Published online May 11, 2004

2566–2577 Nucleic Acids Research, 2004, Vol. 32, No. 8
DOI: 10.1093/nar/gkh580

Predicting transmembrane beta-barrels in proteomes

Henry R. Bigelow[1,*], Donald S. Petrey[2], Jinfeng Liu[1,3,4], Dariusz Przybylski[1,5] and Burkhard Rost[1,3,6]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, [2]Howard Hughes Medical Institute and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032, [3]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, [4]Department of Pharmacology, Columbia University, 630 West 168th Street, New York, NY 10032, [5]Department of Physics, Columbia University, 538 West 120th Street, New York, NY 10027 and [6]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA

Received January 21, 2004; Revised March 14, 2004; Accepted April 12, 2004

ABSTRACT

Very few methods address the problem of predicting beta-barrel membrane proteins directly from sequence. One reason is that only very few high-resolution structures for transmembrane beta-barrel (TMB) proteins have been determined thus far. Here we introduced the design, statistics and results of a novel profile-based hidden Markov model for the prediction and discrimination of TMBs. The method carefully attempts to avoid over-fitting the sparse experimental data. While our model training and scoring procedures were very similar to a recently published work, the architecture and structure-based labelling were significantly different. In particular, we introduced a new definition of beta-hairpin motifs, explicit state modelling of transmembrane strands, and a log-odds whole-protein discrimination score. The resulting method reached an overall four-state (up-, down-strand, periplasmic-, outer-loop) accuracy as high as 86%. Furthermore, accurately discriminated TMB from non-TMB proteins (45% coverage at 100% accuracy). This high precision enabled the application to 72 entirely sequenced Gram-negative bacteria. We found over 164 previously uncharacterized TMB proteins at high confidence. Database searches did not implicate any of these proteins with membranes. We challenge that the vast majority of our 164 predictions will eventually be verified experimentally. All proteome predictions and the PROFtmb prediction method are available at http://www.rostlab.org/services/PROFtmb/.

INTRODUCTION

Beta-strand membrane proteins have unique structural features

Transmembrane beta-barrel (TMB) proteins reside in the outer membranes of gram-negative bacteria, mitochondria and chloroplasts. The functions are diverse, including bacterial adhesion (OmpX) (1), structural integrity of the cell wall (OmpA) (2), colicin release (Phospholipase A), general (OmpF, PhoE) and substrate-specific (LamB, ScrY) diffusion of hydrophilic molecules and sugars, respectively (3–5), and transport of large iron-siderophore complexes and vitamin B12 (FhuA, FepA) (6,7). The first evidence for a TMB was electron microscope data of phoE porin in 1989, followed by a high-resolution structure (8,9). To date, 56 structures have been solved, all from Gram-negative bacteria. They cluster into 11 sequence-structure families (Materials and Methods). All known single-chain TMB proteins are described by a simple 'grammar' (1,10) (Fig. 1): N-terminal signal sequence, M repeats of (upward strand, extra-cellular loop, downward strand, periplasmic hairpin), and possibly a C-terminal region. The number of beta-strands ranges from 8 to 22 between different types of TMBs. Some single-chain barrels exist as monomers (FhuA, FepA, phospholipase A, OmpX, OmpA), others as trimeric porins (Sucrose Specific Porin, Maltoporin, OmpF Matrix Porin, 'porin' from *Rhodopseudomonas blastica*). Multi-chain TMBs are composed of a repeated group with some number of beta hairpins from each monomer. For example, the TolC receptor is a trimer in which each monomer contributes two adjacent hairpins (four transmembrane strands) to the barrel, totalling to 12 strands. Alpha haemolysin is a heptamer in which each monomer contributes a single strand–hairpin–strand motif to form a beta-barrel with 14 strands. Since the grammar of these multimeric barrels differs completely from that of the single-chain barrels, we ignored them for this study.

*To whom correspondence should be addressed. Tel: +1 212 305 4018; Fax: +1 212 305 7932; Email: bigelow@cubic.bioc.columbia.edu
Correspondence may also be addressed to Burkhard Rost. Tel: +1 212 305 4018; Fax: +1 212 305 7932; Email: rost@cubic.bioc.columbia.edu

 © Oxford University Press 2004; all rights reserved

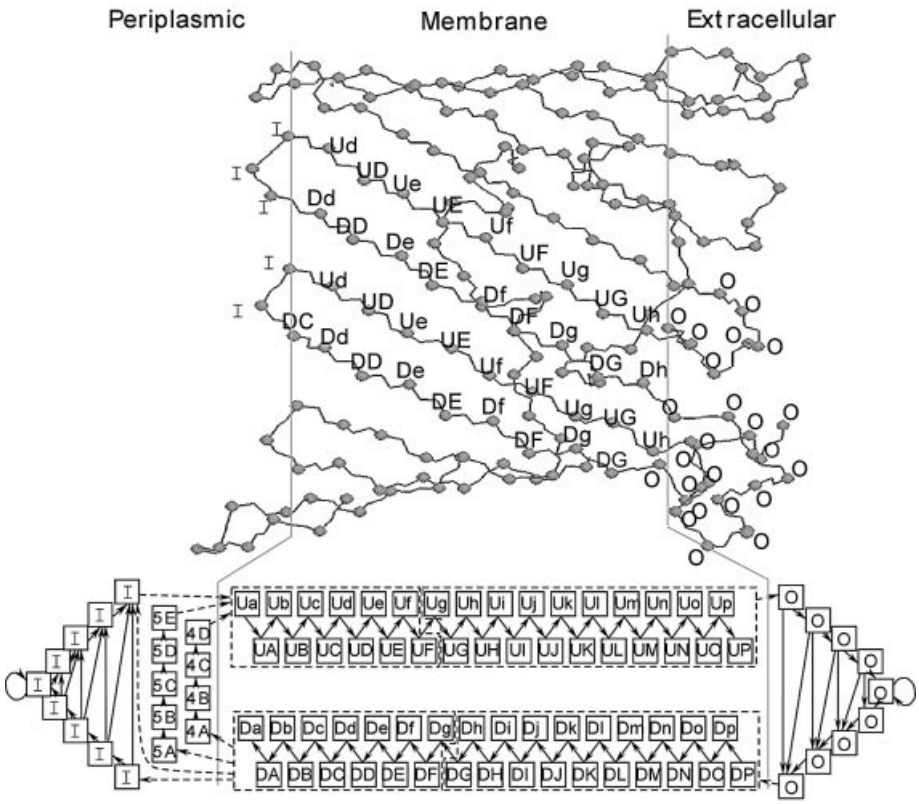

Figure 1. Model architecture and structure-based labelling. Dashed rectangles/arrows are shorthand for showing an arrow connection for each node in the rectangle. Strands are shown to depict the alternating lipid/water environments of the residues. State labels indicate two things: (i) the sequence label for which the state is valid and (ii) the set of emission parameters. Note that states sharing a label also share the same set of emission parameters. The structure-based labelling is illustrated for *R.blastica* porin [PDB code 1prn (49)].

Diversity of function

The functional roles of beta-barrel proteins are diverse, including bacterial adhesion, selective import of sugars, voltage-gated import of ions, ligand-gated import of iron siderophore complexes, phospholipase and protease activity, and export of proteins or drugs. The eight beta-strands of OmpX from *Escherichia coli* protrude several residues into the external medium and bind any foreign protein exhibiting external beta-strands (11). The eight strands of OmpA are connected by long, mobile external loops and serve as docking points for bacteriophages. Two known barrels have enzymatic activity, namely the two *E.coli* proteins OmpT and OmpLA. OmpT with 10 strands is a protease whose active centre is facing the extra-cellular side, and is implicated in the pathogenicity of bacteria (12). The proteolytic mechanism depends on the binding of one lipopolysaccharide molecule. Commonly used protease inhibitors only weakly affect the activity. It is hoped that the structure can be used as a template for development of antimicrobial compounds. OmpLA is a 12-stranded phospholipase for which seven different high-resolution structures exist; it is complexed variously with calcium ion among other compounds. Dimerization and calcium binding is required for activity, which results in the deacylation of lipopolysaccharides in the external leaflet of the outer membrane, producing holes, and ultimately allowing the secretion of colicins and virulence factors (13,14). Among the transporters are Maltoporin and Sucrose Specific Porin, TolC channel-tunnel, FepA and FhuA active iron importers, and PhoE and OmpF, anion and cation selective pores, respectively. Maltoporin has 18 strands that selectively transport glucose. The energetics of binding specificity have been examined in detail, revealing an optimal spacing of non-polar and polar residues facing the pore, resulting in smooth gliding of the solute (10). Sucrose Specific Porin is a close homologue of Maltoporin with a much higher permeability, accounted for by differences in composition of pore-facing residues (3). TolC facilitates drug efflux, leading to drug resistance, and the secretion of alpha-haemolysin; this targets cell lysis. The barrel is formed from a homo-trimer (three chains), each contributing four beta-strands; long alpha helices form a periplasmic tunnel, bridging to the inner membrane component of the transport system (15). FhuA and FepA, both with 22 strands, facilitate the import of ferrichrome. FhuA (ferric hydroxamate uptake) employs a signal transduction upon ferrichrome binding to an extra-cellular 'gating loop' L4, through conformational changes (not in L4). At least one antibiotic is known to be imported through FhuA (albomycin); this mechanism is hoped to be exploited for delivery of other antibiotics (16). Finally, PhoE and OmpF are 16-stranded, trimeric porins with a water-filled channel in each monomer. The selectivity filter is a restriction at half the height of the barrel with a strong transverse electric field. Extensive mutational and electrophysiological studies have revealed the roles of individual charged residues in each of these proteins affecting the critical gating potential (17,18). The outer membranes of mitochondria and chloroplasts were

