairpin and intramolecular lipoyl domain is one possible mechanism for closing the distance between membranes in this model (Figure 6(c), top). Such a model also provides a mechanism for the proposal that substrate extrusion occurs at regions of close membrane apposition (Holland *et al*., 1990), or the ''membrane adhesion sites'' observed by Bayer (1991).

The second model of PEP function posits that PEPs form part of a substrate channel connecting the two membranes. The outer membrane efflux protein TolC is believed to have a substantial periplasmic domain (Koronakis *et al*., 1997), which may also contain coiled coils (discussed below) and could interact with the coiled coils of the PEP HlyD to form the channel. Such a channel could incorporate the α-hairpin as a part of the channel, or the OEP could interact with hairpin helices that have been ''opened'' by substrate binding (Figure 6(c), bottom). This conformational change is consistent with an increase in protease accessibility in HlyD upon substrate binding, measured by Thanabalu *et al*. (1998). The dissociation of M_C to participate in the interaction with OEP helices is also consistent with the CD data showing that M_C

of MexC has a partly helical character in solution. Also indicating that the PEP coiled coils may form part of a larger helical assembly, either by trimerization or by interaction with OEPs, is residue similarity at the *f* position of the PEP heptad repeat (Figure 4), which is located opposite the hydrophobic packing faces of the helices. Finally, database searches with hidden Markov models (discussed below) found a likely PEP-family protein in *Bacillus subtilis*, which has no outer membrane. This suggests that the PEP molecule has a function which does not require contact with the OEP. One possible function is to coordinate substrate transport across the peptidoglycan layer, which is present in both Gram-negative and Gram-positive bacteria.

A very different role for the lipoyl domain in PEPs has been proposed by Neuwald *et al*. (1997), based on the lipoyl-PEP homology, in which the sequence between M_N and M_C binds and swings the substrate from one membrane to the next, analogous to the mechanism by which lipoyl domains may convey lipoyl groups between enzymatic components. However, as we have shown, this intervening region consists almost entirely of sequence predicted to form coiled coils, suggesting a structural role rather than functions such as substrate binding and translocation. Moreover, previous studies of ATP-binding cassette exporters indicate that the IEP alone is responsible for substrate specificity (Akatsuka *et al*., 1997; Binet & Wandersman, 1995).

Outer membrane efflux proteins

OEP motifs and alignment: OEPs have a tandem repeat

For the OEP family we employed a similar strategy of finding common sequence motifs, using the motifs to locate new family members and guide construction of a multiple alignment, and then using the multiple alignment to identify common structural features. Although OEP sequences are highly divergent, two subtle but significant motifs were found by iterative use of Gibbs sampling algorithms. Each of these motifs uniquely defines the OEP family; i.e. there are no high-scoring matches for these motifs to non-OEPs in the database.

The first motif is found in two copies in all OEPs and is terminated by the pattern ΦΦPxΦxΦ, where Φ is hydrophobic, P is a proline residue, and x is a position of lower residue similarity (Figure 7(a) and (b)). The central part of this motif has a heptad repeat pattern suggestive of coiled-coil structure, with alanine residues favored at the *d* positions. A second motif is also found in two copies in every

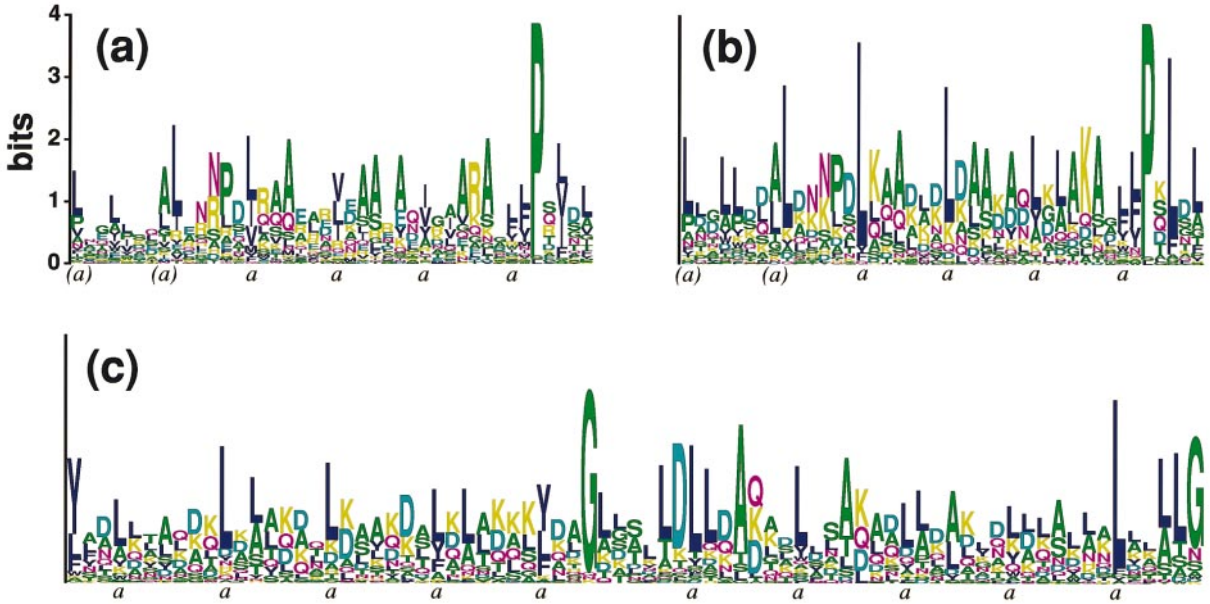

Figure 7. Sequence logo display (Schneider & Stephens, 1990) for the two highest scoring OEP motifs, made from alignments of 56 sequences representing 28 OEPs. Residue information content at each sequence position is shown for the first repeated motif (a), corresponding to residues 27-68 and 233-274 of mature TolC. The same motif is shown in (b), where the global substitutions IVM → L, E → D, and R → K have been made in the alignment. This reduced alphabet allows the hydrophobic positions of the heptad repeat to be seen more easily. These substitutions are also used to display the second motif (c), which represents the regions from 137-211 and 355-429 of TolC. Many of the OEPs have high coiled-coil probabilities for these two motif regions. Predicted *a* positions of the heptad repeat are indicated below the horizontal axis; these tend to have a high proportion of hydrophobic residues (purple). All *d* positions in both motifs show a preference for alanine residues, except the two N-terminal heptads of the first motif (shown in parentheses). Only in some OEP sequences are these 14 residues predicted to form coiled coils.

OEP (Figure 7(c)). This motif contains two heptad-repeat regions flanking a central, conserved G(x5)D pattern. A strong preference for small residues, generally alanine, is again observed at the *d* positions in the second region of heptad repeats. The coiled-coil-forming potential of these sequences is discussed further below. The two motifs of Figure 7 are tandemly repeated in all OEP sequences, suggesting that they have arisen from a duplication event. Importantly, this also suggests that the two halves of OEP molecules have similar three-dimensional structures. Tandem sequence similarity has been observed for three closely related OEPs in prior studies (Gross, 1995), but could not be detected in TolC with the same method, and has not been observed previously to be a general feature of OEPs. There are also no PROSITE, PRINTS (Attwood *et al.*, 1994), or Pfam (2.1) motifs which represent the OEP family as a whole, although the Pfam database contains an entry (B_1036) which aligns C-terminal fragments of 11 OEPs closely related to TolC.

OEP multiple alignment

A whole protein alignment of 39 members of the OEP family was made with an iterative alignment-building process involving PSI-BLAST, HMMs, and the tandemly repeated OEP motifs (see Materials and Methods). Careful choice of parameters was required, since the default parameters of PSI-BLAST lead to inclusion of repetitive helical proteins such as tropomyosin within a few iterations. The alignment stretches from just beyond the N-terminal signal cleavage site of TolC at residue 25 to the conserved glycine residue at TolC residue 429, near the C terminus of most OEPs. An alignment of the tandemly repeated N and C-terminal halves of OEP sequences was also made in the same way. The tandem repeat corresponds approximately to TolC sequences 35-210 and 241-429.

Phylogenetic analysis of the tandem repeat revealed that among Gram-negative bacteria, the N-terminal halves are more closely related to one another than to the C-terminal halves and *vice versa*. Among more distant relatives such as *Aquifex* and *Synechocystis*, the relationships are not as clear. The N-terminal half of an OEP from *Synechocystis*, for example, shares higher sequence similarity with the C-terminal domain from the same protein than with the N-terminal domain from any other protein in the multiple alignment. Multiple duplication events, horizontal transfer, and concerted evolution are among the possible interpretations of these data.

Combining multiple alignment and structure prediction

An alignment of highly divergent proteins with a common function is a rich source of information, since any vestigial sequence similarity must be important to the structure and mechanism of the family. In addition to identifying conserved individual residues like the glycine and proline residues mentioned above, one can look for other conserved characteristics of the alignment. We first used the OEP alignments to find regions predicted to conserve secondary structure with the PHD program (Rost & Sander, 1993). Each half-alignment of the tandem repeat was run separately, and the output shows six strongly predicted α-helical regions, and several likely loops (Figure 8). The loop prediction is particularly strong for sequences between helices h1 and h2, and between h3 and h1′. Surprisingly, very little β-structure is predicted; the most prominent peaks are tandemly repeated and located C-terminal to h1/h1′ and N-terminal to h2/h2′. These results are discussed further below.

