doi:10.1016/j.jmb.2006.07.095 J. Mol. Biol. (2006) 363, 611–623

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Sequence Motifs and Antimotifs in β-Barrel Membrane Proteins from a Genome-Wide Analysis: The Ala-Tyr Dichotomy and Chaperone Binding Motifs

Ronald Jackups Jr, Sarah Cheng† and Jie Liang*

Department of Bioengineering SEO, MC-063, University of Illinois at Chicago, 851 S. Morgan Street, Room 218 Chicago, IL 60607-7052, USA

**Corresponding author*

β-barrel membrane proteins are found in the outer membrane of gram-negative bacteria, mitochondria, and chloroplasts. Although sequence motifs have been studied in α-helical membrane proteins and have been shown to play important roles in their assembly, it is not clear whether over-represented motifs and under-represented anti-motifs exist in β-barrel membrane proteins. We have developed probabilistic models to identify sequence motifs of residue pairs on the same strand separated by an arbitrary number of residues. A rigorous statistical model is essential for this study because of the difficulty associated with the short length of the strands and the small amount of structural data. By comparing to the null model of exhaustive permutation of residues within the same β-strand, propensity values of sequence patterns of two residues and *p*-values measuring statistical significance are calculated exactly by several analytical formulae we have developed or by enumeration. We find that there are characteristic sequence motifs and antimotifs in transmembrane (TM) β-strands. The amino acid Tyr plays an important role in several such motifs. We find a general dichotomy consisting of favorable Aliphatic-Tyr sequence motifs and unfavorable Tyr-Aliphatic antimotifs. Tyr is also part of a terminal motif, YxF, which is likely to be important for chaperone binding. Our results also suggest several experiments that can help to elucidate the mechanisms of *in vitro* and *in vivo* folding of β-barrel membrane proteins.

Keywords: β-barrel membrane protein; combinatorial model; sequence pattern; sequence motif; sequence antimotif

Introduction

Integral membrane proteins can be categorized into two structural classes: α-helical proteins and β-barrel proteins. The structural properties of helical membrane proteins are well characterized, including amino acid composition,[1,2] inter-helical spatial interactions,[3,4] and the packing of helical bundles.[5] β-barrel membrane proteins are found in the outer membrane of gram-negative bacteria, mitochondria, and chloroplasts. Recent studies of β-barrel membrane proteins have revealed much insight on a number of issues, including general structural architecture,[6] characteristic amino acid preferences,[7–15] and spatial strand interaction patterns.[7] Nevertheless, our structural knowledge of the organizational principles of β-barrel membrane proteins lags behind that of α-helical membrane proteins. For example, an important development in the study of helical membrane protein assembly has been the identification of sequence motifs by computational analysis.[16] These motifs play important roles in the folding and assembly of transmembrane (TM) helices. Examples include the well-known GxxxG motifs that promote the dimerization of Glycophorin A,[16,17] as well as other Small-xxx-Small motifs.[17] In contrast, very little is known about sequence motifs in β-barrel membrane proteins or their roles in maintaining protein stability and function.

It is conceivable that there are sequence motifs in β-barrel membrane proteins that aid in their

† Summer student from Illinois Mathematics and Science Academy. Current address: Dept. of Electrical Eng. and Computer Sci, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

E-mail address of the corresponding author: jliang@uic.edu

0022-2836/$ - see front matter

structural and functional integrity in the lipid environment of the outer membrane, as well as antimotifs, forbidden patterns that disrupt the integrity of these proteins. Recent experimental studies suggest that such sequence motifs do exist and are functionally important.[18]

It is a challenging task to identify possible sequence motifs in β-barrel membrane proteins and to infer their roles. Because TM strands are short, methods for motif discovery based on more approximate distributions such as the binomial[19] and χ^2 [20] distributions are not applicable. We have developed a rigorous statistical model based on the combinatorics of residues in TM β-strands to estimate the propensity for the occurrence of patterns of two residues separated by an arbitrary number of residues within a strand. The propensities allow us to identify sequence motifs that are favored and antimotifs that are disfavored. We are also able to calculate *p*-values for measuring the statistical significance of these motifs and antimotifs. In addition, we have also developed a method for identifying motifs and antimotifs located in the loop region.

We use a two-pronged approach. First, we use our method on a large set of sequences in gram-negative bacterial genomes predicted to be TM β-strands. Second, we rationalize the motifs identified by the first study by using the same method on a small set of β-barrel membrane proteins of known structure and by examining the structural features of these motifs. The sequence motifs identified from known structures agree well with those derived from the larger genome-wide set of putative TM strands, suggesting that our statistical model works even when structural data is very limited, and our results are therefore robust.

We have succeeded in identifying many sequence motifs. We find that the amino acid Tyr contributes to many sequence motifs, including the general dichotomy of favorable N-to-C Aliphatic-Tyr sequence motifs and unfavorable Tyr-Aliphatic antimotifs. We also find that Tyr-x-Phe is a favored motif with a high propensity for occurrence in the interstrand loop region. It is known that SurA, a chaperone important for the folding of β-barrel membrane proteins, recognizes a specific sequence pattern involving aromatic residues.[18] This pattern includes the Tyr-x-Phe motif we discovered. In addition, the propensity scale of intrastrand residue pairs we have created will be useful for further studies of β-barrel membrane protein folding. This paper is organized as follows: we first discuss motifs and antimotifs found in an analysis of genomic sequences. We then compare these results with those obtained using only sequences of β-barrel membrane proteins with known structures. This is followed by results of motifs and antimotifs in the full protein sequences. We conclude with discussion of the role of Tyr in forming motifs, and the motifs potentially recognizable by the chaperone SurA.

Results

Development of genomic dataset

The dataset we use for genome-wide analysis is based on the prediction results obtained from 78 gram-negative genomes using a hidden Markov model of TM β-strand sequences developed by Bigelow *et al.*[10] The final dataset after cleaning-up (see Methods) consists of 7968 strands.

Sequence motifs and antimotifs within TM β-strands

We examine the two-residue sequence pattern *XYk* as the occurrence of a specific residue type *X* separated by *k* residues in the N-to-C direction from another specific residue type *Y in sequence order* along the strand. For example, AL3 represents the pattern of Ala to Leu in the N to C direction with 2 residues in between (AxxL), and AA1 is a pair of Ala residues immediately next to each other in sequence. We calculate *intrastrand residue pair propensity* as the odds ratio of the observed frequency of occurrences of sequence pattern *XYk* compared to the expected frequency, which is the frequency of this pattern that would occur by chance if the residues in a single strand are exhaustively permuted, and each permutation is equally likely. Senes *et al.* have already developed statistical tools based on enumeration for this null model,[16] as described in Methods. The new contributions of our study to this null model are a direct analytical form for calculation of the mean and an explicit probability distribution useful for *p*-value calculation in most cases.[21] Using this model and computational tools, we have calculated propensities and *p*-values measuring statistical significance for all possible *XYk* patterns in our genomic database, where *X* and *Y* are amino acid types from the alphabet of 20 amino acids, and $k = 1$–4. Table 1 lists statistically significant intrastrand motifs (propensity ≥ 1.10) and antimotifs (propensity ≤ 0.90) for different values of *k*, as well as *p*-values, with $p < 3.125 \times 10^{-5}$. This *p*-value is selected after correcting for multiple hypotheses; since there are $20 \times 20 \times 4 = 1600$ sequence patterns being tested, a Bonferroni-corrected *p*-value of $p = 3.125 \times 10^{-5}$ corresponds to an effective *p*-value of 0.05 (see below). Full tables of propensities, as well as observed and expected frequencies, are listed in Supplementary Material.