hypothesized according to the endosymbiosis theory to contain only barrels. Studies on individual proteins revealed, however, that they also contain alpha-helical proteins. At the end of 2003, there are still no high-resolution structures available for any mitochondrial or chloroplast TMB, presumably due to the lower stability of proteins whose outer loops face a cytosolic rather than external environment. A more detailed introduction into TMBs is given in two recent, comprehensive reviews by Schulz (1) and Wimley (19).

Membrane strands can be predicted from sequence

Predicting membrane helices has been pursued for over two decades (20–27). In contrast, attempts to predict membrane strands has long been limited to single cases and/or alignment-based techniques (28). One reason for this asymmetry may have been the limited availability of experimental data. Two groups ventured to apply neural networks to this problem despite the very limited data (29,30). Other methods used a combination of rules (31), or a rapid screening algorithm based on bio-physical features (32). A semi-manual pipeline involving a combination of different prediction methods was designed to find chloroplast outer envelope proteins in *Arabidopsis thaliana*; four unknown membrane proteins were unravelled in this way (33). Another group used the residue composition of predicted beta-strands in a linear classifier called the 'pocket algorithm' to discriminate membrane beta proteins from beta-containing non-membrane proteins (34). A combination of hydropathy, amphipathicity and predicted beta-strand content was used in the Beta-barrel finder (BBF) program to identify 118 putative TMB proteins in *E.coli* (35). However, the authors did not estimate how accurately their method distinguishes TMB from non-TMB proteins. Martelli and co-workers implemented a profile-based hidden Markov model (HMM) that appears to successfully predict membrane strands and distinguishes between proteins with and without membrane strands, in particular between strands in membranes and those in globular proteins (36). Several groups have used predictions in conjunction with experimental data to build models for newly studied proteins (37–41).

Here, we present a novel, fully automated HMM-based method to predict membrane strands and to distinguish beta-membrane from all other proteins. Although conceptually similar to the one pioneered by Martelli *et al.* for membrane strands (36), our method differed significantly in the design details. Whole protein discrimination of our model slightly outperforms that published previously (36); per-residue performance is overall similar. We propose our method as a complement. We also provide a large-scale prediction of TMB proteins for 72 entire proteomes from Gram-negative bacteria and for a few Gram-positive bacteria with membranes that are unusual for Gram-positive bacteria. In particular, we identify 164 proteins that could not have been identified through sequence similarity, i.e. have never been implicated with membranes.

MATERIALS AND METHODS

Data sets

Sequence-unique subsets. Fifty-six TMB structures were in PDB (42) when we began our work. Many of these are closely related in sequence. We reduced this bias by creating the largest possible subset of the 56 that fulfilled the following condition: no pair in the set had a sequence similarity above an HVAL of +3 [distance from the Sander-Schneider curve (43–45)]:

$$\mathrm{HVAL} = \mathrm{PID} - \begin{cases} 100 & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32 \cdot \{1+\exp^{-L/100}\}} & \text{for } L \leq 450 \\ 19.5 & \text{for } L > 450 \end{cases} \quad \mathbf{1}$$

in which L was the number of residues aligned between two proteins, PID the percentage of pairwise identical residues, and the functional shape the revised Sander-Schneider curve (44). An HVAL of 0 defines the line, above which (almost) no two naturally evolved proteins differ grossly in their three-dimensional structures. To illustrate the curve: for alignment lengths around 100 residues, 33% pairwise sequence identity suffices to infer structure, above 250 residues 21% is significant, and below 11 residues even 100% identity is not enough to infer structural (or functional) similarity. Technically, we used our server UniqueProt (46) to identify this training set. The final set had 11 members, of these, 1ek9 and 7ahl were excluded, since their barrels are made up of three or seven identical chains, respectively. 1pvl was also excluded since it was a quite flattened and atypical structure. We were finally left with the following eight proteins (labelled as SetTMB): Sucrose Specific Porin 1a0s_P (3), Maltoporin from *E.coli* 1af6_A (4), OmpF Matrix Porin from *E.coli* 1bt9_ A (47), FepA active transporter from *E.coli* 1fep_A (48), 'porin' from *R.blastica* 1prn (49), phospholipase A from *E.coli* 1qd5_A (50), OmpA from *E.coli* 1qjp_A (2), and OmpX from *E.coli* 1qj9_A (11). For training our HMM, we used the sequences given in the Protein Databank (PDB) records. PDB structures are often cleaved versions of the native proteins. Therefore, we also tested our method with the corresponding full sequences of the native proteins taken from SWISS-PROT (51). Relative to SWISS-PROT (SetTMBfull), seven of the eight PDB sequences lacked N-terminal residues and one structure (1qjp) lacked 154 C-terminal residues.

Testing discrimination from non-TMB proteins. We created a new data set (SetROC) to evaluate how well our methods distinguished between proteins with and without membrane strands. This set constituted a subset of SWISS-PROT generated in the following way. We started with the 143 521 proteins in SWISS-PROT by November 2003 and excluded all proteins with less than 140 residues (116 013 proteins). We then filtered these using an HVAL of 0 (equation **1**) as a metric in a greedy clustering algorithm identical to the protocol of UniqueProt (46). This resulted in 10 089 sequence-structure families. We further removed all proteins that had contributed to any of the alignments used to build profiles for training (remaining 10 085 families). We separated all proteins from Gram-negative bacteria (2585). Then, we calculated the percent of low-complexity regions using the program SEG (52) run with default options, and discarded all proteins with more than 15% low-complexity content, after observing that the highest content of low-complexity regions of one of the IOM proteins (see below) was 8% (OM1L_CHLTR, major outer membrane protein). This last step filtered out 214

proteins, leaving 2371 representative proteins. Finally, we discarded all proteins with keywords 'hypothetical protein' (1143), yielding 1228 proteins (SetROC). We used the program Meta_A(nnotator) (53,54) with a modification of the library to classify these proteins into six categories. Meta_A(nnotator) is a program which uses an extensive set of phrase pattern rules, combined with the SWISS-PROT fields in which these keywords occur, to classify proteins according to the experimentally observed sub-cellular localization. The categories (with number of proteins in parentheses) were: 'IOM' (13), 'peripheral outer membrane' (20), 'inner membrane' (102), 'single membrane' (118), and 'non-membrane' (975). Our second data set reflected both our more conservative criteria for bias reduction and the latest version of public databases.

Additional data sets used for comparison. In order to facilitate comparisons with the most accurate method previously published (36), we also tested our method on the sets used by Martelli and colleagues. In particular, to analyse per-residue accuracy, they used a set with 15 TMB structures, filtered at 30% identity—SetTMBcomp—comprising the following PDB chains: 1a0s_A, 1bxw_A, 1e54_A, 1ek9_A, 1fcp_A, 1fep_A, 1i78_A, 1k24_A, 1kmo_A, 1prn_A, 1qd5_A, 1qj8_A, 2mpr_A, 2omf_A, and 2por_A. To evaluate the distinction between proteins with and without membrane strand, they used 1572 proteins (http://www.cbrc.jp/papia/papia.html)—SetROCcomp: 145 outer- and 188 inner-membrane proteins taken from SWISS-PROT (51) that were filtered at a threshold of 30% sequence identity (corresponding to homology-derived secondary structure of protein distances around 5–10), and 1239 globular proteins from PDB filtered at 25% sequence identity.