In principle, any protein sequence property can be calculated for all proteins in an alignment and the results combined to improve prediction power. This idea is used by the PHD program and other secondary structure prediction methods (Barton *et al.*, 1991; Crawford *et al.*, 1987; Levin *et al.*, 1993; Salamov & Solovyev, 1995; Zvelebil *et al.*, 1987). To extend this principle beyond secondary structure prediction, the MAPF program was developed. MAPF is automated and general, taking as its input a multiple alignment and a list of structural prediction algorithms to be performed. Each algorithm is performed on each aligned sequence and synchronized with the multiple alignment. The MAPF output is thus an aligned set of structural predictions for each prediction algorithm. The average of each prediction function over the family is also calculated, such that the contribution of

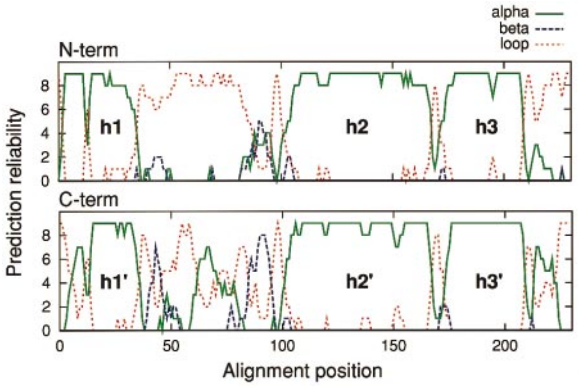

Figure 8. Secondary structure prediction for an alignment of 34 OEPs, including OEPs identified here. The alignment of OEPs was separated into N and C-terminal groups and submitted to the PHD neural-network prediction server. One N-terminal sequence (g2314661) with a 48-residue insert at alignment position 65 was removed for display purposes. Prediction reliability from zero to 9 is indicated, and predicted helices are labeled. A reliability score of 9 corresponds to prediction accuracy >96%, although the method is presumably less accurate for membrane proteins (Rost & Sander, 1993).

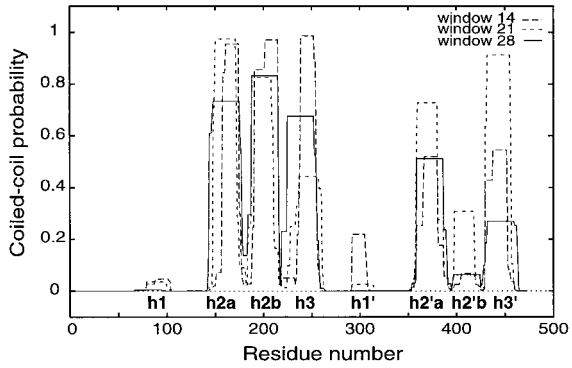

Figure 9. Coiled-coil probability for the OprN protein (Köhler *et al*., 1997), an OEP in *P. aeruginosa*, shown for three different window lengths, where the eight putative coiled coils have been labeled. This example shows the tandem repeat of OEP sequences and the uniform size of the predicted helices.

each protein to the average is weighted to remove bias from highly similar sequences. Programs for secondary structure prediction, coiled-coil prediction, and for other sequence characteristics indicative of transmembrane (TM) β-strands were included in MAPF for analysis of the OEP family.

OEPs contain coiled coils

Since the OEP family appeared to contain a large proportion of helical structure, and since a heptad repeat pattern was evident in the motifs common to the family, we looked for potential coiled coils in the multiple alignment using MAPF and the COILS program. High coiled-coil probability was found in the sequences of each of the six OEP helices identified by secondary structure prediction. Putative helices h2 and h2′ are generally subdivided into two distinct coiled-coil probability peaks. This means there are eight possible coiled-coil sections, four from each half of the tandem repeat structure, which is clearly perceptible in the coiled-coil predictions (Figure 9). At least one member of the OEP alignment has coiled-coil probability $\geqslant 0.95$ for each of the eight potential coiled-coil segments, although some segments are more strongly predicted than others (Table 1). The strong consensus prediction for h2b, h3, h2′a, and h2′b makes a coiled-coil or helical bundle structure extremely likely for these regions. Sequences representing these four helices also scored highly as coiled coils using the PAIRCOIL program, which is believed to have a lower false-positive rate (Berger *et al*., 1995). On the basis of these data, we conclude that at least four of the eight helical sections

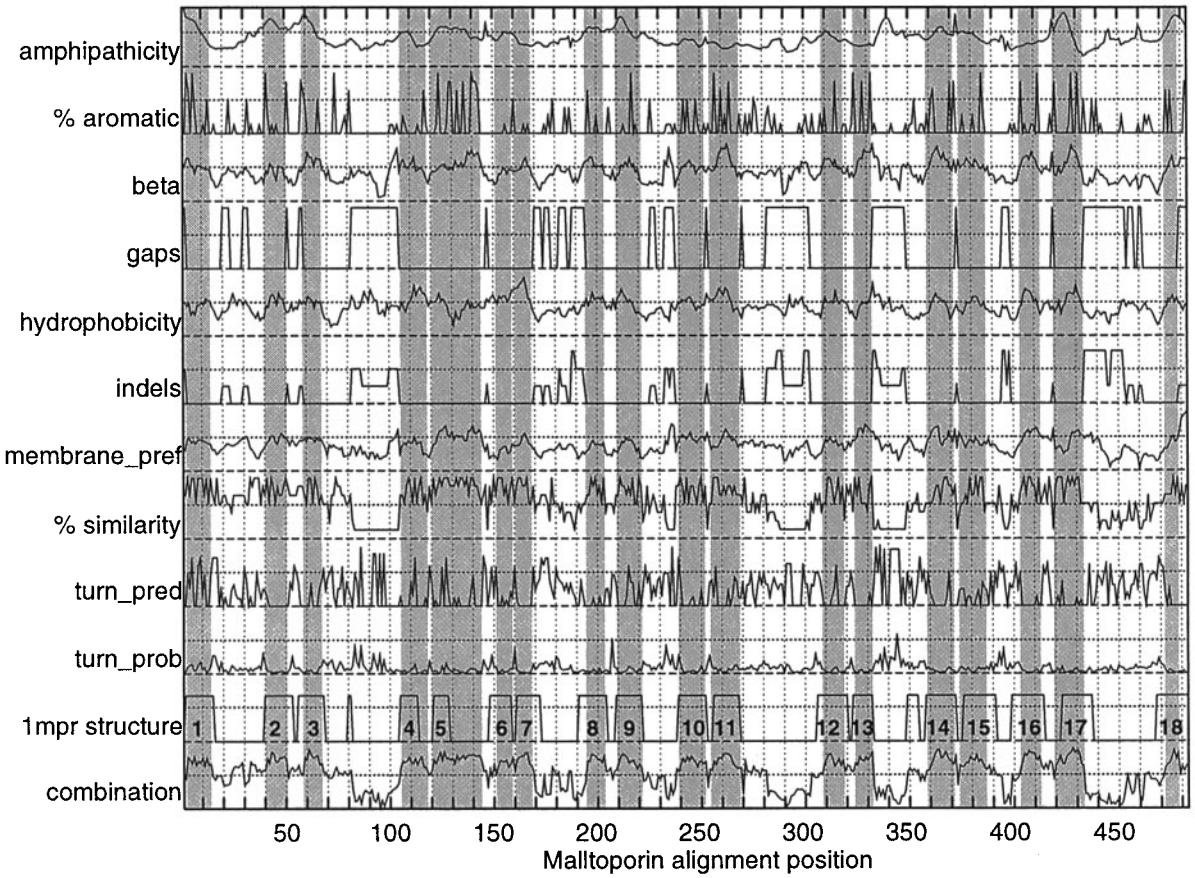

Figure 10(a) (*legend overleaf*)