Several important intrastrand motifs emerge from this study. In general, Aliphatic-Tyr patterns are strongly favored. For example, AY2 is the most significant sequence motif (propensity 1.56, *p*-value 2×10^{-38}). Surprisingly, Tyr-Aliphatic patterns, in which the order of the two residues is reversed, are strongly disfavored. For example, YA2 is a significant antimotif (propensity 0.69, *p*-value 8×10^{-15}). Since $k = 2$, the residues are on the same side of the strand, and thus their side-chains are the closest physically compared to any other value of *k*. Our results suggest that there is a clear physical

Table 1. Pairwise intrastrand sequence motifs, drawn from a genome-wide distribution, with propensities and *p*-values listed

$k=1$			$k=2$			$k=3$			$k=4$		
Pair	Odds	*p*-Value	Pair	Odds	*p*-Value	Pair	Odds	*p*-Value	Pair	Odds	*p*-Value
Motifs											
GV	1.25	1.7×10^{-11}	AY	1.56	2.0×10^{-38}	GY	1.67	2.5×10^{-34}	LY	1.56	4.1×10^{-37}
WQ	1.53	8.7×10^{-9}	LA	1.31	1.6×10^{-29}	YP	1.95	3.9×10^{-13}	VY	1.47	2.9×10^{-14}
IG	1.25	1.5×10^{-7}	GV	1.29	1.3×10^{-12}	LG	1.23	3.5×10^{-12}	AY	1.39	5.1×10^{-14}
YA	1.22	4.1×10^{-7}	LG	1.23	2.3×10^{-12}	SY	1.31	4.6×10^{-9}	AQ	1.43	3.0×10^{-8}
YR	1.25	2.6×10^{-6}	VY	1.31	3.7×10^{-10}	AR	1.24	2.9×10^{-6}	SA	1.25	2.5×10^{-6}
KP	1.57	4.6×10^{-6}	VG	1.26	4.8×10^{-10}	TA	1.19	1.4×10^{-5}	VI	1.36	3.0×10^{-6}
RI	1.32	5.3×10^{-6}	VA	1.20	1.9×10^{-8}	LR	1.17	2.6×10^{-5}	LV	1.19	3.5×10^{-6}
GM	1.42	5.9×10^{-6}	LY	1.19	2.2×10^{-8}				GP	1.46	6.9×10^{-6}
FD	1.25	7.4×10^{-6}	AV	1.20	3.4×10^{-8}				GK	1.29	7.3×10^{-6}
GK	1.27	9.1×10^{-6}	LL	1.10	2.8×10^{-7}				FY	1.28	1.6×10^{-5}
YQ	1.28	1.4×10^{-5}	GY	1.24	1.6×10^{-6}				LA	1.14	2.6×10^{-5}
EL	1.19	1.5×10^{-5}	WP	1.75	1.6×10^{-6}						
LG	1.11	2.4×10^{-5}	YQ	1.43	2.4×10^{-6}						
			IG	1.25	8.2×10^{-6}						
			WI	1.42	1.9×10^{-5}						
Antimotifs											
FL	0.62	5.2×10^{-17}	YL	0.74	4.6×10^{-15}	YR	0.52	9.7×10^{-17}	YL	0.43	4.2×10^{-50}
IL	0.63	5.5×10^{-12}	YA	0.69	7.7×10^{-15}	YF	0.50	1.8×10^{-12}	YV	0.55	2.1×10^{-16}
LL	0.84	4.9×10^{-8}	LR	0.68	1.4×10^{-11}	AL	0.79	5.0×10^{-9}	YF	0.57	3.2×10^{-13}
YP	0.49	8.6×10^{-8}	FY	0.70	1.8×10^{-9}	PL	0.60	2.4×10^{-7}	YI	0.49	6.7×10^{-13}
WG	0.68	2.4×10^{-7}	HY	0.41	1.4×10^{-8}	RS	0.65	4.6×10^{-7}	YA	0.67	2.4×10^{-12}
HG	0.62	7.9×10^{-7}	RL	0.73	2.1×10^{-8}	YS	0.77	7.8×10^{-7}	YW	0.47	3.6×10^{-8}
TI	0.74	9.7×10^{-7}	WY	0.55	4.3×10^{-8}	PV	0.62	3.0×10^{-5}	RY	0.58	5.1×10^{-7}
VL	0.80	3.1×10^{-6}	HA	0.58	7.1×10^{-7}				YY	0.68	2.4×10^{-6}
FR	0.76	7.2×10^{-6}	YV	0.77	1.2×10^{-6}				AL	0.85	7.4×10^{-6}
FF	0.73	9.7×10^{-6}	YT	0.75	1.8×10^{-6}						
II	0.62	1.0×10^{-5}	YG	0.77	2.2×10^{-6}						
NF	0.77	1.6×10^{-5}	LI	0.81	1.9×10^{-5}						
PP	0.45	1.6×10^{-5}	IR	0.66	3.1×10^{-5}						
AH	0.64	2.1×10^{-5}									
IQ	0.69	2.1×10^{-5}									
FW	0.55	2.5×10^{-5}									

Only motifs significant at the threshold *p*-value of 3.125×10^{-5} are listed.

preference for AY2 and a preference against YA2. Examples of the AY2 motif are shown in Figure 1(a).

There are several additional motifs in which Tyr is the C-terminal (second) residue, as well as several antimotifs in which Tyr is the N-terminal (first) residue. In addition to the AY2-YA2 motif-antimotif dichotomy, there are eight similar complementing pairs containing Tyr: VY2-YV2, LY2-YL2, GY2-YG2, SY3-YS3, AY4-YA4, VY4-YV4, LY4-YL4, and FY4-YF4. The most significant pairs of this type involve the aliphatic residues Ala, Val, and Leu, with even k; all such pairs are significant. However, another aliphatic residue, Ile, is under-represented in these motifs, and appears in only one such pattern, the YI4 antimotif.