Whole proteome prediction data sets. We identified all experimentally known 'integral outer membrane' (IOM) proteins in SWISS-PROT using Meta-(A)nnotator (53). Then, we identified all outer membrane proteins with no annotation as regarding integral or peripheral, from SWISS-PROT and TrEMBL files (from ftp://ftp.ebi.ac.uk/pub/databases/SPproteomes/swissprot_files/proteomes/) merely requiring the phrase 'outer membrane protein' to appear in the description field (DE) of the SWISS-PROT file (OM, 748 proteins). We then collected all proteins from 72 Gram-negative, 15 'typical' Gram-positive and five 'atypical' Gram-positive (Mycolata) proteomes, taken from ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/ and ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ (256 432 proteins). Among these, we identified all homologues to IOM proteins with PSI-BLAST *E*-values <0.01 (set IOM_homo with 1388 proteins) and homologues to the OM (OM_homo, 3829 proteins). In cases where a protein had homology to at least one protein in IOM and one in OM, the protein was classified as IOM_homo, not OM_homo. Then, we removed from all sets, proteins with greater than 20% low-complexity content [as determined by SEG with default parameters (52)]. Finally, we identified all proteins in these four sets and the remaining proteome proteins with PROFtmb whole protein discrimination scores above 8, calling this set 'PROFtmb' (Fig. 4). Thus, sets IOM, IOM_homo, OM and OM_homo were all disjoint. Set PROFtmb had intersections with all four sets, and with the remaining proteins not classified in those four sets.

Evaluation

Cross-validation experiment. For testing the performance of our method on membrane strand proteins, we applied a 'leave-one-out' jack-knife procedure on all unbiased data sets SetTMBcomp, SetTMB and SetTMBfull. All values given constituted averages over the performance for the protein left out, i.e. we never show results from the training data. The objective was that we did not use the information of any homologue to decide on any of the parameters of our model since this is likely to yield over-estimates of prediction accuracy. Our testing scheme implied that we actually generated eight or 15 copies of our prediction method. For testing the whole-protein discrimination performance, we trained the HMMs on all the training data and tested on the SWISS-PROT derived sets from which these had been removed.

Measures for performance. In order to evaluate the per-residue and a per-segment performance, we converted our model to a two-state prediction (strand, non-strand). As standard-of-truth for the strand assignments of known structures, we used the DSSP state 'E' (55) (the few strands in non-membrane regions were converted to 'non-strand'). We investigated the performance according to the two-state per-residue accuracy:

$$Q_2 = \frac{100}{\text{Nprot}} \cdot \sum_{k=1}^{\text{Nprot}} \frac{\text{number of residues correctly predicted in protein } k}{\text{length of protein } k} \quad \mathbf{2}$$

where Nprot was the number of proteins in our test set. We also measured separately the success in predicting strands by the following two scores:

$$Q_\beta^{\%\text{obs}} = \frac{100}{\text{Nprot}} \cdot \sum_{k=1}^{\text{Nprot}} \frac{p(k)}{p(k)+u(k)} \quad \mathbf{3}$$

$$Q_\beta^{\%\text{prd}} = \frac{100}{\text{Nprot}} \cdot \sum_{k=1}^{\text{Nprot}} \frac{p(k)}{p(k)+o(k)} \quad \mathbf{4}$$

where $p(k)$ was the number of strand residues correctly predicted in protein k, $u(k)$ the number of under-predicted strand residues (note: $p + u$ is the total number of strand residues observed), and $o(k)$ the number of over-predicted strand residues (note: $p + o$ is the total number of strand residues predicted). Considering also $n(k)$, i.e. the number of correctly predicted non-strand residues, we finally used the standard Matthew's correlation coefficient (56) (C):

$$C = \frac{p \cdot n - o \cdot u}{\sqrt{(p+o)\cdot(p+u)\cdot(o+n)\cdot(u+n)}} \quad \mathbf{5}$$

Due to the mathematical features of this coefficient, we have to sum over the entire data set, rather than summing over the coefficient for each protein. Note that the Matthew's correlation coefficient is usually defined separately for all secondary

structure states. However, for a two-state model (strand/non-strand), *C*(strand) and *C*(non-strand) are numerically identical. All these per-residue scores ignore the segment nature of beta strands, for example, a method that correctly predicts the termini of six strands and entirely misses two short strands may appear more accurate under a per-residue measure than another one that correctly predicts all eight strands, however, never identifies the termini correctly. This problem is addressed by scores that measure the per-segment performance (57,58). In particular, we measure the performance of the segment overlap SOV(β) (58). We estimated standard deviations by monitoring the distribution of accuracy for all proteins. Due to the small data sets of high-resolution structures, our estimates for standard deviations probably constitute lower limits (59,60).

Whole-protein discrimination accuracy. To test the ability of PROFtmb to discriminate between TMBs and non-TMBs, we sorted all proteins in the test set according to their whole protein discrimination score that we calculated as:

$$D = \text{bits} - (0.066 * \text{protein length} - 6) \quad \mathbf{6}$$

where we estimated the slope 0.066 and the intercept –6 by the optimal separating line between the positives and negatives in a cluster plot of length versus bits scores (see Supplementary Material). We used the truncated receiver operator characteristic (ROC) score (61) to measure whole-protein discrimination accuracy in the following way. Traverse the list of predictions, sorted from best to worst score. The ROC curve is a series of points:

$$\left(\frac{i}{\text{Nneg}}, \frac{\text{Npos}(i)}{\text{Npos}}\right), \quad 1 \leq i \leq \text{Nneg} \quad \mathbf{7}$$

with Nneg given the number of negative examples considered (which may not be all available negative examples), Npos the total number of positive examples, and Npos(*i*) the number of positive examples with scores higher than the *i*th negative. The ROC_{Nneg} curve is then calculated as:

$$\text{ROC}_{\text{Nneg}} = \frac{1}{\text{Npos} \cdot \text{Nneg}} \sum_{i}^{\text{Nneg}} \text{Npos}(i) \quad \mathbf{8}$$

In particular, we showed the usual $\text{ROC}_{n\%}$ values, i.e. the values at which *n*% of the proteins identified, were correct. To generate standard deviations for ROC_n calculations, we re-sampled with replacement from the data set in question (i.e. SetROCcomp or SetROC) 1000 times, each time generating a re-sampled set of the same size. For each of the 1000 re-sampled sets, we calculated the ROC_n score, and then the standard deviation of the resulting distribution of 1000 ROC_n scores (Fig. 2).

Hidden Markov model

The model consists of the 91 states connected as shown (Fig. 1). States are labelled to indicate both the assignment of tied parameters and the labelling for which the state is valid during training (see Supplementary Material). For example, state Ua has its own parameters, and only has one valid label. In contrast, any of the nine states 'I' (inner loop) or 'O' (outer loop) have the same set of tied emission parameters, and they are all valid for the same residue position. The set of connections to and from each end of transmembrane strands allow modelling the variable-length strand overhangs explicitly. When modelling the periplasmic loop region, we consider explicitly existing four or five residue beta-turns since we observed these two types of turns to be abundant in periplasmic regions of TMBs (below).

Transmembrane beta-strand. To identify the up- and down-strands, we defined a latitude corresponding to the extra-cellular 'aromatic cuff' (1,10). We aligned all membrane strands to this 'aromatic cuff'. Only six sequence-consecutive positions were common to all TMBs. The majority of membrane strands extended several (average 4, maximum 15) residues further into the extra-cellular region and a few (average 1, maximum 8) residues into the periplasmic side.

Beta-hairpins. We defined beta-hairpins as the smallest stretch of contiguous, non-helical residues enclosed by a pair of double-backbone hydrogen-bonded residues, in the pattern of an anti-parallel beta-sheet bridge-partner pair. A modified version of the molecular browser Troll (62) was used to identify such beta-hairpins. Hydrogen bonds were defined using the following parameters: dipole-dipole energy maximum of –0.5 kcal/mol, <1.5 Radians angle between the N–H bond and the hydrogen bond, or between the C=O bond and the hydrogen bond.