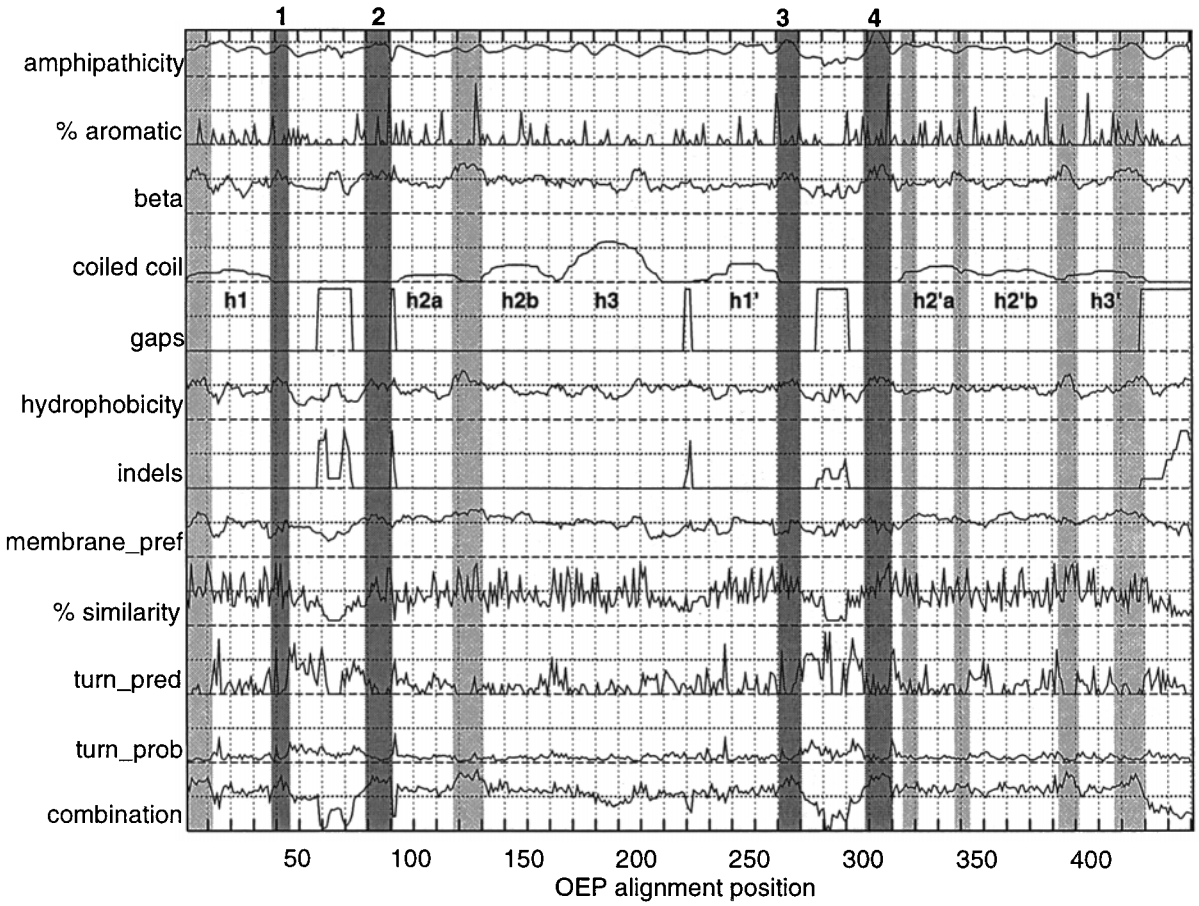

Figure 10. (a) Transmembrane β-strand prediction for the maltoporin family. MAPF output is shown for an input alignment of seven maltoporin family members. Weighted alignment averages of nine predictors hypothesized to correlate or anticorrelate with transmembrane β structure are shown, along with the numbered positions of the β-strands of *E. coli* maltoporin (1mpr). Two of the strands are split due to alignment gaps. Chou-Fasman turn probability, turn prediction, and β propensity (Chou & Fasman, 1978), hydrophobicity (Kyte & Doolittle, 1982), amphipathicity (Eisenberg *et al*., 1984), porin membrane preference (Gromiha *et al*., 1997), percentage similarity, insertions/deletions (indels), and gap positions were incorporated into the combined predictor (combination). See Materials and Methods for details. The ''indel'' measurement is the minimum number of insertions or deletions observed at each position. Gray bars represent combination function values above 0.65 over a continuous span of more than six consecutive residues. The percentage of aromatic residues is also shown for each alignment position, since these are often found at the membrane surface of porin β-strands (Kreusch *et al*., 1994). The Figure was generated by MAPF using the plotting program GNUPLOT. (b) MAPF output and candidate TM β-strands for outer membrane efflux proteins. Shaded in gray are regions which match the criteria used for porin TM β-strand prediction as in (a). Four of these regions (dark gray) are symmetric with respect to the tandem repeat (see the text for discussion). The input to MAPF used for this Figure was a set of 13 OEPs closely related to TolC of *E. coli* and OprJ of *P. aeruginosa*. The same set of structural predictors was used as in (a), with the addition of coiled-coil probability (window = 21). Putative helices are labeled, and the percentage of aromatic residues in the alignment is shown for reference.

are highly likely to adopt a coiled-coil-like structure. A solid case can also be made for h3′ based on sequence similarity with h3. Coiled coil formation is less certain for h2a, which is less similar to h2′a than the other tandemly repeated helix pairs are to one another. Putative helices h1 and h1′ also have fewer representatives with high coiled-coil probability, in part due to the conserved proline residues at the ends of the helical regions (Figure 7).

The drop in coiled-coil probability between h2a and h2b (and between h2′a and h2′b) observed for many OEP sequences is generally coincident with a shift in the register of the heptad repeat by omission of three residues. This conserved stutter is consistent with ''*x*-layer'' or ''*da*-layer'' packing arrangements which decrease the extent of supercoiling and are often found in helical bundles (Lupas, 1996a; Lupas *et al*., 1995). Thus, the break in coiled-coil probability observed here may not indicate a break in helical structure, particularly since it does not coincide with an increase in loop probability (Figure 8), a decrease in sequence conservation, or an increase in alignment gaps.

Table 1. The numbers of OEPs in a multiple alignment of 34 OEPs which have coiled-coil probabilities above 0.95, 0.70, and 0.50 (for COILS window 21 or 28) within each predicted helical region

Predicted helix	$P \geqslant 0.95$	$P \geqslant 0.70$	$P \geqslant 0.50$
h1	2	7	7
h2a	2	5	8
h2b	6	11	14
h3	8	16	22
h1′	2	12	14
h2′a	5	14	16
h2′b	4	7	12
h3′	1	4	6

A second region of the alignment between predicted coiled coils h2b and h3 (and between h2′b and h3′) also has a conserved spacing between heptad repeats, corresponding to a deletion of four residues. Although this heptad break is predicted to form a loop, the coiled-coil structure could continue across the break instead, forming a ''stammer'' (Brown *et al*., 1996). This theoretical packing arrangement also involves an *x*-layer, in which residues from opposing coils point directly at one another, requiring a very small residue. At this *x*-layer position in the OEP alignment is a highly conserved glycine residue (Figure 7(c)), consistent with this hypothesis. Thus, it is also a possibility that h2a-h2b-h3 and h2′a-h2′b-h3′ are capable of forming contiguous coiled coils over 100 residues in length.

Although all of the putative OEP helices contain a distinct heptad repeat, the pattern is not the canonical leucine-zipper signature found in most two-stranded, parallel coiled coils. The aligned sequences display a clear preference for alanine in the *d* position of the heptad repeat. The presence of alanine in the hydrophobic core appears to promote an antiparallel orientation in tetrameric coiled coils, to create packing layers of the type Leu-Ala-Leu-Ala as opposed to Ala-Ala-Ala-Ala and Leu-Leu-Leu-Leu (Monera *et al*., 1996). Further, alanine in the *d* position has been theorized to correlate with the type of antiparallel four-helix bundle found in the structure of the Rop protein (Banner *et al*., 1987; Gernert *et al*., 1995). These considerations suggest that an antiparallel orientation may be favored for some of the OEP helices. It can also be seen in Figures 9 and 10(b) that the predicted coiled-coil regions have approximately the same length, about 25-30 residues. The uniform size of the coiled-coil regions reinforces the suggestion that some or all of them may associate in a larger helical bundle or multimeric coiled-coil assembly. This could also explain why the coiled-coil predictions are not uniformly high for all members of the family, since the COILS program does not always detect higher order coiled-coil associations. OEP helices could also form transient complexes with PEP helices for substrate efflux, as we proposed above. Another candidate for interaction with TolC helices is the TolA protein, which has a periplasmic helical domain (Levengood *et al*., 1991) with a high probability of coiled-coil formation.

Prediction of porin transmembrane β-strands using MAPF

TolC has been predicted by others to form a porin-like β-barrel (Koronakis *et al*., 1997; Paulsen *et al*., 1997), in apparent conflict with the secondary structure and coiled-coil predictions for the OEP family described here. To help resolve this conflict, we applied the MAPF sequence analysis method to the task of identifying transmembrane β-strands. We trained the method on families of known porins and subsequently applied it to the OEPs. Our goals were twofold: to test whether TolC and its relatives have sequence characteristics typical of porins, and to identify potential transmembrane β-strands in the OEP alignment.

TM β-strands are more difficult to identify than TM α-helices, since a β-strand can cross the membrane in as few as six residues, and only a few of these may be non-polar. Sequence characteristics such as hydrophobicity, amphipathicity, secondary structure, and sequence identity have been used previously to predict TM strands, as have conjunctions of these criteria (Fischbarg *et al*., 1994; Gromiha *et al*., 1997; Jeanteur *et al*., 1991) and neural networks (Diederichs *et al*., 1998). The method presented here was developed, trained, and tested on multiple alignments of three porin families comprising 29 proteins. One member of each family has a known structure: OmpF from *E. coli* (Cowan *et al*., 1992), maltoporin from *E. coli* (Schirmer *et al*., 1995), and porin from *Rhodopseudomonas blastica* (Kreusch *et al*., 1994). The alignments were generated automatically using PSI-BLAST and hidden Markov models in the same way as for the OEPs (without knowledge of the structures) and given as input to MAPF, along with algorithms for nine relevant sequence characteristics or prediction methods. MAPF generates separate alignments of each sequence characteristic as well as a plot showing weighted averages of all characteristics together. An example of the latter for the maltoporin family (Figure 10(a)), shows that alignment-averaged functions of residue similarity and β-structure correlate well with known TM β-strands, as expected, whereas alignment gaps and turn predictions are strongly anticorrelated with TM strands.

To see if these individual functions could be combined into a more accurate predictor of TM β-structure, a cross-validation procedure was then used to weight each predictor in proportion to its accuracy in each of the three families and add them to create a combined predictor function (see Materials and Methods). The correlation coefficient between the combined predictor and the actual structure was greater than any of the individual predictors for each family, demonstrating that combining structural criteria can improve prediction accuracy. Further, when a simple threshold

was applied to the combination function, such that continuous regions of more than six residues with values >0.65 are predicted to be TM β-strand, this resulted in correct assignment of 77% of the residues from maltoporin (see Figure 10(a)), 73% for OmpF, and 78% for *R. blastica* porin. The average prediction accuracy of 76% for these three protein families is higher than that of other described methods (Paul & Rosenbusch, 1985; Stoorvogel *et al*., 1991; Vogel & Jähnig, 1986; Gromiha *et al*. 1997; Gromiha & Ponnuswamy, 1993). The only computational methods reporting equal or higher accuracy (of which we are aware) either include the test protein in the training procedure (Ponnuswamy & Gromiha, 1993) or use only a single test protein which has significant sequence similarity to one of the training proteins (E-value $= 9\ e^{-10}$ by BLAST search of PDB; Gromiha *et al*., 1997). In contrast, the three test porins used in our cross-validation procedure have no significant sequence similarity to one another (E-values > 1).