Although the Tyr-Aliphatic dichotomy may be due in part to the individual preference of Tyr for the C-terminal side of TM β-strands,[13] this preference alone does not fully account for several observed exceptions to the dichotomy: YA1, YR1, YQ1, YQ2, and YP3 are favored motifs in which Tyr is the first (i.e. N-terminal) residue, while FY2, WY2, PY2, HY2, RY4, and YY4 are disfavored antimotifs in which Tyr is the second (i.e. C-terminal) residue. Thus, effects other than individual residue preferences are responsible for the placement of Tyr in these intrastrand motifs. The only favorable Tyr-X motif when $k=2$ is YQ2 (propensity 1.43, *p*-value 2×10^{-6}). Visual inspection of PDB structures containing YQ2 patterns indicates that the polar side-chain of Gln comes in close proximity to the hydroxyl group of the Tyr residue, forming a stable side-chain hydrogen bond. Two examples of the YQ2 motif exhibiting this behavior are shown in Figure 1(b). This observation may offer an additional explanation of the rotamer preference of Tyr to extend in the N-C direction, as described by Chamberlain and Bowie.[15] It is also interesting to note that while Aliphatic-Tyr patterns are strongly favored when $k=2$, Aromatic-Tyr patterns (FY2, WY2, and HY2) are disfavored.

Propensity and sequence motifs in known structures

The propensities calculated in the first study are based on a large dataset of sequences predicted to be TM β-strands. Putative TM strands are used because the number of crystal structures of TM β-barrels is very small (<30). However, it is useful to attempt to identify motifs and antimotifs in a small dataset of known structures, as this helps to validate the predictions and propensities calculated from the genome-wide dataset and to provide structural rationalization for the resulting sequence motifs.

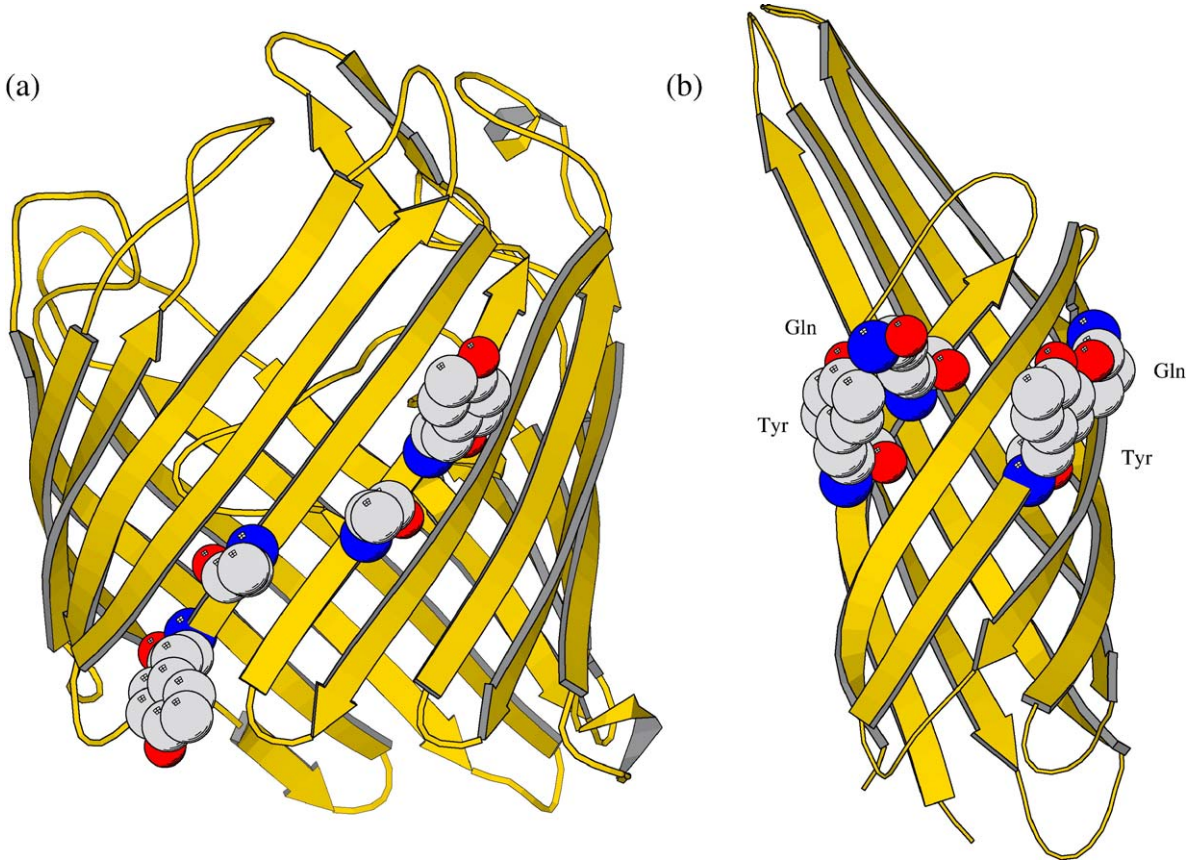

Figure 1. Two examples of intrastrand sequence motifs in β-barrel membrane proteins: (a) An instance of the AY2 motif in OmpF. The preference for tyrosine's side-chain to face the N-C direction is preserved. (b) Two instances of the YQ2 motif, an example of a Tyr-Polar motif, in OmpX. The side-chains of both residues in each pair form H-bonds.

We have compiled a dataset of 19 non-homologous β-barrel membrane proteins of known structure, consisting of 262 TM strands. We apply the same statistical methods to this list as to the genome-wide list of putative TM β-strands. Patterns significant at the level of $p < 0.05$ are listed in Table 2. Full tables of propensities, as well as observed and expected frequencies, are listed in Supplementary Material. Although there is some disagreement between these results and those of the genome-wide analysis, a majority of the Aliphatic-Tyr dichotomy patterns is preserved in the structural analysis (AY2, YA2; VY2, YV2; YL2; LY4, YL4; and YV4). The significant AY2 pattern occurs in 15 of the 19 proteins in the dataset. Additionally, YQ2 is also a favored motif (propensity 1.74, p-value 0.05), and one Aromatic-Tyr antimotif, WY2, is preserved (propensity 0.18, p-value 0.03). Overall, propensity values tend to be more extreme in the structural dataset (e.g. 1.75 *versus* 1.56 for AY2), although p-values are less significant. This is not surprising for such a small dataset, in which significant patterns may be more obscure due to the small sample size.

Correction for multiple hypothesis testing

Because the propensities calculated in this study represent all possible ordered pairwise residue combinations, we have calculated p-values for 400 (20×20) hypothesis tests for each k value. Specifically, these are 2-sided tests, in which the null hypothesis is that the propensity is 1.00 and the alternate hypothesis is that the propensity is significantly different from 1.00 (higher or lower). The danger in using so many tests is that there is a high probability that some of the significant p-values discovered resulted from random chance alone. At a significance level of $p < 0.05$, we would naïvely expect 20 (400×0.05) tests to have significant results by chance alone for each k value.

To resolve this problem for the genome-wide analysis, a simple Bonferroni correction is employed, which is to adjust the p-value cut-off for significance by dividing by the total number of tests.[23] Since we use a p-value of 0.05 and have a total of 1600 tests (400×4 for the number of k values), we use an adjusted p-value of $0.05/1600 = 3.125 \times 10^{-5}$. After this correction, there are 91 significant patterns adhering to this stringent requirement, as listed in Table 1.