Extra-cellular (outer) loops. All residues C-terminal to the up-strand and N-terminal to the down-strand were labelled as extra-cellular loop. This included short stretches of beta-strand residues and alpha-helical residues. These were all modelled with one state labelled 'O' for 'outer'. Using this definition, we found 14 four-residue and 7 five-residue periplasmic hairpins in our training set, out of a total of 51 periplasmic hairpins. There were other sizes, but these were by far the most abundant, so we chose to explicitly model these two types of hairpins, while modelling all other sizes with a set of nine states labelled 'I' for 'inner'.

N- and C-termini extra sequence. All extra N- or C-terminal sequences that were not part of the barrel proper was merely labelled as 'I'.

Interpretation of model architecture. Before settling on the present architecture we tried several others with notable differences, all of which performed less well than the present model. The first most surprising improvement came from separately modelling up-strands and down-strands, rather than having matching latitudes share the same emission parameters. We observed shortly after this that the up-strands and down-strands differed markedly in their composition at the aromatic cuffs. Another previous version modelled all pore-facing strand residues with the same emission parameters. The decision to also model the strand overhang regions as linear chains with each state having its own transition to the loop states also improved the results relative to a previous model in which the overhangs were represented as states with a

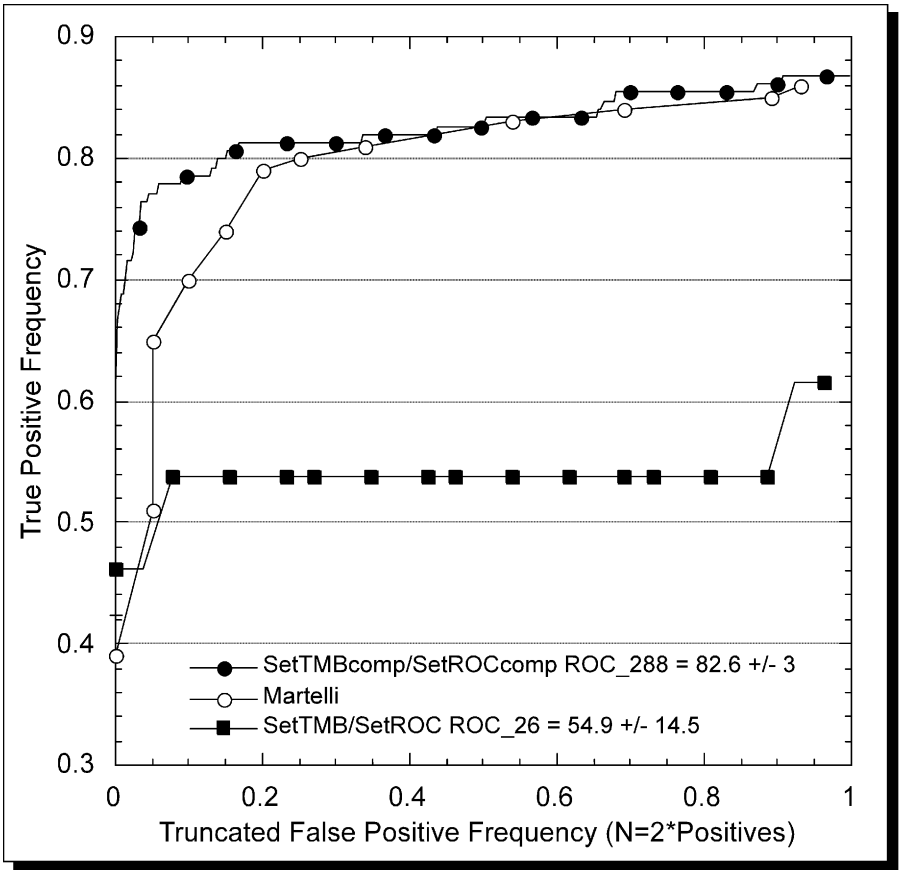

Figure 2. Discrimination between TMB and non-TMB proteins. The ROC curves truncated at a value with false positives = 2* true positives (Materials and Methods) suggested a slight improvement of our model (dark lines with circles) over the previous work by Martelli *et al.* (36) (grey lines with crosses). Note that the thin lines indicate tests on the data set previously used (SetROCcomp), while the thick line marks our much larger set (SetROC). The larger set suggested lower performance (open versus filled circles). Standard deviations of ROC_n scores (Materials and Methods) reveal the amount of noise.

self-transition. Finally, the explicit modelling of four- and five-residue hairpins improved the results by a few percent relative to a model with just the nine 'I' states.

Availability

Web-server. A web server is available at http://cubic.bioc.columbia.edu/services/proftmb/. Users can submit a file with sequences for single proteins or alignments and obtain whole-protein and/or two-state per-residue predictions. The prediction method is also integrated into the PredictProtein server (63) available at http://www.predictprotein.org.

Software. All software is available for download. This includes the source code for PROFtmb (GNU general public license), and one Perl script for parsing PSI-BLAST profiles into the readable format. Advanced users can retrain the HMM on any structural family of proteins by providing individual model specification files and a set of labelled training sequences. More commonly, the user will use the model specification files and labelled training sequences that are provided. PROFtmb is written in ANSI C++. It handles approximately 1000 proteins per minute in whole-protein discrimination mode (TMB, not-TMB), and approximately 100 per minute in per-residue mode, on a 2.8 GHz Xeon processor. Pre-compiled executables are available for Windows (win32), LINUX, SGI-IRIX and MAC OSX.

Predictions and data sets. Through http://cubic.bioc.columbia.edu/services/proftmb/ all predictions and data sets are available. In particular, we give the raw results of our method for 1641 proteins of unknown structure or function and scoring above -5 according to PROFtmb, from the 78 Gram-negative bacterial genomes. For those scoring above a certain whole-protein score threshold the corresponding per-residue predictions are also available. All predictions are also integrated into the PEP Predictions for Entire Proteomes (PEP) database (64) giving results from a variety of prediction methods for entirely sequenced proteomes.

RESULTS AND DISCUSSION

Particular architecture used

The key information for developing the PROFtmb architecture originated from a careful interpretation of the details in eight experimental high-resolution structures of barrels. The idea was to encode the 3D structures through a discrete set of 'structure states'. In doing this, we assumed that residues in similar micro-environments (such as the 'aromatic cuff') or

Table 1. Per-residue accuracy for different methods and different data sets[a]

Method	Data	Q_2	$Q_\beta^{\%prd}$	$Q_\beta^{\%obs}$	C	SOV_2	SOVβ
Estimated σ		±6	±9	±14	±0.11	±13	±11
Martelli	SetTMBcomp	84	87	80	0.69	91	94
PROFtmb	SetTMBcomp	83	87	80	0.69	79	93
PROFtmb	SetTMB	83	84	85	0.70	87	92
PROFtmb	SetTMBfull	86	86	85	0.74	88	94

[a]Method: PROFtmb marks the method introduced here as tested on different data sets. Martelli *et al.* marks results from a previously published method [40] (note: we re-generate Martelli's measures using the raw data courteously provided). Data: SetTMBcomp is a set of non-sequence unique beta-membrane proteins from PDB used by Martelli *et al.*, SetTMB is a sequence-unique subset of SetTMBcomp, with eight representative high-resolution TMBs, SetTMBfull is the same proteins, however, consisting of the full protein sequence as found in SWISS-PROT rather than on the reduced fragments used to determine the structure. Note: this test is important since we do not have PDB sequences. Measures: Q_2 is the two-state per-residue accuracy (equation **2**), $Q_\beta^{\%obs}$ is the percentage of known TMB residues correctly predicted (equation 3), $Q_\beta^{\%prd}$ the percentage of predicted TMB residues that are correctly predicted (equation **2**), C the Matthews correlation coefficient (equation **5**), SOV_2 the percentage segment overlap between all predicted and observed membrane and non-membrane segments, and SOVβ the corresponding segment overlap for membrane strands. Note that results with $Q_\beta^{\%obs} > Q_\beta^{\%prd}$ mark over-predictions, those with $Q_\beta^{\%obs} < Q_\beta^{\%prd}$ under-predictions of membrane strands. Estimates for standard deviation (σ): All data sets were too small to allow for adequate estimates of standard deviations. We provided what may constitute lower limits through a bootstrap analysis of the data. Although technically the bootstrap values differed slightly for different data sets, we quoted only the highest values found for each score.

following similar structural 'grammars' (e.g. all residues in the first position of a four-residue periplasmic beta-hairpin) would share selective pressure, and thus have a strongly biased residue composition. This was a natural extension from the observations of the 'aromatic cuff' and 'hydrophobic belt' (1,5,19). Each discrete structural state was represented as an 'architectural state' in the HMM (rectangles, Fig. 1). Having labelled each residue in the training set with a particular structure state, the grammar was specified as a consequence. If the state 'aromatic cuff' is followed by the state 'extra-cellular loop', this directed connection is specified in the HMM architecture (arrows, Fig. 1). A few specific features were naturally modelled by this approach. First, the aromatic cuff states faced outward toward the lipid bilayer by definition. We defined all other beta-strand states (whether embedded in the membrane or overhanging on either side) relative to this position. Thus, the alternating pattern (pore-facing, lipid-facing, …) was implicitly modelled. Secondly, variable length of strands overhanging, on either side of the outer membrane were modelled simply by the presence of additional states that were all connected directly to the states 'outer loop', 'inner loop' and 'hairpin' (Fig. 1, dashed rectangles/lines). Note that we did explicitly use the observation of the enrichment of tyrosine and phenylalanine in the latitude described as the extra-cellular aromatic cuff, to help determine its position. Technically, we gave them two-letter names (more details in Materials and Methods and Supplementary Material).