The false positive residues overpredicted by the combination function are found mostly in internal loops of the barrel structure which adopt an extended conformation but do not cross the membrane. Residues 130-144 of the maltoporin alignment, for instance, are incorrectly predicted to be in a TM segment (Figure 10(a)). These residues, which have many of the characteristics of TM β-strands, form a highly conserved loop which packs into the porin channel to constrict the pore. Most false negatives occur at the extracellular ends of TM β-strands and may be due to the fact that residues were counted as TM β if they participated in β-barrel hydrogen bonds in the crystal structures. Since some hydrogen-bonded strands extend out of the membrane, a few extracellular β-strand residues may not have transmembrane sequence characteristics and would thus have lower combination function values. Other false negatives result from the automated alignment procedure which can occasionally insert a gap between residues of the same TM segment (e.g. strand 13 of Figure 10(a)).

Recently, several new structures of porin-like, β-barrel proteins have been solved, including *E. coli* proteins FhuA (Ferguson *et al*., 1998), FepA (Buchanan *et al*., 1999), and OmpA (Pautsch & Schulz, 1998), which have 22, 22, and eight transmembrane β-strands, respectively. As an additional test of the prediction method, alignments of these families were also submitted to the MAPF program, using average weights from the three porin families in the training set. However, unlike the training set porins, each of these proteins has an additional domain lacking TM β-strands. In particular, nearly half of the residues in OmpA form a periplasmic domain at the C terminus, which is more strongly conserved and has a higher average hydrophobicity than the OmpA transmembrane domain. Using the same criteria as for the porin families tested above, the prediction accuracy for the FhuA, FepA, and OmpA families was 74%, 73%, and 57%, respectively. The lower score for the OmpA family results from false positive predictions in the C-terminal domain. If the analysis is performed on the OmpA transmembrane domain only, the accuracy improves to 81%. Accuracy for the FhuA and FepA predictions also increases when analyzing only their transmembrane domains, to 77% and 79%, respectively. Thus, although the method is clearly biased toward highly conserved, hydrophobic regions, it performs well on outer membrane domains which are substantially larger or smaller than the porins in the training set.

Prediction of transmembrane β-strands for OEPs

The same method for predicting porin β-strands was then applied to multiple alignments of the OEPs, using predictor weights derived from the three porin families. The MAPF output showed regions of the alignment with many of the characteristic features of TM β-strands, but the combined predictor function which was found to be useful in the porin-like proteins did not contain many well-defined peaks. The major difference in the combination function appeared to be due to the predicted OEP coiled-coil regions. These do not have many gaps or predicted turns, and have fairly uniform sequence characteristics, which led to a relatively high, flat combination function over the length of each predicted helix. Since the training set was composed of porins which do not contain coiled coils or long helices, this result might have been anticipated. Given that coiled coils and TM β-strands are mutually exclusive structures, we incorporated coiled-coil prediction into the combination function, using a weight equal to that of the ''gaps'' feature. The resulting MAPF output (Figure 10(b)) reveals persistent and substantive differences from the porin results. In particular, the strength and frequency of predicted turns in OEPs is much lower, the number of aromatic residues is lower, and there are many fewer peaks in the β-strand prediction and TM-β combined predictor functions. Moreover, predicted β-segments of the porin alignments generally appeared in pairs, representing two TM strands and a periplasmic turn (Figure 10(a)). Similar patterns are not observed in the OEP family combined predictor.

Despite these differences, several regions of the OEP alignment, generally adjacent to the putative helices described above, show strong indications of TM β-structure. Region 4 of Figure 10(b), for instance, is highly hydrophobic, amphipathic, has strong beta prediction, is void of gaps and turn predictions, has a high degree of residue similarity, and has a correspondingly high combination function. Several candidate β-strands, including strand 4, also have conserved aromatic residues near their membrane borders, often observed in porin structures. In addition, candidate strands 1 and 2 have homologous sequences to segments 3 and 4, respectively, a consequence of the tandem repeat.

These four regions were also assigned a relatively high probability of β-structure by the PHD program (Figure 8) and are the most likely TM strands.

In general, the two halves of the OEP tandem repeat appear to share most structural characteristics, supporting the claim that they have similar structures. The tandemly repeated features include predicted helices and coiled coils, two candidate β-strands, several conserved aromatic residues, and patterns of high and low turn probability. Ideally, the tandem repeat provides two independent data sets which can be compared to improve structural predictions. Predictions of β-strands in only one half of the tandem repeat are probably less reliable than strand predictions at identical positions in both halves. For example, the region between h2′b and h3′ has conserved features typical of TM β-strands, but the symmetric section between h2b and h3 is much less hydrophobic and has many predicted turns. The characteristics of the sequence between h2a and h2b are also different in each half of the repeat, including a more prominent gap in coiled-coil probability for the N-terminal half. It is not clear whether these differences reflect variations in the folded structure of the two halves, or merely reflect asymmetries in packing arrangement, interaction with the membrane, or interaction with other proteins. Two other regions of high combination function, located N-terminal to h2′a and C-terminal to h3′ are also potential TM β-strands, although they do not have strong predictions in both repeats. The candidate strand N-terminal to h1 lies outside the tandem repeat.

Most of the candidate TM β-strands indicated by the MAPF procedure contain 12 or fewer residues and probably represent single β-strands. Pairs of β-strands in porins tend to span 15-30 residues or more (see Figure 10(a)). Extra β-strands could be formed by N or C-terminal residues outside of the alignment, and one or two additional β-strands could be present between the tandem repeats (between h3 and h1′) where the alignment is least certain. Likewise, some sequence regions with large combination function may not be TM β-strands, but may instead be internally packing loops, which as we have noted share many of the sequence features of TM β-strands. Further, the procedure used here was trained to identify porin-like β-strands and may not apply if the fold of the OEP family is significantly different.

Interpretation for OEP structure

There are several reasons to believe that OEPs are porins. They reside in the outer membrane, have transport functions, and can serve as phage receptors, all like porins. TolC also forms pores in artificial bilayers which can be blocked by the addition of polypeptides (Benz *et al.*, 1993), and appears to be a trimer by gel filtration and electron microscopy (Koronakis *et al.*, 1997). However, the evidence presented here indicates that while these proteins have several regions which have sequence characteristics of TM β-strands, OEPs are unlikely to adopt the 16-18 strand β-barrel fold observed for porins. First of all, the combined predictor function for OEPs does not show the distinct, two-peak regions of probable TM β-structure seen in all six porin-like families. In addition, secondary structure predictions show that the OEP structure is apparently dominated by α-helices. The secondary structure of porins, in comparison, is predicted (correctly) to be dominated by β-strands and loops (Table 2). Further, OEPs are highly likely to contain coiled coils, whereas none of the 29 porins from the three aligned families has coiled-coil probability above 0.06 for any residue (window 28). Thus, while it is possible that OEPs have a TM β-barrel composed of a small number of strands (perhaps similar to OmpA), it is unlikely that there is a continuous sequence region forming this domain, as is the case for OmpA and the porins.

Another clear difference between OEPs and porins is evident in the distribution of gaps in the alignments. It is unusual for highly divergent protein sequences to share a long sequence motif without gaps such as we observe in the OEP family (Figures 7 and 10). The porin family alignments, in contrast, tend to have gaps interspersed regularly between positions of adjacent β-strands, despite the families being more closely related than OEPs. While this could mean that the lengths of OEP loops are highly conserved, a more likely interpretation is that there are many fewer loops in the OEP fold, and that individual secondary structural elements are composed of larger numbers of residues, what one would expect for helices. This interpretation is also supported by the contrast in

Table 2. Distribution of secondary structure types predicted by the PHD server for OEP alignments in comparison to alignments of three porin families

Family	Proteins	% α-helix	% β-strand	% Loop
OEP	33	61.4	3.8	34.8
OEP (N-term only)	33	62.7	2.2	35.1
OEP (C-term only)	33	68.9	6.2	24.8
Porin (*R. blastica*)	5	12.1	40.0	47.9
Maltoporin	7	9.7	23.0	67.3
OmpF	17	5.6	30.7	63.7

the number of predicted turns for the two families. Moreover, unlike all known porin structures, TolC appears to have a substantial periplasmic domain by electron microscopy of two-dimensional crystals (Koronakis *et al.*, 1997). We propose that this domain is formed in part from helices, in a coiled-coil or helical bundle structure, and speculate that these helices interact with PEP helices to form a *trans*-periplasmic channel.

Finally, we investigated the possibility that porins and OEPs may be evolutionarily related by looking for sequence similarity with Gibbs sampling and other methods described above. No significant motifs were found linking the two families, indicating that any sequence relationship is very distant. We also did not find evidence that porin sequences have tandem symmetry, which might also be expected if they share the same fold with OEPs.