However, for the analysis on the smaller structural dataset, there are no more than 20 statistically significant results at the level of $p < 0.05$ for each of the k values 1-4 in Table 2. Thus, a more robust analysis is needed to justify the significance of these results. We therefore calculate the false discovery

Table 2. Pairwise intrastrand sequence motifs, drawn from a dataset of known TM β-barrel structures, with propensities and *p*-values listed

$k=1$			$k=2$			$k=3$			$k=4$		
Pair	Odds	*p*-Value	Pair	Odds	*p*-Value	Pair	Odds	*p*-Value	Pair	Odds	*p*-Value
Motifs											
GV	1.59	5.7×10^{-4}	GR	2.14	5.6×10^{-6}	GY	1.81	5.0×10^{-5}	LY	1.90	8.2×10^{-5}
SY	1.58	9.2×10^{-3}	AY	1.75	5.6×10^{-4}	TM	3.25	2.3×10^{-3}	WV	2.79	1.1×10^{-3}
GL	1.37	2.0×10^{-2}	LG	1.63	1.0×10^{-3}	WG	2.09	1.0×10^{-2}	TY	1.93	9.3×10^{-3}
VG	1.37	2.5×10^{-2}	LA	1.61	1.1×10^{-3}	GR	2.00	1.5×10^{-2}	TG	1.68	1.0×10^{-2}
EM	2.58	2.8×10^{-2}	AA	1.47	2.3×10^{-2}	AG	1.45	1.8×10^{-2}	GD	1.86	1.3×10^{-2}
RY	1.71	3.9×10^{-2}	IL	1.73	2.7×10^{-2}	VY	1.69	2.3×10^{-2}	HR	5.27	2.0×10^{-2}
VK	1.72	4.0×10^{-2}	ND	2.16	3.0×10^{-2}	EL	1.70	3.6×10^{-2}	GN	1.88	2.9×10^{-2}
TW	1.74	4.2×10^{-2}	VY	1.43	3.2×10^{-2}	VA	1.60	4.2×10^{-2}	VG	1.60	3.3×10^{-2}
LG	1.31	4.3×10^{-2}	IA	1.60	3.8×10^{-2}	AW	2.01	4.6×10^{-2}	FA	1.75	3.6×10^{-2}
TV	1.51	4.9×10^{-2}	KW	3.31	4.0×10^{-2}	GQ	1.85	4.9×10^{-2}	IG	1.75	4.5×10^{-2}
			VP	2.24	4.2×10^{-2}						
			YQ	1.74	4.9×10^{-2}						
Antimotifs											
VY	0.36	3.7×10^{-3}	YA	0.35	6.9×10^{-4}	YG	0.55	1.0×10^{-2}	YF	0.15	8.9×10^{-3}
SS	0.31	2.4×10^{-2}	YV	0.46	3.1×10^{-3}	YE	0.16	1.7×10^{-2}	AR	0.00	1.3×10^{-2}
YY	0.50	2.5×10^{-2}	YT	0.36	1.2×10^{-2}	GS	0.32	1.9×10^{-2}	YL	0.50	1.6×10^{-2}
DG	0.18	2.8×10^{-2}	YL	0.62	2.5×10^{-2}	KA	0.17	2.3×10^{-2}	GV	0.42	2.3×10^{-2}
MT	0.00	3.1×10^{-2}	WY	0.18	2.9×10^{-2}	WQ	0.00	2.8×10^{-2}	YV	0.48	2.3×10^{-2}
YW	0.00	3.6×10^{-2}	VK	0.00	3.4×10^{-2}	YR	0.28	3.1×10^{-2}	TI	0.18	3.5×10^{-2}
RA	0.36	4.0×10^{-2}	EY	0.00	4.4×10^{-2}	NW	0.00	3.9×10^{-2}	KT	0.00	4.0×10^{-2}
TD	0.19	4.4×10^{-2}				FT	0.36	4.2×10^{-2}	YI	0.29	4.1×10^{-2}
LV	0.47	4.7×10^{-2}				LW	0.00	4.2×10^{-2}	YQ	0.19	4.8×10^{-2}
						YQ	0.37	4.7×10^{-2}			

Only motifs significant at the threshold *p*-value of 0.05 are listed.

rate (FDR), which is the expected proportion of test results incorrectly declared significant.[24]

To calculate the FDR for a certain k value, we follow an approach similar to that used in the Significance Analysis of Microarrays.[25] We randomly scramble the residues within our dataset among all strands and then recalculate *p*-values. We replicate this process 1,000 times and average the number of significant *p*-values ($p<0.05$) detected in each replicate. The results are listed in Table 3. The FDRs calculated for our dataset for intrastrand sequence motifs and antimotifs range from 38–45%, and suggest that 7–9 of the results found for each k value may be erroneously declared statistically significant by chance alone.

Propensity and sequence motifs in full proteins

Sequence motifs and antimotifs in TM strands may not be the only important ones in β-barrel membrane proteins. It is possible that motifs also occur in non-TM regions, or straddle across two different TM strands. For instance, a recent study shows that the periplasmic molecular chaperone SurA preferentially binds to an Aromatic-x-Aromatic motif, which may aid in its binding to outer membrane β-barrel proteins.[18] However, our analysis shows that there are no favorable Aromatic-x-Aromatic sequence motifs (when $k=2$) in TM strands, suggesting that if this pattern is favored in β-barrel membrane proteins, it does not occur in TM strands.

Table 3. False discovery rates, at different values of k, for motif discovery in the dataset of TM strands from known structures

k	Random	Actual	FDR (%)
1	8.5	19	45
2	7.9	19	42
3	7.8	20	39
4	7.3	19	38

To further investigate this, we apply the same sequence motif model to entire peptide sequences in the dataset of known structures. The only difference is that instead of using single strands, we use the full protein chain. Although the protein chains are too long to allow for the exact calculation of *p*-values, it is possible to calculate the exact odds ratios and approximate *z*-values, from which approximate *p*-values can be estimated (see Methods). Table 4 shows significant motifs and antimotifs from this analysis when $k=2$. Full tables of propensities, as well as observed and expected frequencies, are listed in Supplementary Material. The most significant motif is GR2 (odds ratio 1.70, approximate *p*-value 2×10^{-5}), which is also the most significant motif in the TM-only study of known structures. The second most significant motif, YF2 (odds ratio 1.97, approximate *p*-value 7×10^{-5}), is of the type Aromatic-x-Aromatic, and occurs in 16 of the 19 proteins in the dataset, a total of 33 times. Visual inspection reveals that this motif occurs frequently near the C-terminal end of the barrel, and that the α-carbon of the Phe residue is often in the periplasm, hence excluding

Table 4. Odds ratios for full protein sequence motif analysis when $k=2$ for the dataset of known structures