Per-residue performance: most residues predicted correctly

In terms of per-residue accuracy, our method predicted ~80% of all strand residues correctly, reaching a Matthew's correlation coefficient (equation **5**) as high as 0.7 (Table 1). Many methods developed on high-resolution structures are evaluated on the protein sequence deposited in the PDB. These sequences often constitute only fragments of the full-length protein (65). For TMBs, the major differences are that PDB sequences miss N-terminal residues including, but not restricted to, the signal peptide. Additionally, the sequence for OmpA (PDB identifier: 1qjp) lacked 154 C-terminal residues. Therefore, we also examined our method on the full-length protein sequences taken from SWISS-PROT. As it turned out, the performance was rather similar between the two sets (SetTMB versus SetTMBfull in Table 1). The overall two-state per-residue accuracy was higher for the full-length sequences simply because most additional residues were trivially recognized as being 'not membrane strand'. Overall, our method behaved very differently for full-length proteins: for PDB sequence fragments it over-predicted and for full-length sequences from SWISS-PROT it slightly over-predicted residues in membrane strands. At the same time, the observed strands were predicted even more accurately for full-length proteins (Table 1). The non-realistic over-prediction on the PDB data set was also the main difference in per-residue accuracy between our HMM-based method and the one published previously by Martelli and colleagues (36) (Table 1: SetTMBcomp). Martelli and colleagues did not evaluate their methods on full-length proteins. Since their HMM-based method was not publicly available during this work's development, we could not explore whether the difference that we observed is generic or particular to our method.

Detailed four-state model surprisingly accurate

Although our HMM internally represents many structural states that correspond to the barrel grammar (Fig. 1), the actual per-residue predictions were obtained by collapsing all these states into two (membrane-strand/other). Usually, two-state predictions reach numerically higher values than, e.g. four-state predictions due to the higher level for the random background (over-simplified: random is 50% for two states and 25% for four). We were thus surprised to observe that our four-state model was almost as accurate as the two-state reduction (Table 2). In fact, PROFtmb was extremely successful in distinguishing between upward- and downward-strands and between periplasmic- and outer loops (boldface in Table 2). For example, 1171 were correctly predicted as membrane-strand, and only 14 of these confused the states up-strand and down-strand. Similarly, only 15 of the 1706 correctly predicted non-membrane strand residues confused periplasmic and outer-membrane. The latter may be due to the strong difference in length distributions of the (short) periplasmic loops and (long) outer loops.

Table 2. Confusion matrix on four-state predictions[a]

Predicted → Observed ↓	Up-strand	Down-strand	Peri-loop	Outer-loop	SUM	Pok
Up-strand	**563**	9	27	57	656	86
Down-strand	5	**594**	58	33	690	86
Peri-loop	39	20	**763**	13	835	91
Outer-loop	77	104	2	**790**	973	81
SUM	684	727	850	893	3154	86
Pok	82	82	90	88	86	86

[a]In Table 1, we converted the actual states of our HMM onto two states: membrane-strand/other. Here, we showed more detail (see Fig. 1 for a visualization of the four states; SetTMB). For example, of the 656 residues observed in up-strand, 563 residues were correctly predicted (86%); in total PROFtmb predicted 684 up-strand residues, i.e. 82% of the predicted up-strand residues were correct. Note: Pok (percentage correct) = diagonal values/SUM over that row (observed) or column (predicted).

Multiple sequence alignments improved performance

Profiles from multiple sequence alignments contain important information about protein structure. In particular, using alignment information improved our model on average by about 18 percentage points in terms of two-state per-residue accuracy. Nevertheless, such profiles also constitute a source of noise, arising from alignment errors. Given the dramatic improvement due to the use of profiles, it is likely that additional improvement may be attainable through a more clever construction of alignments and profile extraction.

About 45% coverage at 100% accuracy in finding TMB

Although we made a special effort to curate our large data set used to establish the false positive rates of our method (SetROC), we found only minor differences between this large SWISS-PROT-based set and the data set taken from the PDB (SetROCcomp, Fig. 2). The two data sets gave significantly different ROC scores (82.5 versus 54.9%). Also, the larger SWISS-PROT data (SetROC) yielded a much larger standard deviation in a simple bootstrap (66) experiment (14.5 versus 3%). We used the larger data set to estimate the accuracy for whole-protein discrimination.

High accuracy and coverage above whole-protein score of 10

Applying our method to entirely sequenced proteins required introducing a threshold in the reliability of the prediction. This threshold reduced the number of incorrectly predicted membrane barrels. Our largest data set (SetROC) suggested that all proteins identified above whole-protein discrimination scores of 10 (equation **6**) were indeed TMBs (100% accuracy, Fig. 3). At this threshold ~45% of the TMB proteins in the data set were correctly identified (45% coverage, Fig. 3). Although we did not thoroughly characterise the distribution of scores using standard statistical methods (such as *Z*-scores), we feel that the whole protein score suffices as a reasonable estimate for accuracy and coverage, especially considering the very conservative choice of cut-off thresholds that we used.

Case study for known TMBs

We ran PROFtmb on six known TMBs proposed to us. These were (number of transmembrane strands predicted in parentheses): adhesin AIDA-I precursor from *E.coli* (20 strands), S-layer protein, putative from *Deinococcus radiodurans* (30 strands), hypothetical protein TM0476 from *Thermotoga maritima* (36 strands), putative exported protein from *Yersina pestis* (10 strands), SomB from *Synechococcus* sp. (16 strands) and S-layer protein precursor from *Thermus aquaticus* (no per-residue prediction given). PROFtmb identifies the first four with scores over 8, corresponding to 95% accuracy. SomB was given a score of 3.15, corresponding to 90% accuracy. Finally, S-layer protein precursor obtained a score of –4.8 (accuracy 25%), too low for PROFtmb to provide a per-residue prediction. It is apparent that PROFtmb over-predicts the number of transmembrane strands in at least two of these proteins (TM0476 and S-layer protein), but gives reasonable per-residue predictions for the other three proteins.

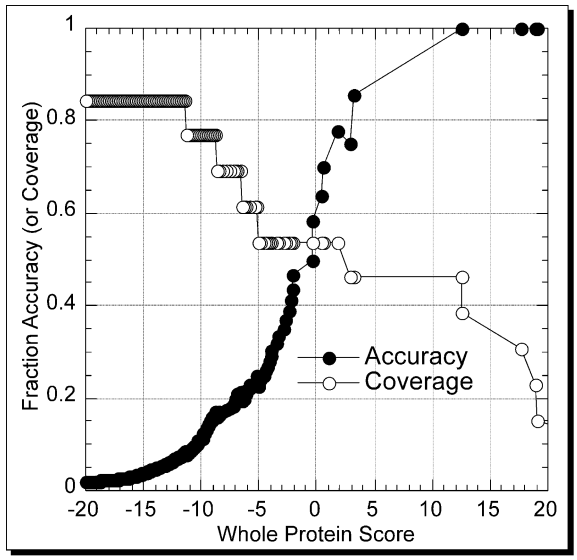

Figure 3. Threshold for accurate discrimination. Higher whole-protein discrimination scores (equation **6**) yielded higher accuracy (correctly predicted TMBs/predicted TMBs) in discriminating between TMBs and non-TMBs (black line with filled circles). The flipside of this was low coverage (correctly predicted TMBs/observed TMBs) at high accuracy (dotted grey lines mark thresholds used in Fig. 4). Nevertheless, this analysis performed on the largest possible sequence-unique data set (SetROC) suggested an impressive performance: 100% accuracy at levels of 45% coverage. (Note that the density of proteins—indicated by symbols—was much higher toward lower whole protein scores since most proteins in SetROC did not have beta-barrels.)