Thus far, only two classes of integral membrane proteins have been observed by crystallography: helical bundles and β-barrel proteins (von Heijne, 1997). As we have shown, TolC and its relatives have characteristics of both classes of proteins, preventing straightforward classification as either type. The large fraction of helices predicted with a high degree of confidence, combined with several likely TM β-strands, suggests that OEPs adopt a mixed α/β fold unlike any outer membrane protein of known structure. Since the hydrophobic regions of the OEP family are typically less than 20 residues long (Figure 10(b)), the transmembrane segments of OEPs are likely to be β-strands rather than α-helices. However, to test whether any of the predicted helices might be involved in spanning the membrane, MAPF was used to identify hydrophobic and polar faces of each helix in the multiple alignment (data not shown). The putative helices h2a, h2b, h2′a, h2′b, and h1′ are consistently amphipathic, and could theoretically reside in the membrane as part of a helical bundle, with polar faces away from the membrane. This would be highly unusual, however, since transmembrane helices are typically hydrophobic on all sides and are thought to be individually stable in the membrane (Hunt *et al.*, 1997). It is more likely that these helices pack against other parts of the efflux pump structure or lie parallel with the membrane surface, rather than crossing it. Perhaps the most likely transmembrane helix is the region between h2a and h2b, which on average is apolar at all heptad positions. If, as discussed previously, h2a and h2b form a continuous helix connected by a stutter in the heptad register, then it is possible that the center section of h2 could span the outer membrane as a helix. However, this sequence region is also a strong TM β-strand candidate (Figure 10(b)). The transmembrane topology of the TolC protein is currently being investigated using a novel site-specific proteolysis technique in conjunction with the theoretical methods presented here (Ehrmann *et al.*, 1997; M. Mondigler *et al.*, unpublished results).

HMM search for new PEP and OEP family members

The efflux pumps are polyphyletic

Final alignments of the two protein families were also used to assess how widespread this tripartite efflux system is among bacteria. Since the PEP and OEP alignments contain some gaps, it is more appropriate to search for new family members with profile hidden Markov models than with simple weight matrices which do not incorporate gaps as effectively. HMMs were built from the PEP/lipoyl motif (Figure 2), and from an alignment of PEPs comprising two lipoyl half-domains and the intervening coiled coil. Searching with these HMMs against the non-redundant (NR) protein database and completed bacterial genomes allowed several hypothetical proteins to be added to the PEP family (Table 3). These include four each from *Synechocystis* and *A. aeolicus*, indicating that cyanobacteria and two-membrane hydrogenobacteria have similar efflux mechanisms. A putative PEP could also be confirmed in the spirochete *Treponema pallidum*, and new PEPs were also identified in *Haemophilus influenzae*, *Helicobacter pylori*, *E. coli*, *Sphingomonas*, and *Coxiella burnetii*.

Surprisingly, a significant match ($E = 0.0018$) to the second PEP hidden Markov model was found in the genome of the Gram-positive bacterium *B. subtilis*. This hypothetical protein (g2635842) had strong matches to the M_N motif and to the coiled-

Table 3. PEP and OEP homologues identified by HMM search ($E < 0.01$), which are annotated as putative or hypothetical proteins in the SWISS-PROT and GenBank (NR) databases

	Species	GenBank ID
PEP	*Aquifex aeolicus*	2982989
		2983555
		2983499
		2984016
	Haemophilus influenzae	1573913
	Helicobacter pylori	2314660
	Escherichia coli	1789900
		1789637
		1787013
		1790024
		1790012
	Synechocystis sp.	1651721
		1001690
		1653651
		1652466
	Sphingomonas sp.	1314576
	Coxiella burnetii	3248942
	Bacillus subtilis	2635842
OEP	*Aquifex aeolicus*	2983286
		2983760
		2983554
		2983579
		2983607
	Borrelia burgdorferi	2688031
	Haemophilus influenzae	1574800
		1574302
	Helicobacter pylori	2314661
		2313728
	Synechocystis sp.	1653357

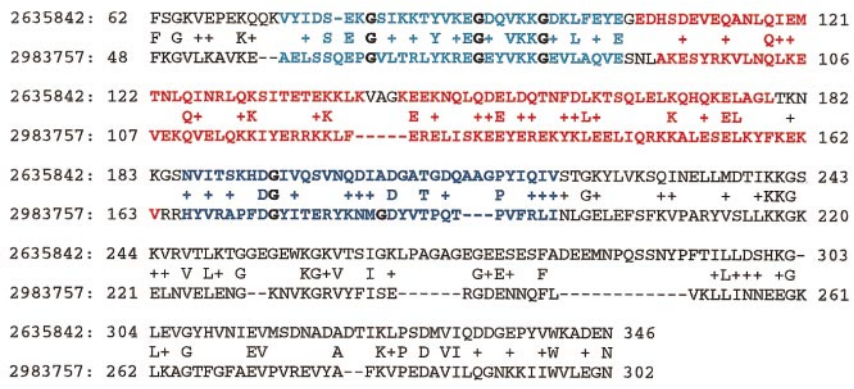

Figure 11. Alignment of a putative Gram-positive PEP homologue from *B. subtilis* (g2635842) with the most similar protein among known Gram-negative PEPs (g2983757 from *A. aeolicus*). The two proteins are 42% similar over 284 residues. Predicted M_N and M_C lipoyl-fold regions are in cyan and blue, respectively, with matches to the three conserved lipoyl glycine residues in black boldface. Residues with probability > 0.995 of coiled-coil formation (COILS window 28) are in red.

coil region, but the alignment to M_C was poor. The most similar previously identified PEP family member is from *A. aeolicus* (see Figure 11). Subsequent searches of the NR database using PSI-BLAST showed that homologues of g2635842 occur in *Streptococcus crista* (TptB), *Enterococcus faecalis* (BacG), *Lactobacillus sake* (SapE), and another in *B. subtilis* (g2633806). The three named proteins are putative components of ABC transport complexes (Axelsson & Holck, 1995; Correia *et al*., 1997; Tomita *et al*., 1997). All four of these proteins also have clear first lipoyl motifs (M_N) and are immediately followed by regions with a high probability (>0.9) of forming coiled coils. The sequences following the coiled-coil sections are consistent with second lipoyl (M_C) motifs, although many are missing one or more of the conserved glycine residues. It is therefore probable that Gram-positive bacteria have efflux pumps which contain PEP-like proteins with lipoyl-coiled-coil structural modules. It follows that the function of this module may not be solely to connect inner and outer membranes of two-membrane bacteria, but may also be to interact with the inner membrane transporter to guide passage across the peptidoglycan layer.

HMMs were also created for N-terminal and C-terminal alignments of OEPs, and for a combined alignment of the N and C-terminal tandem repeats. Searches with these HMMs identified new OEP family members in *Synechocystis, A. aeolicus*, *H. influenzae*, and *H. pylori*, as well as in *Borrelia burgdorferi*. No significant OEP or PEP matches were found in eukaryotic sequences, and as expected, no high-scoring OEP motifs were found in Gram-positive sequences.

Detection of database sequence errors

A final use of the alignment process and HMM search we have described is the detection of frameshift or other sequencing errors in database sequences. Several protein sequences were identified which strongly matched part of an OEP or PEP motif, but were missing another part of the motif. In many cases, when the nucleotide sequences for these genes were examined, the sequence match to the motif continued in another reading frame, indicating a probable error in the deposited sequence.

For example, a portion of the N-terminal OEP motif of Figure 7(a) is missing from the sequence for FusA of *Burkholderia cepacia* (Utsumi *et al*., 1991). Scanning the nucleotide sequence with the program BLASTX (NCBI) shows that the rest of the motif for FusA is found upstream in a different reading frame. Likewise, the deposited sequence for YohG of *E. coli* has its N terminus in the middle of the same OEP motif. Sequence similarity to other OEPs continues upstream of YohG into the sequence for hypothetical protein YohH, and includes the first conserved proline residue of the motif and a likely export signal sequence. Combining the two open reading frames, starting at the second methionine residue of YohH yields an OEP sequence which contains all motifs and aligns well with the family. Finally, the sequence reported for the outer membrane efflux protein AprF in *Pseudomonas fluorescens* (Liao & McCallus, 1998) is 42% identical with AprF from *P. aeruginosa* over 257 residues, but curiously has several 20-40 residue sections with nearly 90% identity. Examination of the nucleotide sequence reveals that the most likely protein sequence is scattered among the three 5′-3′ reading frames. Introduction of 13 frameshifts would yield a protein sequence which is 84% identical with *P. aeruginosa* AprF over a much longer sequence span (473 residues).

Conclusion

We have described a series of techniques for sequence analysis and structure prediction in highly divergent protein families, focusing on two families involved in multidrug resistance in Gram-negative bacteria. Building up multiple alignments from shorter motifs, we have made highly accurate whole-protein alignments of the PEP and OEP families, and used these alignments for structure prediction. For the PEP family, we identified a 2-fold symmetric region of high coiled-coil probability, flanked by two halves of a lipoyl domain. CD spectroscopy showed that lipoyl domain peptides from MexC preferentially adopt a β-conformation, and we suggest that the

intervening helical region could form an α-helical hairpin to unite the two lipoyl domain halves. The PEP helices, in a closed, antiparallel form, or in an open, extended form, could interact with the OEP periplasmic domain to create a *trans*-periplasmic channel.

For the OEP family, motifs were found containing conserved glycine, proline, aspartate, and hydrophobic residues likely to be important to the common structure and function of these proteins. The arrangement of the motifs suggested a tandem-repeat structure, which was confirmed by multiple alignment and by the demonstration that structural properties of the sequences, such as secondary structure and coiled-coil propensity, are also tandemly repeated. A computational method (MAPF) was developed to automate the identification of conserved structural features of protein families. MAPF was first applied to the prediction of transmembrane β-strands in porins and shown to be an improvement over existing methods. Application of MAPF to the OEP family revealed regions of the OEP alignment likely to form transmembrane β-strands and coiled coils. These observations led to the prediction that OEPs constitute a structural class of proteins unlike outer membrane proteins of known structure, in contrast with previous predictions. By searching the protein database with hidden Markov models, this class was shown to be polyphyletic, including representatives from the cyanobacteria and hydrogenobacteria. The PEP family was also extended to these species and shown to include homologues in several species of Gram-positive bacteria. We hope these insights into the structural features of bacterial efflux pumps will encourage further research into overcoming this form of multidrug resistance in bacteria. It is also our hope that these or similar methods for generating accurate multiple alignments and making maximum use of their inherent structural information will become a routine complement to protein experiments and to the annotation of new genomes.