Motifs			Antimotifs		
Pair	Odds	*p*-Value	Pair	Odds	*p*-Value
GR	1.70	2.1×10^{-5}	GP	0.49	1.5×10^{-2}
YF	1.97	6.8×10^{-5}	RQ	0.36	2.0×10^{-2}
VY	1.68	5.4×10^{-4}	VK	0.47	2.3×10^{-2}
EG	1.57	7.7×10^{-4}	EY	0.46	2.6×10^{-2}
AT	1.49	8.9×10^{-4}	YM	0.00	2.8×10^{-2}
PS	1.79	1.5×10^{-3}	YS	0.61	2.8×10^{-2}
RD	1.62	3.5×10^{-3}	KL	0.55	3.0×10^{-2}
NI	1.65	4.9×10^{-3}	RR	0.42	3.5×10^{-2}
HH	3.68	6.7×10^{-3}	SA	0.69	3.8×10^{-2}
YQ	1.64	8.0×10^{-3}	YT	0.63	3.9×10^{-2}
VL	1.42	1.9×10^{-2}	WY	0.25	4.6×10^{-2}
SF	1.50	2.2×10^{-2}	EP	0.32	4.6×10^{-2}
EK	1.63	2.3×10^{-2}			
LA	1.34	2.4×10^{-2}			
KQ	1.63	2.6×10^{-2}			
WP	2.13	3.4×10^{-2}			
GE	1.36	3.9×10^{-2}			

Only patterns with a *p*-value less than 0.05 are listed.

this pattern from the TM-only analysis. Struyvé *et al.* discovered a characteristic 10-residue C-terminal pattern that terminated with Phe in many bacterial outer membrane proteins, and showed that mutation of the terminal Phe impaired proper assembly into the outer membrane.[22] This is consistent with the hypothesis that the YF2 motif is biologically important.

Although YF2 is a strong motif, it is the only favorable Aromatic-x-Aromatic motif. In fact, another Aromatic-x-Aromatic pattern is an antimotif (WY2, odds ratio 0.25, *p*-value 0.05). Combined with experimental data,[18] this suggests that the chaperone SurA binds to the YF2 motif specifically, and that its affinity for other Aromatic-x-Aromatic sequences has minimal effects because of the sparsity of these other motifs (such as WY2).

The experimental study that discovered the Aromatic-x-Aromatic binding motif also found an Aromatic-x-Pro motif.[18] These motifs were found in tandem (Aromatic-x-Aromatic-x-Pro). A favorable Aromatic-x-Pro motif is found in our study, WP2 (odds ratio 2.13, *p*-value 0.03). This motif is occasionally found on the periplasmic side of the protein, where Trp is a component of the aromatic girdle and Pro participates in short loops. This position is in the same vicinity as the YF2 motif mentioned above, although there are only six instances of a full Aromatic-x-Aromatic-x-Pro motif in the structural dataset.

There are several motifs and antimotifs that appear on both the list of known TM strand sequence patterns and the list of full protein sequence patterns, e.g. motifs GR2, VY2, YQ2, and LA2, and antimotifs VK2, EY2, YT2, and WY2, when $k=2$. This suggests that these preferences either are not isolated to TM strands, or that they are frequent enough in TM strands that their significance is not attenuated by the addition of non-TM regions to the analysis. In contrast, the AY2 motif and several Tyr-Aliphatic antimotifs from the TM strand analysis are missing in the full protein analysis. AY2 has an odds ratio of only 1.30 and an approximate *p*-value of 0.09, below the significance threshold of 0.05. This suggests that the complementing motif-antimotif pairs involving Tyr are specific to the TM regions, and thus may be more important for transmembrane stability and less relevant for recognition by molecular chaperones.

Effect of position-dependent individual residue bias

The amino acid composition of the two datasets (genomic and known structures) used in the study of intrastrand pairwise propensities is described in Table 5. In addition to amino acid frequencies for the full datasets, Table 5 also displays frequencies for three cross-sections of the strands in the datasets: the N-terminal, central, and C-terminal thirds.

Overall, the amino acid composition of the smaller dataset derived from known structures is more skewed, showing a bias of Thr for the N-terminal third, a bias of Ala and Gly for the central third, and of Tyr for the C-terminal third. Similar biases are seen in the genomic dataset, but to a much smaller degree. These residue biases have been documented in β-barrel membrane proteins in previous studies.[7,8,13] In addition, it is also well-known that residues whose sidechains face the interior of the β-barrel have very different amino acid preferences from external residues facing the surrounding lipid bilayer.[7,8]

The existence of these position-dependent individual residue preferences may affect our study of pairwise propensities. For example, the bias of Ala for the central third of the strand and of Tyr for the C-terminal third may increase the propensity of the AY2 motif without providing extra information about the relationship between the two residues. It is also possible that the situation may work in reverse, and the high preference of AY2 pairs increases the individual positional preferences of Ala and Tyr.

To determine whether either situation has a confounding effect on our statistical study, we have taken two measures. First, we have modified our statistical model to treat internal and external residues separately, as described in Methods. The propensities in Tables 1 and 2 are presented after this correction. Second, we have developed another statistical model to examine the effect of single-residue positional bias. Whereas the null model described earlier is based on the exhaustive permutation of residues within a strand, this new null model is based on the permutation of residues across strands that are in the same position within their strands.

We are most concerned about the effect of the positional bias of Tyr, as it occurs in the most significant pairs in our study. We observed the

Table 5. Amino acid composition, in percent, for both the genome-wide dataset of putative TM strands, comprising 74,380 residues, and the dataset of TM strands from known structures, comprising 2565 residues

	Genome-wide dataset				Structure dataset			
A.A	N-term	Central	C-term	Whole	N-term	Central	C-term	Whole
A	9	12	8	10	6	15	6	9
C	0	0	0	0	0	0	0	0
D	4	3	4	4	3	3	5	4
E	3	3	3	3	4	2	3	3
F	7	5	6	6	6	3	6	5
G	9	11	8	9	9	17	9	12
H	1	1	2	2	2	1	2	1
I	4	4	4	4	6	4	4	5
K	3	3	4	4	4	2	4	3
L	13	12	9	12	10	13	7	10
M	2	2	1	2	2	2	2	2
N	4	4	6	5	4	3	4	4
P	2	2	3	2	1	1	1	1
Q	3	3	5	4	4	3	7	4
R	5	4	6	5	3	3	5	4
S	7	7	7	7	6	7	6	6
T	8	7	8	8	10	6	6	8
V	7	8	6	7	9	9	5	8
W	3	2	3	2	4	1	4	3
Y	4	6	8	6	6	5	15	9

Compositions are calculated for the first (N-term.), second (Central), and last (C-term) thirds of each strand, as well as the whole strand (Whole).

results of using this new model on Aliphatic-Tyr motifs and Tyr-Aliphatic antimotifs when $k=2$, listed in Table 6 for the genomic database. Individual residue bias does not have a confounding effect on the motifs studied: the propensity of AY2 decreased slightly (1.56 to 1.46), while LY2 increased (1.19 to 1.40) and VY2 remained unchanged at 1.31. There is, however, a noticeable confounding effect on antimotifs: all propensities increased, and one (YL2) increased past 1.00. Full tables of propensities based on the positional model for both datasets are listed in Supplementary Material.