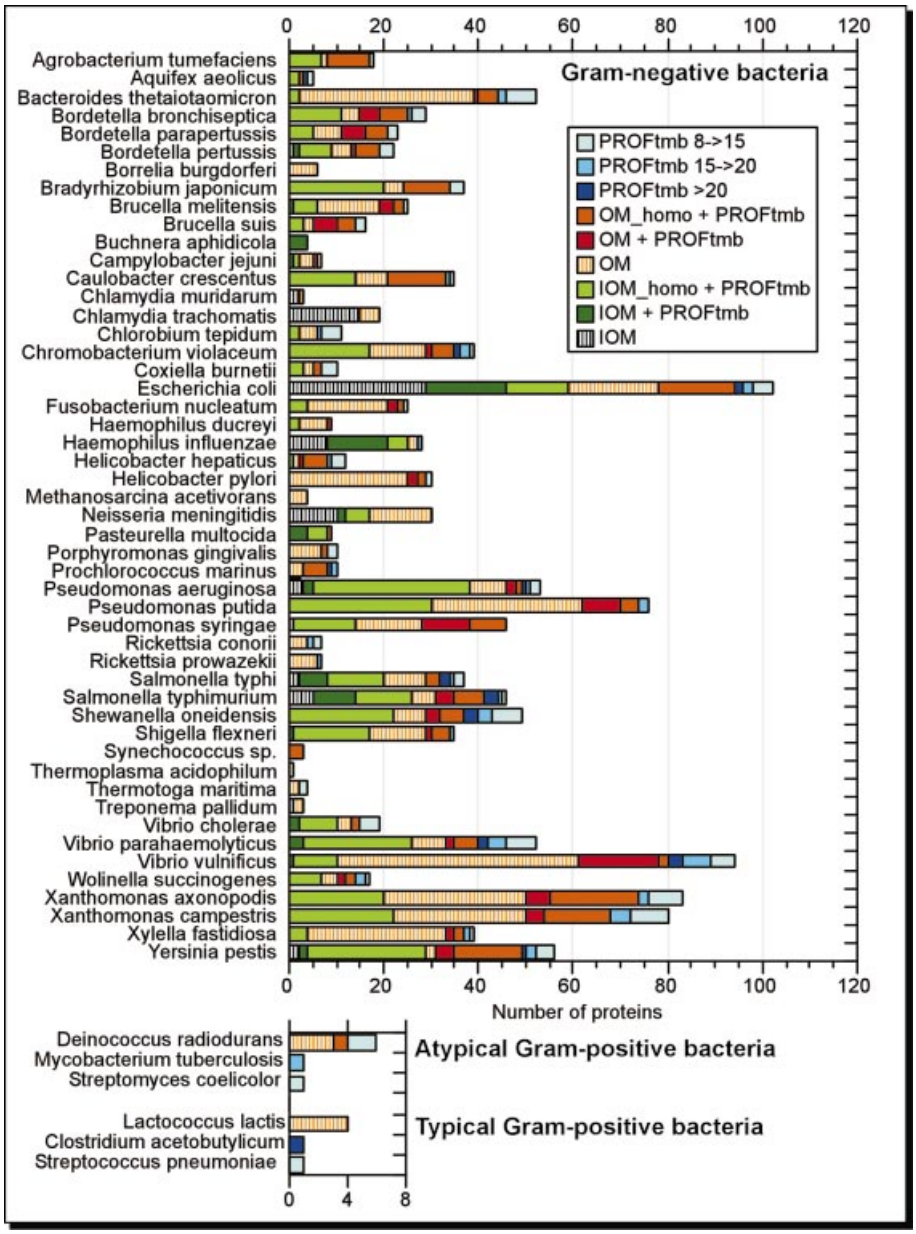

Figure 4. Transmembrane barrels predicted in entire proteomes. For each proteome, we reported the numbers of proteins in each set (or intersection of sets). Sets IOM, IOM and their homologues (IOM_homo), outer membrane (OM) and OM_homo are disjoint (Materials and Methods), while the set 'PROFtmb' denotes all proteins achieving a whole protein score above 8 (equation **6**), and is not disjoint with the four sets above. Thus, categories in this graph are named according to which sets the proteins belong. For example, 'PROFtmb 8→15' denotes all previously un-annotated proteins achieving a score between 8 and 15, and similarly for 'PROFtmb 15→20' and 'PROFtmb >20'. 'OM + PROFtmb' denotes all proteins annotated as outer membrane which also achieve a PROFtmb score above 8. 'IOM' denotes all proteins in IOM but with PROFtmb scores <8 (i.e. not in set PROFtmb). Note that all findings in typical Gram-positive bacteria constitute false positives. Categories IOM_homo and OM_homo without PROFtmb predictions are not reported, since many proteins in these sets are likely not generic TMBs. The following proteomes had neither known IOMs nor yielded any PROFtmb predictions: Gram-negative: *Archaeoglobus fulgidus*, *Blochmannia floridanus*, *Buchnera aphidicola_Sg*, *Buchnera* sp., *Halobacterium* sp., *Leptospira intrerrogans*, *Mesorhizobium loti*, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Methanosarcina mazei*, *Nostoc* sp., *Pirellula* sp., *Pyrococcus abyssi*, *Pyrococcus furiosus*, *Pyrococcus horikoshii*, *Sinorhizobium meliloti*, *Sulfolobus solfataricus*, *Sulfolobus tokodaii*, *Thermoplasma volcanium*, *Thermosynechococcus elongatus*, *Ureaplasma urealyticum*, *Wigglesworthia brevipalpis*; atypical Gram-positive: *Mycobacterium bovis*, *Mycobacterium leprae*; typical Gram-positive: *Bacillus subtilis*, *Clostridium perfringens*, *Enterococcus faecalis*, *Listeria innocua*, *Mycoplasma gallisepticum*, *Mycoplasma genitalium*, *Mycoplasma penetrans*, *Mycoplasma pneumoniae*, *Mycoplasma pulmonis*, *Oceanobacillus iheyensis*, *Staphylococcus aureus*, *Streptococcus pyogenes*.

Most TMBs appear to be known

We collected all proteins in each fully sequenced proteome of 72 Gram-negative, 15 'typical' and five 'atypical' (Mycolata) Gram-positive bacteria (Fig. 4). We also defined sets with 'integral outer membrane' proteins (IOM), their homologues (IOM_homo), outer-membrane proteins (OM), and their homologues (OM_homo; see Materials and Methods). We applied PROFtmb to all proteins in all proteomes and retained predictions with scores >8 (corresponding to 95% accuracy

and 45% coverage, Fig. 3). While PROFtmb identified 46% (69/148) of the experimentally known IOM proteins, it identified only 28% (388/1388) of the proteins that might have been labelled as 'IOM' based on sequence similarity alone (set of homologues). The significant discrepancy between these two results (IOM/IOM_homo) suggested that homology-based inference alone is likely to generate too many false positives. In contrast, PROFtmb identified only ~16% (91/560) of the proteins labelled as 'outer membrane' and only 5% (191/3829) of their homologues. Most likely this low percentage is a combination of actual TMB proteins missed by PROFtmb and peripheral outer membrane proteins that were not annotated precisely enough. Finally, PROFtmb found 164 new proteins at the 95% accuracy score cut-off which had—to the best of our knowledge—previously not been annotated as outer membrane nor were sequence similar to any outer membrane protein, not even at a liberal PSI-BLAST *E*-values <0.01.