Materials and Methods

Motif detection

Starting sets of known PEP and OEP family members were obtained from the literature (Dinh *et al.*, 1994; Létoffé *et al.*, 1996; Paulsen *et al.*, 1996, 1997). The PURGE procedure (Neuwald *et al.*, 1995) with threshold score 200 was then used to remove highly similar sequences in order to avoid bias. To estimate the number of significant motifs and their sequence lengths, the ASSET algorithm (Neuwald & Green, 1994) was used. Output motif alignments were compared as weight matrices against protein sequence databases (Bairoch & Boeckmann, 1991; Bleasby & Wootton, 1990) with the SCAN algorithm (Neuwald *et al.*, 1995) to identify family members overlooked in the starting set. These were added, and the procedure repeated with PURGE threshold 150. Once the approximate length of the motif was established by ASSET, the Gibbs Motif Sampler (Neuwald *et al.*, 1995) was used to refine the alignment and determine the number of occurrences per input sequence. To further address the question of how many motif sites occur in each protein, the Gibbs Site Sampler (Lawrence *et al.*, 1993) was used to look for fixed numbers of sites in each sequence. The Gibbs Site Sampler was particularly useful for identifying regions where gaps were required, since it could extract lower-scoring motif copies which had insertions or deletions relative to the main motif. A computer program which groups together several Gibbs sampling techniques is now publicly available (Neuwald *et al.*, 1997), but was not used for this work.

Coiled coils were predicted using the COILS algorithm (Lupas, 1996b; Lupas *et al.*, 1991), which was run on all studied sequences, using 21 and 28-residue windows and the MTIDK matrix. Results were checked for consistency with predictions with increased weight (2.5-fold) on the *a* and *d* positions of the heptad repeat, and with predictions of the PAIRCOIL program (Berger *et al.*, 1995). The MAPF program was used to align and visualize coiled-coil predictions with respect to multiple alignments of PEP and OEP families.

Multiple alignment and detection of remote homologues

To build protein alignments longer than the identified motifs and which included gaps, we used the iterated BLAST algorithm PSI-BLAST as a first step (Altschul *et al.*, 1997). Starting from single PEP and OEP sequences, multiple alignments were constructed by iterative search of the NR protein database. Parameters were adjusted to avoid inclusion of false positives (typically myosin and other coiled coil containing proteins) in the search matrices, although this required exclusion of many true positives. True and false positives were judged primarily by the presence or absence of the signature motifs identified by Gibbs sampling. A few gaps were removed from the N and C termini of the PSI-BLAST output alignments and from poorly aligned, highly divergent regions using the program SEAVIEW (Galtier *et al.*, 1996). The alignments were then converted to hidden Markov models and the sequences realigned using either HMMER 1.8.4 (for maltoporin, *R. blastica* porin and OmpF alignments) or HMMER 2.0 (for all other alignments and HMM searches; Eddy, 1996, http://hmmer.wustl.edu). Finally, the refined motif alignments were scanned against the database using HMM representations. Matches were considered significant if they had estimated *E*-values <0.01, were aligned across all motif regions, and had coiled-coil patterns similar to those of known family members. Highly similar sequences (>90% identity) were removed from the multiple alignments before structure prediction analysis. Alignments of OEP and PEP families are available at http://arep.med.harvard.edu.

Phylogeny

Phylogenetic analyses were performed using Phylo_win (Galtier *et al.*, 1996). The neighbor-joining method was used with 200 bootstrap replicates, shuffling the sequence order before each replicate. PAM distance and percentage identity were used on separate trials to cluster 60 PEP motifs from 30 protein sequences, incorporating 38 alignment positions. The two distance

measures gave nearly identical results. For the OEP calculation, 68 motifs (34 protein sequences) of 174 alignment positions were compared using percentage identity as the distance measure.

CD spectroscopy

N-terminal and C-terminal peptides of MexC protein (>95% pure) were obtained from the Biological Chemistry and Molecular Pharmacology Biopolymers Facility at Harvard Medical School. Peptide samples of concentration 0.1-2 mg/ml in 20 mM sodium phosphate buffer (pH 7.0), with a range of TFE concentrations from 0-60% (v/v), were scanned from 185-240 nm (step size 0.5 nm, averaging time one second) in an Aviv 62DS CD spectrometer at 25°C. All samples were passed through a Millipore 0.22 μm spin column before measurement. Cells of path-length 1.0 mm and 0.1 mm were used, and the buffer spectrum was subtracted from the data, averaged over five to 15 scans. Secondary structure estimates were derived from the spectra using the ridge regression algorithm by Provencher & Glöckner (1981), with a 16 protein reference set. Results were checked for consistency with reference sets including poly(L-glutamate) and denatured proteins, since the inclusion of denatured proteins can assist in discriminating between sheet and random coil (Venyaminov *et al*., 1993).

β-Strand prediction

Structure prediction programs in C or Perl were obtained from cited authors or downloaded from public repositories and organized into Perl subroutines. New code was written for several prediction functions. Test-set alignments for the porins were created using the methods described above for the PEP and OEP families. These alignments were not corrected using knowledge of the solved structures in order to provide an accurate test of the method for families of unknown structure. Each porin family served as the test set for training sets composed of the other two families. Contributions from each structural predictor to the combination function were weighted by the correlation coefficient between the prediction and the actual positions of TM β-strands, averaged over the training sets. All three porin families were used as the training set for the OEP combination function. The β-sheet periodicity angle used to calculate hydrophobic moment (amphipathicity) was 160°. A window of seven residues was used to smooth the beta, hydrophobicity, amphipathicity, and membrane preference functions. Beta and membrane preference functions were scaled to have minima at 0 and maxima at 1. Similarity was defined according to the following residue groups: RKH, ED, AST, G, C, P, NQ, ILVM, WFY. The turn prediction threshold was 0.000075 for the product of four consecutive turn probability values (see Chou & Fasman, 1978). Contributions of each protein to the average function for each predictor were divided by the average percentage identity of that protein to the others in the alignment in order to counteract the bias from similar proteins. Output plots for each selected predictor, as well as a plot showing them all in combination, were generated automatically by MAPF using the program GNUPLOT. Computational analysis and molecular modeling were performed on a Silicon Graphics Octane workstation. MAPF code for β-strand prediction is available at http://arep.med.harvard.edu.

Acknowledgments

We gratefully acknowledge M. Ehrmann, T. Schirmer, K. Diederichs, and J. Fischbarg for helpful correspondence, H. Wendt and C. Elkin for assistance with CD spectrometry, C. Dahl for peptide synthesis, and members of the Church lab and M. Ehrmann for comments on the manuscript. J.M.J. was supported in part by a grant from the National Science Foundation. Additional support to J.M.J. and G.M.C. was provided by a grant from Hoechst Marion Roussel.

References

Akatsuka, H., Binet, R., Kawai, E., Wandersman, C. & Omori, K. (1997). Lipase secretion by bacterial ATP-binding cassette exporters: molecular recognition of the LipBCD, PrtDEF, and HasDEF exporters. *J. Bacteriol.* **179**, 4754-4760.

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

Aono, R., Tsukagoshi, N. & Yamamoto, M. (1998). Involvement of outer membrane protein TolC, a possible member of the *mar-sox* regulon, in maintenance and improvement of organic solvent tolerance of *Escherichia coli*. *J. Bacteriol.* **180**, 938-944.

Athappilly, F. K. & Hendrickson, W. A. (1995). Structure of the biotinyl domain of acetyl-coenzyme A carboxylase determined by MAD phasing. *Structure,* **3**, 1407-1419.

Attwood, T. K., Beck, M. E., Bleasby, A. J. & Parry-Smith, D. J. (1994). PRINTS—a database of protein motif fingerprints. *Nucl. Acids Res.* **22**, 3590-3596.

Axelsson, L. & Holck, A. (1995). The genes involved in production of and immunity to sakacin A, a bacteriocin from *Lactobacillus sake* Lb706. *J. Bacteriol.* **177**, 2125-2137.

Bairoch, A. & Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **19**, 2247-2249.

Bairoch, A., Bucher, P. & Hofmann, K. (1997). The PROSITE database, its status in 1997. *Nucl. Acids Res.* **25**, 217-221.

Banner, D. W., Kokkinidis, M. & Tsernoglou, D. (1987). Structure of the ColE1 Rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196**, 657-675.

Barton, G. J., Newman, R. H., Freemont, P. S. & Crumpton, M. J. (1991). Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur. J. Biochem.* **198**, 749-760.

Bayer, M. E. (1991). Zones of membrane adhesion in the cryofixed envelope of *Escherichia coli*. *J. Struct. Biol.* **107**, 268-280.

Behrends, H. W., Folkers, G. & Beck-Sickinger, A. G. (1997). A new approach to secondary structure evaluation: secondary structure prediction of porcine kinase and yeast guanylate kinase by CD spectroscopy of overlapping peptide segments. *Biopolymers,* **41**, 213-231.

Benson, D., Lipman, D. J. & Ostell, J. (1993). GenBank. *Nucl. Acids Res.* **21**, 2963-2965.