Discussion

We have described a number of sequence patterns for β-barrel membrane proteins in this study. The occurrence of some of them can be rationalized by the physicochemical properties of peptides and lipids as revealed in previous studies.[8,26] However, many of the patterns discovered in this study may provide helpful novel information about the assembly and folding process of β-barrel membrane proteins.

Table 6. Comparison of propensities derived from different null models when $k=2$ for the genome-wide dataset

Motifs			Antimotifs		
Pair	Strand	Pos.	Pair	Strand	Pos.
AY	1.56	1.46	YA	0.69	0.88
LY	1.19	1.40	YL	0.74	1.12
VY	1.31	1.31	YV	0.77	0.98

The original null model (Strand) is based on the permutation of residues within individual strands, while the positional null model (Pos.) is based on the permutation of residues across strands but in the same posision on the strand.

Role of tyrosine

Among the most significant sequence patterns, many involve the amino acid tyrosine. It is part of several complementing sequence motif-antimotif pairs (e.g. AY2-YA2, VY2-YV2, and LY4-YL4). All aromatic residues have important roles in the aromatic girdle of β-barrel membrane proteins, but Tyr stands out as the most unique.

Rotamer preferences provide some explanation for the unique properties of Tyr. Chamberlain and Bowie determined that Tyr has a distinct preference to form the (180,90) rotamer in TM β-barrels, usually occurring at the C-terminal end of TM strands.[15] This rotamer aligns the side-chain in the N-C direction so that it nearly coincides with the membrane normal, thus maximizing the distance of the hydroxyl group from the center of the membrane bilayer and resulting in the most stable placement of the amino acid. This preference is shown in Figure 1(a), in which Tyr adopts the (180,90) rotamer in both the AY2 and YQ2 motifs.

The tendency for Tyr to adopt an N-C rotamer may explain some of the significant sequence patterns discovered in this study, or, alternatively, the patterns help to explain the rotamer preference. In the (180,90) rotamer, Tyr extends its polar hydroxyl group toward the C terminal, and leaves its nonpolar aromatic group relatively closer to the N terminal. Therefore, it would be reasonable to expect aliphatic residues such as Ala, Val, and Leu to be on the N-terminal side of Tyr, and polar residues such as Gln to be on the C-terminal side. However,

the small separation ($k=2$ or 4) between residues may suggest a further relationship, such as protein-protein or protein-lipid interactions. For example, the aliphatic residue may contact the nonpolar tail of a membrane lipid while the Tyr residue contacts the polar head. Also of note is the absence of Ile in this analysis, which, despite being aliphatic, may have a side-chain unsuitable for such interactions with lipids. Any explanation of this phenomenon should also further incorporate the observation of Aromatic-Tyr antimotifs when $k=2$ (FY2, HY2, and WY2).

Comparison to soluble β-sheets and TM α-helical proteins

Senes *et al.* studied pairwise sequence motifs in α-helical membrane proteins.[16] The most significant motifs and antimotifs contain aliphatic residues and Gly, which are the most abundant residues in these proteins. Several sequence motifs and antimotifs in β-barrel membrane proteins also include these residues, but no discernible relationship exists between the results of these two analyses. The different structural properties of α-helices and β-strands most likely result in different motifs and emphasize different residue separations. For instance, residue side-chains are closest to each other at $k=4$ in α-helices, but are closest at $k=2$ in β-strands.

Sequence motifs play important roles in α-helical membrane protein folding, where motifs such as GxxxG, AxxxA, and those containing Ser and Thr are known to promote the dimerization of TM α-helices.[16,17,27,28] Intrastrand sequence motifs in β-barrel membrane proteins may play important roles as well. Unlike TM helices, which may interact with up to 5-6 other helices, each TM strand only interacts with two other strands, one on each side. If a sequence motif is essential for the thermodynamic stability of a barrel protein, its mutation is likely to have a more profound direct consequence which may lead to observable changes in protein stability or flexibility near the region of the mutated strands. This may be different from helical proteins, where the contribution of one sequence motif is necessarily modified by the interactions of the helix with other helices.

The importance of sequence motifs in chaperone binding

The observation of assisted *in vivo* folding by the periplasmic chaperone SurA[18] implies that there may be sequence motifs that are recognized by chaperones. Our results suggest that the YF2 motif is a statistically significant motif that might be specifically recognized by SurA in *E. coli* and other bacteria. Although the panning of a phage-display peptide library selects the strongest peptide binder, biological systems may have relatively labile interactions for chaperone activity to ensure the timely release of the peptide, and it is likely that some of the motifs identified in this study may be relevant for efficient and rapid *in vivo* chaperone binding.

Struyvé *et al.* discovered a characteristic 10-residue C-terminal pattern in many bacterial outer membrane proteins.[22] Among 30 such proteins, Phe occupies the C-terminal position in 28 proteins, and Tyr occupies the third position from the C-terminal in 18.[22] Therefore, the YF2 motif is a common feature in many outer membrane proteins. Mutation or deletion of the terminal Phe in the β-barrel membrane protein PhoE from *E. coli* results in a dramatic impairment of the protein's ability to assemble into outer membranes correctly, though it does not affect transport across the inner membrane.[22] Since SurA is located in the periplasm between the inner and outer membranes, this finding is consistent with the hypothesis that the SurA chaperone recognizes the YF2 motif.

The analysis of motifs in the loop regions is based on the set of 19 β-barrel membrane proteins with known structures. A natural extension would be to use a genomic database of predicted loops. It is expected that with a large amount of data, more subtle motifs in the loop regions might be uncovered.

Experimentally testable hypotheses

Mutational studies that measure the structural stability or folding behavior of β-barrel membrane proteins in which sequence motifs are substituted may elucidate their roles. For example, to test the hypothesis that the AY2 motifs are important for *in vivo* sorting of β-barrel membrane proteins but not for intrinsic protein stability, one can measure the *in vitro* folding behavior of mutants where AY2 motifs are changed to YA2 antimotifs. If such mutants fold normally in a test tube, it would suggest that Aliphatic-x-Aromatic motifs play roles mostly for sorting and *in vivo* folding.

Another type of experiment might be the use of double double-mutants. When two high-propensity pairwise motifs are interacting spatially on neighboring strands (e.g. I355-L357 and A377-Y379 in LamB), one can replace these pairs of motifs with low-propensity pairs of antimotifs. Experimental assays on the folding and sorting behavior of these mutant proteins will help to clarify their roles in maintaining protein stability and in promoting *in vivo* folding. Additional mutation studies on motifs in the loop region will further help to specify the role of sequence motifs in SurA chaperone binding.

Summary

In this study, we have developed statistical models for the discovery of sequence motifs in the strands and loop regions of β-barrel membrane proteins. Our results show that there are strong motifs and antimotifs in transmembrane β-strands. The amino acid Tyr plays an important role in such

motifs. A general dichotomy consists of favorable Aliphatic-Tyr sequence motifs and unfavorable Tyr-Aliphatic antimotifs. We also find that the terminal motif YxF may be an important part of a sequence recognition pattern for chaperone binding. Our results also suggest several experiments that can help to elucidate the mechanisms of *in vitro* and *in vivo* folding of β-barrel membrane proteins.