Closer inspection of 164 new findings

Of the 164 completely novel finds, only two were from 'typical' Gram-positive bacteria, and thus false positives; all others originated from only 34 of the 72 Gram-negative proteomes. Those with six or more new proteins were: *Vibrio vulnificus* (14 proteins), *Vibrio parahaeomlyticus* (67) (12 proteins), *Xanthomonas campestris* (68) (12 proteins), *Shewanella oneidensis* (69) (12 proteins), *Xanthomonas axonopodis* (68) (nine proteins), *E.coli* (70) (eight proteins), *Bacteroides thetaiotaomicron* (71) (eight proteins) and *Yersinia pestis* (72) (seven proteins). 'Atypical' Gram-positive bacteria have outer membranes nearly twice as thick as those of Gram-negatives, composed of mycolic acid and a variety of extractible lipid, and contain pore-forming proteins (73,74). A 1.7 nm electron microscopic image of MspA from *Mycobacterium tuberculosis* revealed the pore to be 10 nm, in contrast to the ~4 nm pores of 'typical' Gram-positive pores (75). Recently, the first structure of a mycobacterial outer membrane protein was solved (76). Though it presents invaluable new information, we did not attempt to include it in our model for this study. Among the 'atypical' Gram-positives, PROFtmb identified four previously unidentified proteins. These were: conserved hypothetical protein (GI 15805996) and hypothetical protein (GI 15805156) from *D.radiodurans*, PPE (GI 15610479) from *M.tuberculosis*, and secreted *endo*-1,4-beta-xylanase B (GI 21220761) from *Streptomyces coelicolor*. Additionally, PROFtmb correctly identified a single protein, S-layer protein, putative (GI 15807560) from *D.radiodurans*. However, very permissive sequence searches picked 32 proteins from these five proteomes that had some sequence similarity to known IOMs. PROFtmb detected none of these, possibly because sequence similarity was too permissive (hence the findings constituted false positives), but more likely because TMBs traversing thicker membrane may have to differ in detail and hence might not be modelled accurately by PROFtmb.

Comparison to other methods

We compared our findings in *E.coli* to those of Zhai and Saier (35). Their BBF program identifies 118 proteins: 47 previously known TMBs and 71 additional unknown proteins. PROFtmb identifies 54 proteins in *E.coli*: 30 IOMs, 16 OMs (with no annotation as regards 'integral or peripheral'; see Materials and Methods) and eight previously unknown. Between BBF and PROFtmb, only 24 proteins were commonly identified, with only one of those (yjbH protein precursor, GI 7451212) previously unknown. While this discrepancy reflects the substantial differences in the two procedures and stringency of cut-off thresholds used, the small overlap (24 out of 118 or 54) is still surprising. But, if accurate, BBF is complementary to PROFtmb, at least for *E.coli*. We applied PROFtmb to eukaryotic proteins, but found that the program failed to accurately identify putative TMB proteins in these organisms. Since PROFtmb was trained exclusively on bacterial TMBs, failure to detect any eukaryotic TMBs is most likely due to the significantly different statistics of these structures. Attempting to address this problem, Schleiff *et al.* (33) use a pipeline to identify TMBs in the outer membrane of the chloroplast of *A.thaliana*. However, the authors did not report any explicit predictions.

CONCLUSIONS

The HMM for the structure of TMBs implemented in our novel method PROFtmb proved rather accurate. Although our method did not outperform a simpler HMM (36) in terms of a two-state model (membrane/not-membrane), it excelled in modelling more detailed states (distinction between up- and down-strands). This is certainly a benefit of using detailed structural and residue preference observations (1,5,19) to guide the initial model specification and structure-based labelling. In contrast to other available methods (34,35), ours is fully automated. Our improved estimate for log-odd values in discriminating TMB from non-TMB proteins allowed the application of PROFtmb to entirely sequenced proteomes. Analysing 72 Gram-negative bacteria, we found 164 putative TMB proteins that had previously not been implicated with membranes. Although a few of these predicted TMB proteins will turn out to be false positives, this set certainly constitutes a good starting point in the hunt for experimental studies of unknown outer membrane proteins.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

Thanks to Megan Restuccia (Columbia) for computer assistance. One discussion with Chris Bystroff helped towards understanding HMMs and parameter estimation in particular. Numerous discussions with Yanay Ofran and Trevor Siggers (Columbia) are gratefully acknowledged. Thanks also to Pier Luigi Martelli (Bologna) for graciously providing their data sets upon request. We are also grateful to both anonymous reviewers who helped considerably in improving our manuscript. This work was supported by the grants RO1-GM63029-01 and GM64633-01 from the National Institutes of Health (NIH), the grant LM07329-01 from the National Library of Medicine, and the grant DBI-0131168 from the National Science Foundation (NSF). Last, but not least, thanks to Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil

Bourne (San Diego University), and their crews for maintaining excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