Benz, R., Maier, E. & Gentschev, I. (1993). TolC of *Escherichia coli* functions as an outer membrane channel. *Int. J. Med. Microbiol. Virol. Parasitol. Infect. Dis.* **278**, 187-196.

Berg, A. & de Kok, A. (1997). 2-oxo acid dehydrogenase multienzyme complexes. The central role of the lipoyl domain. *Biol. Chem.* **378**, 617-634.

Berg, A., Vervoort, J. & de Kok, A. (1996). Solution structure of the lipoyl domain of the 2-oxoglutarate dehydrogenase complex from *Azotobacter vinelandii*. *J. Mol. Biol.* **261**, 432-442.

Berg, A., Vervoort, J. & de Kok, A. (1997). Three-dimensional structure in solution of the N-terminal lipoyl domain of the pyruvate dehydrogenase complex from *Azotobacter vinelandii*. *Eur. J. Biochem,* **244**, 352-360.

Berger, B., Wilson, D. B., Wolf, E., Tonchev, T., Milla, M. & Kim, P. S. (1995). Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl Acad. Sci. USA,* **92**, 8259-8263.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319-324.

Binet, R. & Wandersman, C. (1995). Protein secretion by hybrid bacterial ABC-transporters: specific functions of the membrane ATPase and the membrane fusion protein. *EMBO J.* **14**, 2298-2306.

Binet, R., Létoffé, S., Ghigo, J. M., Delepelaire, P. & Wandersman, C. (1997). Protein secretion by Gram-negative bacterial ABC exporters—a review. *Gene,* **192**, 7-11.

Bleasby, A. J. & Wootton, J. C. (1990). Construction of validated, non-redundant composite protein-sequence databases. *Protein Eng.* **3**, 153-159.

Brown, J. H., Cohen, C. & Parry, D. A. D. (1996). Heptad breaks in α-helical coiled coils: stutters and stammers. *Proteins: Struct. Funct. Genet.* **26**, 134-145.

Buchanan, S. K., Smith, B. S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., van der Helm, D. & Diesenhofer, J. (1999). Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. *Nature Struct. Biol.* **6**, 56-63.

Chothia, C. & Murzin, A. G. (1993). New folds for all-β proteins. *Structure,* **1**, 217-222.

Chou, P. Y. & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annu. Rev. Biochem.* **47**, 251-276.

Correia, F. F., Lamont, R., Bayer, M., Rosan, B. & DiRienzo, J. M. (1997). Cloning and sequencing of a mutated locus that affects fimbrial tuft organization and corncob formation in *Streptococcus crista*. *Int. J. Oral Biol.* **22**, 241-248.

Cowan, S. W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R. A., Jansonius, J. N. & Rosenbusch, J. P. (1992). Crystal structures explain function properties of two *E. coli* porins. *Nature,* **358**, 727-733.

Crawford, I. P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the α subunit of tryptophan synthetase. *Proteins: Struct. Funct. Genet.* **2**, 118-129.

Cusack, S., Berthet-Colominas, C., Härtlein, M., Nassar, N. & Leberman, R. (1990). A second class of synthetase structure revealed by X-ray analysis of *Escherichia coli* seryl-tRNA synthetase at 2.5 Å. *Nature,* **347**, 249-255.

Dardel, F., Davis, A. L., Laue, E. D. & Perham, R. N. (1993). Three-dimensional structure of the lipoyl domain from *Bacillus stearothermophilus* pyruvate dehydrogenase multienzyme complex. *J. Mol. Biol.* **229**, 1037-1048.

Diederichs, K., Freigang, J., Umhau, S., Zeth, K. & Breed, J. (1998). Prediction by a neural network of outer membrane β-strand topology. *Protein Sci.* **7**, 2413-2420.

Dinh, T., Paulsen, I. T. & Saier, M. H., Jr. (1994). A family of extracytoplasmic proteins that allow transport of large molecules across the outer membrane of gram-negative bacteria. *J. Bacteriol.* **176**, 3825-3831.

Eddy, S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.* **6**, 361-365.

Ehrmann, M., Bolek, P., Mondigler, M., Boyd, D. & Lange, R. (1997). TnTIN and TnTAP: Mini-transposons for site-specific proteolysis *in vivo*. *Proc. Natl Acad. Sci. USA,* **94**, 13111-13115.

Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA,* **82**, 140-144.

Engel, A. M., Cejka, Z., Lupas, A., Lottspeich, F. & Baumeister, W. (1992). Isolation and cloning of Omp alpha, a coiled-coil protein spanning the periplasmic space of the ancestral eubacterium *Thermotoga maritima*. *EMBO J.* **11**, 4369-4378.

Fath, M. J., Skvirsky, R. C. & Kolter, R. (1991). Functional complementation between bacterial MDR-like export systems: colicin V, alpha-hemolysin and *Erwinia* protease. *J. Bacteriol.* **173**, 7549-7556.

Felmlee, T., Pellett, S. & Welch, R. A. (1985). Nucleotide sequence of an *Escherichia coli* chromosomal hemolysin. *J. Bacteriol.* **163**, 94-105.

Ferguson, A. D., Hofmann, E., Coulton, J. W., Diederichs, K. & Welte, W. (1998). Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide. *Science,* **282**, 2215-2220.

Fischbarg, J., Cheung, M., Li, J., Iserovich, P., Czegledy, F., Kuang, K. & Garner, M. (1994). Are most transporters and channels beta barrels? *Mol. Cell. Biochem.* **140**, 147-162.

Fujinaga, M., Berthet-Colominas, C., Yaremchuk, A. D., Tukalo, M. A. & Cusack, S. (1993). Refined crystal structure of the seryl-tRNA synthetase from *Thermus thermophilus* at 2.5 Å resolution. *J. Mol. Biol.* **234**, 222-233.

Galtier, N., Gouy, M. & C, G. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* **12**, 543-548.

Gernert, K. M., Surles, M. C., Labean, T. H., Richardson, J. S. & Richardson, D. C. (1995). The Alacoil: a very tight, antiparallel coiled-coil of helices. *Protein Sci.* **4**, 2252-2260.

Gray, L., Baker, K., Mackman, N., Haigh, R. & Holland, I. B. (1989). A novel C-terminal signal sequence targets *E. coli* hemolysin directly to the medium. *Mol. Gen. Genet.* **205**, 127-133.

Green, J. D. F., Laue, E. D., Perham, R. N., Ali, S. T. & Guest, J. R. (1995). Three-dimensional structure of a lipoyl domain from the dihydrolipoyl acetyltransferase multienzyme complex of *Escherichia coli*. *J. Mol. Biol.* **248**, 328-343.

Griffith, J. K., Baker, M. E., Rouch, D. A., Page, M. G. P., Skurray, R. A., Paulsen, I. T., Chater, K. F., Baldwin, S. A. & Henderson, P. J. F. (1992). Membrane transport proteins: implications of sequence comparisons. *Curr. Opin. Cell Biol.* **4**, 684-695.

Gromiha, M. M. & Ponnuswamy, P. K. (1993). Prediction of transmembrane β-strands from hydrophobic characteristics of proteins. *Int. J. Pept. Protein Res.* **42**, 420-431.

Gromiha, M. M., Majumdar, R. & Ponnuswamy, P. K. (1997). Identification of membrane spanning β strands in bacterial porins. *Protein Eng.* **10**, 497-500.

Gross, R. (1995). Domain structure in the outer membrane transporter protein CyaE of *Bordella pertussis*. *Mol. Microbiol.* **17**, 1219-1220.

Hofmann, K. & Stoffel, W. (1993). TMbase—a database of membrane spanning protein segments. *Biol. Chem. Hoppe-Seyler,* **374**, 166.

Holland, I. B., Kenny, B. & Blight, M. (1990). Haemolysin secretion from *E. coli*. *Biochimie,* **72**, 131-141.

Hunt, J. F., Earnest, T. N., Bousche, O., Kalghatgi, K., Reilly, K., Horvath, C., Rothschild, K. J. & Engelman, D. M. (1997). A biophysical study of integral membrane protein folding. *Biochemistry,* **36**, 15156-15176.

Hwang, J., Zhong, X. & Tai, P. C. (1997). Interactions of dedicated export membrane proteins of the colicin V secretion system: CvaA, a member of the membrane fusion protein family, interacts with CvaB and TolC. *J. Bacteriol.* **179**, 6264-6270.

Jeanteur, D., Lakey, J. H. & Pattus, F. (1991). The bacterial porin superfamily: sequence alignment and structure prediction. *Mol. Microbiol.* **5**, 2153-2164.

Köhler, T., Michéa-Hamzehpour, M., Henze, U., Gotoh, N., Curty, L. K. & Pechére, J.-C. (1997). Characterization of *mexE-mexF-oprN*, a positively regulated multidrug efflux system of *Pseudomonas aeruginosa*. *Mol. Microbiol.* **23**, 345-354.

Koronakis, V., Koronakis, E. & Hughes, C. (1989). Isolation and analysis of the C-terminal signal directing export of *Escherichia coli* hemolysin protein across both bacterial membranes. *EMBO J.* **8**, 595-605.

Koronakis, V., Li, J., Koronakis, E. & Stauffer, K. (1997). Structure of TolC, the outer membrane component of the bacterial type I efflux system, derived from two-dimensional crystals. *Mol. Microbiol.* **23**, 617-626.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **26**, 283-291.

Kreusch, A., Neubüser, N., Schlitz, E., Weckesser, J. & Schulz, G. E. (1994). Structure of the membrane channel porin from *Rhodopseudomonas blastica* at 2.0 Å resolution. *Protein Sci.* **3**, 58-63.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105-132.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science,* **262**, 208-214.