Model and Methods

Datasets

The dataset for genome-wide analysis is based on 3,171 proteins predicted to contain transmembrane β-barrels by a hidden Markov model (HMM) developed by Bigelow *et al.*[10] We use only those proteins listed as integral outer membrane proteins or those with high sequence similarity to integral outer membrane proteins in the authors' online database (PROFtmb, http://www.rostlab.org/services/PROFtmb/). We then extract subsequences predicted to be TM strands. Altogether, we obtain 15,946 putative TM strands.

The subsequences extracted in this way have very different lengths. We limit strands to 10 residues in length because most strands in the dataset of known structures (Table 7) are 10 residues in length or shorter. If a subsequence is longer than 10 residues, we choose the 10 residues closest to the periplasmic side of the protein, because this side usually contains short β-turns that clearly delineate the ends of the strands.

In order to eliminate individual strand sequences that share high sequence similarity with other strands, we use the transmembrane-specific substitution matrix PHAT.[29] We require that no two strands share a pairwise, gapless similarity score higher than 4.5 per residue. This similarity cut-off reduces the original 15,946 strands nearly in half to 7,968 strands without high pairwise similarity, which we use for our genome-wide analysis.

The second dataset based on β-barrel membrane proteins of known structure comprises 19 structures found in the Protein Data Bank (Table 7), totaling 262 β-strands. All proteins share no more than 26% pairwise sequence identity. All structures have a resolution of 2.6 Å or better. To determine which residues in the dataset are transmembrane, the coordinates in the protein's PDB file were translated and rotated so that the *xy*-plane was perpendicular to the vertical axis of the barrel and equidistant to the observed aromatic girdles presumed to be at the membrane interfaces. A residue is declared to be transmembrane if it is located in a β-strand and the *z*-coordinate (vertical distance from bilayer center) of its associated α-carbon is between −13.5 Å and 13.5 Å.

Table 7. Dataset of 19 β-barrel membrane proteins used for this study

Protein	Organism	Architecture	Strands	PDB ID
OmpA	*E. coli*	monomer	8	1BXW[30]
OmpX	*E. coli*	monomer	8	1QJ8[31]
NspA	*N. meningitidis*	monomer	8	1P4T[32]
OpcA	*N. meningitidis*	monomer	10	1K24[33]
OmpT	*E. coli*	monomer	10	1I78[34]
OMPLA	*E. coli*	dimer	12	1QD6[35]
NalP	*N. meningitidis*	monomer	12	1UYN[36]
Porin	*R. capsulatus*	trimer	16	2POR[37]
Porin	*R. blastica*	trimer	16	1PRN[38]
OmpF	*E. coli*	trimer	16	2OMF[39]
Omp32	*C. acidovorans*	trimer	16	1E54[40]
LamB	*S. typhimurium*	trimer	18	2MPR[41]
ScrY	*S. typhimurium*	trimer	18	1A0S[42]
FepA	*E. coli*	monomer	22	1FEP[43]
FhuA	*E. coli*	monomer	22	2FCP[44]
FecA	*E. coli*	monomer	22	1KMO[45]
BtuB	*E. coli*	monomer	22	1NQE[46]
TolC	*E. coli*	trimer	4	1EK9[47]
α-Hemolysin	*S. aureus*	heptamer	2	7AHL[48]

All proteins share no more than 26% pairwise sequence identity. Crystal structures have a resolution of 2.6 A or less. Three identical chains of TolC and seven of α-hemolysin form a single barrel; all other proteins listed form whole barrels with each peptide chain.

Propensity of intrastrand two-residue sequence patterns

We introduce the propensity $P(X, Y \mid k)$ for two ordered intrastrand residues of type X and type Y that are k positions away on the same strand. For example, when $k=1$, $P(X, Y \mid 1)$ represents the propensity that an X residue is immediately followed by a Y residue on a β-strand. We define the propensity as:

$$P(X, Y|k) = \frac{f(X, Y|k)}{\mathbb{E}[f'(X, Y|k)]}, \quad (1)$$

where $f(X, Y \mid k)$ is the observed frequency of XYk patterns in the TM region, and $\mathbb{E}[f'(X, Y|k)]$ is the expected frequency.

The calculation of $\mathbb{E}[f'(X, Y|k)]$ depends on the null model. Here we choose a null model in which the residues within each strand are permuted exhaustively and independently, and each permutation occurs with equal probability. In this null model, an XYk pattern forms if in a permuted strand an X residue happens to be followed by a Y residue at the k-th position down the strand in the N-C direction. $\mathbb{E}[f'(X, Y)]$ is then the expected number of XYk patterns over the entire dataset. This expectation is calculated for a single strand as

$$\mathbb{E}[f'(X, Y|k)] = \frac{xy(l-k)}{l(l-1)}, \quad (2)$$

where l is the length of the strand, x is the number of residues of type X , and y is the number of residues of type Y.

To illustrate this, we can represent $f'(X, Y \mid k)$ as the sum of identical Bernoulli variables $f'(1, Y \mid k)$, each of which equals 1 if one of the y residues of type Y is in the k-th position past a specific residue of type X when the strand is randomly permuted, or 0 if the k-th position is not a Y residue. The probability that the residue of type X is placed in one of the first $l-k$ positions is $(l-k)/l$. If it were placed in one of the last k positions, there would not be enough space for an XYk motif to form. The probability that one of the y residues of type Y is placed in the k-th position past the residue of type X once the latter has been placed is $y/(l-1)$. Thus,

$$\mathbb{E}[f'(1, Y|k)] = \mathbb{P}_{1.Y|k}(1) = \frac{(l-k)}{l} \cdot \frac{y}{(l-1)}.$$

There are x such identical variables (one for each residue of type X), and the expectation of their sum is the sum of their expectations, leading to equation (2).

For *XXk* motifs, *i.e.* two residues of the same type displaced by k residues, the expectation is calculated as

$$\mathbb{E}[f'(X,X|k)] = \frac{x(x-1)(l-k)}{l(l-1)},$$

as there will be $x-1$ residues available to form an *XXk* motif with a specific residue of type *X*. Although these Bernoulli random variables are dependent (*i.e.* the placement of one *XYk* motif will affect the probability of another *XYk* motif), the expectation of their sum is the sum of their expectations, because expectation is a linear operator.

In order to calculate statistical significance in terms of *p*-values, we must determine $\mathbb{P}_{X,Y|k}(i)$, the probability of the occurrence of $i=f'(X,Y\,|\,k)$ *XYk* motifs. Analytical formulae to calculate $\mathbb{P}_{X,Y|k}(i)$ are complex,[21] but because all of the strands in the dataset are short, it is possible to fully enumerate all permutations to determine $\mathbb{P}_{X,Y|k}(i)$, as was done by Senes *et al.* for TM α-helices.[16] Once probability distributions are calculated for each strand, a combined dataset probability distribution can be obtained using the method of Senes *et al.*[16] Two-tailed *p*-values can then be calculated directly from the dataset probability distribution.