REFERENCES

1. Schulz,G.E. (2000) beta-Barrel membrane proteins. *Curr. Opin. Struct. Biol.*, **10**, 443–447.
2. Pautsch,A. and Schulz,G.E. (2000) High-resolution structure of the OmpA membrane domain. *J. Mol. Biol.*, **298**, 273–282.
3. Forst,D., Welte,W., Wacker,T. and Diederichs,K. (1998) Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nature Struct. Biol.*, **5**, 37–46.
4. Wang,Y.F., Dutzler,R., Rizkallah,P.J., Rosenbusch,J.P. and Schirmer,T. (1997) Channel specificity: structural basis for sugar discrimination and differential flux rates in maltoporin. *J. Mol. Biol.*, **272**, 56–63.
5. Koebnik,R., Locher,K.P. and Van Gelder,P. (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. *Mol. Microbiol.*, **37**, 239–253.
6. Ferguson,A.D., Hofmann,E., Coulton,J.W., Diederichs,K. and Welte,W. (1998) Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide [see comments]. *Science*, **282**, 2215–2220.
7. Buchanan,S.K., Smith,B.S., Venkatramani,L., Xia,D., Esser,L., Palnitkar,M., Chakraborty,R., van der Helm,D. and Deisenhofer,J. (1999) Crystal structure of the outer membrane active transporter FepA from *Escherichia coli* [see comments]. *Nature Struct. Biol.*, **6**, 56–63.
8. Jap,B.K. (1989) Molecular design of PhoE porin and its functional consequences. *J. Mol. Biol.*, **205**, 407–419.
9. Jap,B.K., Downing,K.H. and Walian,P.J. (1990) Structure of PhoE porin in projection at 3.5 A resolution. *J. Struct. Biol.*, **103**, 57–63.
10. Meyer,J.E., Hofnung,M. and Schulz,G.E. (1997) Structure of maltoporin from Salmonella typhimurium ligated with a nitrophenyl-maltotrioside. *J. Mol. Biol.*, **266**, 761–775.
11. Vogt,J. and Schulz,G.E. (1999) The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. *Structure*, **7**, 1301–1309.
12. Vandeputte-Rutten,L., Kramer,R.A., Kroon,J., Dekker,N., Egmond,M.R. and Gros,P. (2001) Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site. *EMBO J.*, **20**, 5033–5039.
13. Snijder,H.J., Kingma,R.L., Kalk,K.H., Dekker,N., Egmond,M.R. and Dijkstra,B.W. (2001) Structural investigations of calcium binding and its role in activity and activation of outer membrane phospholipase A from *Escherichia coli*. *J. Mol. Biol.*, **309**, 477–489.
14. Snijder,H.J. and Dijkstra,B.W. (2000) Bacterial phospholipase A: structure and function of an integral membrane phospholipase. *Biochim. Biophys Acta*, **1488**, 91–101.
15. Buchanan,S.K. (2001) Type I secretion and multidrug efflux: transport through the TolC channel-tunnel. *Trends Biochem. Sci.*, **26**, 3–6.
16. Locher,K.P., Rees,B., Koebnik,R., Mitschler,A., Moulinier,L., Rosenbusch,J.P. and Moras,D. (1998) Transmembrane signaling across the ligand-gated FhuA receptor: crystal structures of free and ferrichrome-bound states reveal allosteric changes. *Cell*, **95**, 771–778.
17. Van Gelder,P., Saint,N., Phale,P., Eppens,E.F., Prilipov,A., van Boxtel,R., Rosenbusch,J.P. and Tommassen,J. (1997) Voltage sensing in the PhoE and OmpF outer membrane porins of *Escherichia coli*: role of charged residues. *J. Mol. Biol.*, **269**, 468–472.
18. Van Gelder,P., Saint,N., van Boxtel,R., Rosenbusch,J.P. and Tommassen,J. (1997) Pore functioning of outer membrane protein PhoE of *Escherichia coli*: mutagenesis of the constriction loop L3. *Protein Eng.*, **10**, 699–706.
19. Wimley,W.C. (2003) The versatile beta-barrel membrane protein. *Curr. Opin. Struct. Biol.*, **13**, 404–411.
20. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
21. von Heijne,G. (1981) Membrane proteins: the amino acid composition of membrane-penetrating segments. *Eur. J. Biochem.*, **120**, 275–278.
22. White,S.H. and Wimley,W.C. (1999) Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys Biomol. Struct.*, **28**, 319–365.
23. Ikeda,M., Arai,M., Lao,D.M. and Shimizu,T. (2001) Transmembrane topology prediction methods: A re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.*, **1**, http://www.bioinfo.de/isb/2001/2002/0003/.
24. Nakai,K. (2001) Prediction of in vivo fates of proteins in the era of genomics and proteomics. *J. Struct. Biol.*, **134**, 103–116.
25. Simon,I., Fiser,A. and Tusnady,G.E. (2001) Predicting protein conformation by statistical methods. *Biochim. Biophys Acta*, **1549**, 123–136.
26. Chen,C.P., Kernytsky,A. and Rost,B. (2002) Transmembrane helix predictions revisited. *Protein Sci.*, **11**, 2774–2791.
27. Chen,C.P. and Rost,B. (2002) State-of-the-art in membrane prediction. *Appl. Bioinf.*, **1**, 21–35.
28. von Heijne,G. (1996) Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.*, **66**, 113–139.
29. Diederichs,K., Freigang,J., Umhau,S., Zeth,K. and Breed,J. (1998) Prediction by a neural network of outer membrane beta-strand protein topology. *Protein Sci.*, **7**, 2413–2420.
30. Jacoboni,I., Martelli,P.L., Fariselli,P., De Pinto,V. and Casadio,R. (2001) Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor. *Protein Sci.*, **10**, 779–787.
31. Gromiha,M.M., Majumdar,R. and Ponnuswamy,P.K. (1997) Identification of membrane spanning beta strands in bacterial porins. *Protein Eng.*, **10**, 497–500.
32. Wimley,W.C. (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. *Protein Sci.*, **11**, 301–312.
33. Schleiff,E., Eichacker,L.A., Eckart,K., Becker,T., Mirus,O., Stahl,T. and Soll,J. (2003) Prediction of the plant beta-barrel proteome: a case study of the chloroplast outer envelope. *Protein Sci.*, **12**, 748–759.
34. Liu,Q., Zhu,Y.S., Wang,B.H. and Li,Y.X. (2003) A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. *Comput. Biol. Chem.*, **27**, 69–76.
35. Zhai,Y. and Saier,M.H.,Jr (2002) The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within prokaryotic genomes. *Protein Sci.*, **11**, 2196–2207.
36. Martelli,P.L., Fariselli,P., Krogh,A. and Casadio,R. (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18** [Suppl 1], S46–S53.
37. Fanucci,G.E., Cadieux,N., Piedmont,C.A., Kadner,R.J. and Cafiso,D.S. (2002) Structure and dynamics of the beta-barrel of the membrane transporter BtuB by site-directed spin labeling. *Biochemistry*, **41**, 11543–11551.
38. Heins,L., Mehrle,A., Hemmler,R., Wagner,R., Kuchler,M., Hormann,F., Sveshnikov,D. and Soll,J. (2002) The preprotein conducting channel at the inner envelope membrane of plastids. *EMBO J.*, **21**, 2616–2625.
39. Labesse,G., Garnotel,E., Bonnel,S., Dumas,C., Pages,J.M. and Bolla,J.M. (2001) MOMP, a divergent porin from *Campylobacter*: cloning and primary structural characterization. *Biochem. Biophys. Res. Commun.*, **280**, 380–387.
40. Conlan,S., Zhang,Y., Cheley,S. and Bayley,H. (2000) Biochemical and biophysical characterization of OmpG: a monomeric porin. *Biochemistry*, **39**, 11845–11854.
41. Wong,K.K. and Hancock,R.E. (2000) Insertion mutagenesis and membrane topology model of the Pseudomonas aeruginosa outer membrane protein OprM. *J. Bacteriol.*, **182**, 2402–2410.
42. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
43. Sander,C. and Schneider,R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
44. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
45. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
46. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res.*, **31**, 3789–3791.
47. Phale,P.S., Philippsen,A., Kiefhaber,T., Koebnik,R., Phale,V.P., Schirmer,T. and Rosenbusch,J.P. (1998) Stability of trimeric OmpF porin: the contributions of the latching loop L2. *Biochemistry*, **37**, 15663–15670.
48. Buchanan,S.K., Smith,B.S., Venkatramani,L., Xia,D., Esser,L., Palnitkar,M., Chakraborty,R., van der Helm,D. and Deisenhofer,J. (1999)

Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. *Nature Struct. Biol.*, **6**, 56–63.
49. Kreusch,A. and Schulz,G.E. (1994) Refined structure of the porin from *Rhodopseudomonas blastica*. Comparison with the porin from *Rhodobacter capsulatus*. *J. Mol. Biol.*, **243**, 891–905.
50. Snijder,H.J., Ubarretxena-Belandia,I., Blaauw,M., Kalk,K.H., Verheij,H.M., Egmond,M.R., Dekker,N. and Dijkstra,B.W. (1999) Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. *Nature*, **401**, 717–721.
51. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
52. Wootton,J.F.S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
53. Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.*, **8**, 169–170.
54. Eisenhaber,F. and Bork,P. (1999) Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, **15**, 528–535.
55. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
56. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys Acta*, **405**, 442–451.
57. Rost,B., Sander,C. and Schneider,R. (1994) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.
58. Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.
59. Rost,B. (2001) Protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
60. Rost,B. and Eyrich,V. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins*, **45** [Suppl 5], S192–S199.
61. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
62. Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties and electrostatics of macromolecular structures and sequences. *Methods Enzymol.*, **374**, 492–509.
63. Rost,B. and Liu,J. (2003) The PredictProtein server. *Nucleic Acids Res.*, **31**, 3300–3304.
64. Carter,P., Liu,J. and Rost,B. (2003) PEP: predictions for entire proteomes. *Nucleic Acids Res.*, **31**, 410–413.
65. Liu,J. and Rost,B. (2004) CHOP proteins into structural domain-like fragments. *Proteins*, **55**, 678–688.
66. Diaconis,P. and Efron,B. (1983) Computer-intensive methods in statistics. *Sci. Am.*, **248**, 96–108.
67. Makino,K., Oshima,K., Kurokawa,K., Yokoyama,K., Uda,T., Tagomori,K., Iijima,Y., Najima,M., Nakano,M., Yamashita,A. *et al.* (2003) Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*. *Lancet*, **361**, 743–749.
68. da Silva,A.C., Ferro,J.A., Reinach,F.C., Farah,C.S., Furlan,L.R., Quaggio,R.B., Monteiro-Vitorello,C.B., Van Sluys,M.A., Almeida,N.F., Alves,L.M. *et al.* (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*, **417**, 459–463.
69. Heidelberg,J.F., Paulsen,I.T., Nelson,K.E., Gaidos,E.J., Nelson,W.C., Read,T.D., Eisen,J.A., Seshadri,R., Ward,N., Methe,B. *et al.* (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.*, **20**, 1118–1123.
70. Blattner,F.R., Plunkett,G., 3rd, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
71. Nelson,K.E., Fleischmann,R.D., DeBoy,R.T., Paulsen,I.T., Fouts,D.E., Eisen,J.A., Daugherty,S.C., Dodson,R.J., Durkin,A.S., Gwinn,M. *et al.* (2003) Complete genome sequence of the oral pathogenic *Bacterium porphyromonas gingivalis* strain W83. *J. Bacteriol.*, **185**, 5591–5601.
72. Deng,W., Burland,V., Plunkett,G., 3rd, Boutin,A., Mayhew,G.F., Liss,P., Perna,N.T., Rose,D.J., Mau,B., Zhou,S. *et al.* (2002) Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.*, **184**, 4601–4611.
73. Seltmann,G. and Holst,O. (2002) *The Bacterial Cell Wall*. Springer, Berlin, New York.
74. Niederweis,M. (2003) Mycobacterial porins—new channel proteins in unique outer membranes. *Mol. Microbiol.*, **49**, 1167–1177.
75. Engelhardt,H., Heinz,C. and Niederweis,M. (2002) A tetrameric porin limits the cell wall permeability of *Mycobacterium smegmatis*. *J. Biol. Chem.*, **277**, 37567–37572.
76. Faller,M., Niederweis,M. and Schulz,G.E. (2004) The structure of a mycobacterial outer-membrane channel. *Science*, **303**, 1189–1192.