Létoffé, S., Delepelaire, P. & Wandersman, C. (1996). Protein secretion in Gram-negative bacteria: assembly of the three components of ABC protein-mediated exporter is ordered and promoted by substrate binding. *EMBO J.* **15**, 5804-5811.

Levengood, S. K., Beyer, W. F., Jr & Webster, R. E. (1991). TolA: a membrane protein involved in colicin uptake contains an extended helical region. *J. Bacteriol.* **88**, 5939-5943.

Levin, J. M., Pascarella, S., Argos, P. & Garnier, J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6**, 849-854.

Liao, C.-H. & McCallus, D. E. (1998). Biochemical and genetic characterization of an extracellular protease from *Pseudomonas fluorescens*. *Appl. Env. Microbiol.* **64**, 914-921.

Lupas, A. (1996a). Coiled coils: new structures and new functions. *Trends Biochem. Sci.* **21**, 375-382.

Lupas, A. (1996b). Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**, 513-525.

Lupas, A., Van Dyke, M. & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science,* **252**, 1162-1164.

Lupas, A., Müller, S., Goldie, K., Engel, A. M., Engel, A. & Baumeister, W. (1995). Model structure of the Ompα rod, a parallel four-stranded coiled coil from the hyperthermophilic eubacterium *Thermotoga maritima*. *J. Mol. Biol.* **248**, 180-189.

Ma, D., Cook, D. N., Alberti, M., Pon, N. G., Nikaido, H. & Hearst, J. E. (1993). Molecular cloning and characterization of *acrA* and *acrE* genes of *Escherichia coli*. *J. Bacteriol.* **175**, 6299-6313.

Marger, M. D. & Saier, M. H., Jr. (1993). A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport. *Trends Biochem. Sci.* **18**, 13-20.

McLachlan, A. D. (1978). The double helix coiled coil structure of murein lipoprotein from *Escherichia coli*. *J. Mol. Biol.* **122**, 493-506.

Michéa-Hamzehpour, M., Pechère, J.-C., Plésiat, P. & Köhler, T. (1995). OprK and OprM define two genetically distinct multidrug efflux systems in *Pseudomonas aeruginosa*. *Antimicrob. Agents Chemother.* **39**, 2392-2396.

Monera, O. D., Zhou, N. E., Lavigne, P., Kay, C. M. & Hodges, R. S. (1996). Formation of parallel and antiparallel coiled-coils controlled by the relative positions of alanine residues in the hydrophobic core. *J. Biol. Chem.* **271**, 3995-4001.

Morona, R., Manning, P. A. & Reeves, P. (1983). Identification and characterization of the TolC protein, an outer membrane protein from *Escherichia coli*. *J. Bacteriol.* **153**, 693-699.

Neuwald, A. F. & Green, P. (1994). Detecting patterns in protein sequences. *J. Mol. Biol.* **239**, 698-712.

Neuwald, A. F., Liu, J. S. & Lawrence, C. E. (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* **4**, 1618-1632.

Neuwald, A. F., Liu, J. S., Lipman, D. J. & Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nucl. Acids Res.* **25**, 1665-1677.

Nikaido, H. (1998). Antibiotic resistance caused by gram-negative multidrug efflux pumps. *Clin. Infect. Dis.* **27**, S32-S41.

O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science,* **254**, 539-544.

Paul, C. & Rosenbusch, J. P. (1985). Folding patterns of porin and bacteriorhodopsin. *EMBO J.* **4**, 1593-1597.

Paulsen, I. T., Brown, M. H. & Skurray, R. A. (1996). Proton-dependent multidrug efflux systems. *Microbiol. Rev.* **60**, 575-608.

Paulsen, I. T., Park, J. H., Choi, P. S. & Saier, M. H., Jr. (1997). A family of Gram-negative bacterial outer membrane factors that function in the export of proteins, carbohydrates, drugs, and heavy metals from

Gram-negative bacteria. *FEMS Microbiol. Letters,* **156**, 1-8.

Pautsch, A. & Schulz, G. E. (1998). Structure of the outer membrane protein A transmembrane domain. *Nature Struct. Biol.* **5**, 1013-1017.

Pimenta, A., Blight, M., Clarke, D. & Holland, I. B. (1996). The Gram-negative cell envelope 'springs' to life: coiled-coil *trans*-envelope proteins. *Mol. Microbiol.* **19**, 643-645.

Ponnuswamy, P. K. & Gromiha, M. M. (1993). Prediction of transmembrane helices from hydrophobic characteristics of proteins. *Int. J. Pept. Protein Res.* **42**, 326-341.

Poole, K., Heinrichs, D. E. & Neshat, S. (1993). Cloning and sequence analysis of an EnvCD homologue in *Pseudomonas aeruginosa*: regulation by iron and possible involvement in the secretion of the siderophore pyoveridine. *Mol. Microbiol.* **10**, 529-544.

Poole, K., Gotoh, N., Tsujimoto, H., Zhao, Q., Wada, A., Yamasaki, T., Neshat, S., Yamagishi, J., Li, X.-Z. & Nishino, T. (1996). Overexpression of the *mexC-mexD-OprJ* efflux operon in *nfxB*-type multidrug-resistant strains of *Pseudomonas aeruginosa*. *Mol. Microbiol.* **21**, 713-724.

Provencher, S. W. & Glöckner, J. (1981). Estimation of globular protein secondary structure from circular dichroism. *Biochemistry,* **20**, 33-37.

Rost, B. & Sander, C. (1993). Prediction of secondary structure at better than 70 % accuracy. *J. Mol. Biol.* **232**, 584-599.

Saier, M. H., Jr, Tam, R., Reizer, A. & Reizer, J. (1994). Two novel families of bacterial membrane proteins concerned with nodulation, cell division and transport. *Mol. Microbiol.* **11**, 841-847.

Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11-15.

Schirmer, T., Keller, T. A., Wang, Y. F. & Rosenbusch, J. P. (1995). Structural basis for sugar translocation through maltoporin channels at 3.1 A resolution. *Science,* **267**, 512-514.

Schlör, S., Schmidt, A., Maier, E., Benz, R., Goebel, W. & Gentschev, I. (1997). *In vivo* and *in vitro* studies on interactions between the components of the hemolysin (HlyA) secretion machinery of *Escherichia coli*. *Mol. Gen. Genet.* **256**, 306-319.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097-6100.

Schülein, R., Gentschev, I. & Mollenkopf, H.-J. (1992). A topological model for the haemolysin translocator protein HlyD. *Mol. Gen. Genet.* **234**, 155-163.

Scott, W. G., Milligan, D. L., Milburn, M. V., Prive, G. G., Yeh, J., Koshland, D. E., Jr. & Kim, S. H. (1993). Refined structures of the ligand-binding domain of the aspartate receptor from *Salmonella typhimurium*. *J. Mol. Biol.* **232**, 555-573.

Seiffer, D., Klein, J. R. & Plapp, R. (1993). EnvC, a new lipoprotein of the cytoplasmic membrane of *Escherichia coli*. *FEMS Microbiol. Letters,* **107**, 175-178.

Sonnhammer, E. L., Eddy, S. R. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405-420.

Spencer, M. E., Darlison, M. G., Stephens, P. E. & Duckenfield, I. K. (1984). Nucleotide sequence of the *sucB* gene encoding the dihydrolipoamide succinyltransferase of *Escherichia coli* K12 and homology with the corresponding acetyltransferase. *Eur. J. Biochem.* **141**, 361-374.

Stoorvogel, J., van Bussel, M. J. A. W. M., Tommassen, J. & van de Klundert, J. A. M. (1991). Molecular characterization of an *Enterobacter cloacae* outer membrane protein (OmpX). *J. Bacteriol.* **173**, 156-160.

Thanabalu, T., Koronakis, E., Hughes, C. & Koronakis, V. (1998). Substrate-induced assembly of a contiguous channel for protein export from *E. coli*: reversible bridging of an inner-membrane translocase to an outer membrane exit pore. *EMBO J.* **17**, 6487-6496.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Tomita, H., Fujimoto, S., Tanimoto, K. & Ike, Y. (1997). Cloning and genetic and sequence analyses of the bacteriocin 21 determinant encoded on the *Enterococcus faecalis* pheromone-responsive conjugative plasmid pPD1. *J. Bacteriol.* **179**, 7843-7855.

Utsumi, R., Yagi, T., Katayama, S., Katsuragi, K., Tachibana, K., Toyoda, H., Ouchi, S., Obata, K., Shibano, Y. & Noda, M. (1991). Molecular cloning and characterization of the fusaric acid-resistance gene from *Pseudomonas cepacia*. *Agric. Biol. Chem.* **55**, 1913-1918.

Venyaminov, S. Y., Baikalov, I. A., Shen, Z. M., Wu, C. S. & Yang, J. T. (1993). Circular dichroic analysis of denatured proteins: inclusion of denatured proteins in the reference set. *Anal. Biochem.* **214**, 17-24.

Vogel, H. & Jähnig, F. (1986). Models for the structure of outer-membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction methods. *J. Mol. Biol.* **190**, 191-199.

von Heijne, G. (1997). Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* **66**, 113-139.

Wandersman, C. & Delepelaire, P. (1990). TolC, an *Escherichia coli* outer membrane protein required for hemolysin secretion. *Proc. Natl Acad. Sci. USA* **87**, 4776-4780.

Webster, R. E. (1991). The *tol* gene products and the import of macromolecules into *Escherichia coli*. *Mol. Microbiol.* **5**, 1005-1011.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957-961.

Edited by G. von Heijne

(Received 5 November 1998; received in revised form 12 February 1999; accepted 12 February 1999)