Confounding between conformational residue preferences and intrastrand sequence motifs

There is a strong possibility that intrastrand two-body sequence motifs may be affected by individual residue propensities. We take two measures to examine this potential effect. First, we correct our statistical model for conformational preference (*i.e.* whether the residue's sidechain is facing into the β-barrel or away from it), described here. Second, we observe the effects of a different null model that focuses on position-dependent residue preferences, described below.

In order to correct for conformational residue preferences, we divide each strand into two "substrands" which contain only residues facing the same direction. For even values of k, the substrands can be treated as independent strands, since both residues in each motif will face the same direction and thus occur in the same substrand. For each substrand, the effective k will be half of the original k.

For odd values of k, however, the two residues in a motif will face different directions. Thus, the two substrands derived from a strand must be permuted individually and then "shuffled" back into one strand in order to determine null hypothesis propensities (Figure 2). In this model, every possible combination of each permutation of one substrand and each permutation of its partner substrand is considered equally likely.

In order to calculate $\mathbb{E}[f'(X,Y|k)]$ when k is odd under these conditions, one must calculate two separate expected values and sum them, one for the case when the first residue of a pair is in an odd position of the strand, and one when it is in an even position. Start with the first case and let x_o be the number of residues of type *X* in the odd positions of the strand, and y_e be the number of residues of type *Y* in the even positions of the strand. Then

$$\mathbb{E}[f'(X,Y|k)_o] = \frac{x_o y_e \left\lceil \frac{l-k}{2} \right\rceil}{\left\lceil \frac{l}{2} \right\rceil \left\lfloor \frac{l}{2} \right\rfloor},$$

where $\lceil x \rceil$ represents the *ceiling* function (which equals the lowest integer higher than or equal to x) and $\lfloor x \rfloor$ represents the *floor* function (which equals the highest integer lower than or equal to x). Similar to the derivation for equation (2), this is the sum of x_o Bernoulli variables $f'(1,Y\,|\,k)_o$, each of which equals 1 if one of the y_e number of residues of type *Y* is in the k-th position past a specific residue of type *X*, and 0 otherwise. The probability that the residue of type *X* is placed in one of the first $l-k$ positions is $\lceil l-k/2 \rceil / \lceil l/2 \rceil$, since the x_o residues can only be placed in the $\lceil l/2 \rceil$ odd-numbered positions of the strand. Likewise, the y_e residues of type *Y* can only be placed in the $\lfloor l/2 \rfloor$ even-numbered positions of the strand, and thus the probability that a residue of type *Y* is in the k-th position past an appropriately placed residue of type *X* is $y_e/\lfloor l/2 \rfloor$.

For the second case, in which the residue of type *X* is placed in an even position, the expected value is similar:

$$\mathbb{E}[f'(X,Y|k)_e] = \frac{x_e y_o \left\lfloor \frac{l-k}{2} \right\rfloor}{\left\lfloor \frac{l}{2} \right\rfloor \left\lceil \frac{l}{2} \right\rceil}.$$

Summing the two expected values and simplifying results in the final expected value:

$$\mathbb{E}[f'(X,Y|k)] = \frac{x_o y_e \left\lceil \frac{l-k}{2} \right\rceil + x_e y_o \left\lfloor \frac{l-k}{2} \right\rfloor}{\left\lceil \frac{l}{2} \right\rceil \left\lfloor \frac{l}{2} \right\rfloor}.$$

When $X=Y$, it is necessary only to replace y_e with x_e and y_o with x_o. Simplifying results in the final expected value:

$$\mathbb{E}[f'(X,X|k)] = \frac{x_o x_e (l-k)}{\left\lceil \frac{l}{2} \right\rceil \left\lfloor \frac{l}{2} \right\rfloor}.$$

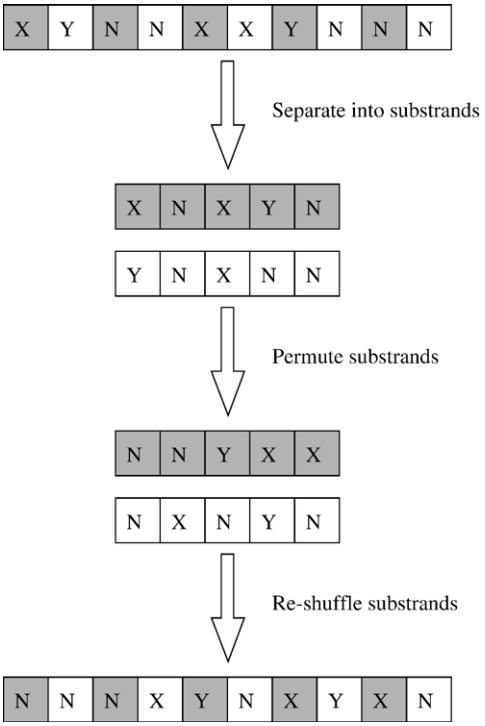

Figure 2. Graphical example of the correction for single-residue conformational preferences in intrastrand sequence pattern analysis when k is odd. In the first step, alternating residues are separated into substrands, in which all residues face the same direction (internal or external). In the second step, an example of one permutation is performed on each substrand. In the third step, the substrands are re-shuffled back into one strand. Each combination of every permutation of each strand is considered equally likely in our null model.

We then calculate propensity as in equation (1). In order to calculate two-tailed *p*-values, we use the method of Senes *et al.* as described above.[16] The results reported in Tables 1 and 2 are obtained after such corrections.

Propensity of intrastrand two-residue sequence patterns for full proteins

The analysis of two-residue sequence patterns can be performed on whole peptide sequences just as with short TM strands using the same model. The difference is that l is now the length of the full protein, and x and y are the numbers of residues of type X and Y, respectively, in the full protein.

However, because x, y, and l are usually too large to allow a full enumeration of permutations, it is not possible to calculate *p*-values exactly for motifs and antimotifs. Nevertheless, if l is much larger than x and y, an approximation using the binomial distribution will be useful. Recall that $f'(X, Y \mid k)$ can be represented as the sum of x identical yet dependent Bernoulli variables $f'(1,Y \mid k)$, with $\mathbb{P}_{1,Y|k}(1) = \frac{y(l-k)}{l(l-1)}$ and $\mathbb{P}_{1,X|k}(1) = \frac{(x-1)(l-k)}{l(l-1)}$. If l is much larger than x and y, the dependence between these variables will be small, and their sum can be approximated as a binomial distribution, with the mean as calculated in equation (2) and the variance:

$$\mathrm{var}[f'(X,Y|k)] = x \cdot \mathbb{P}_{1,Y|k}(1) \cdot [1-\mathbb{P}_{1,Y|k}(1)].$$

If the proteins in the dataset are assumed to be uncorrelated, as is likely the case since their sequences have pairwise identity ≤26%, it is possible to calculate a mean and variance for the entire dataset by summing those values over all proteins. It is then possible to calculate *z*-values as:

$$z(X,Y|k) = \frac{f'(X,Y|k) - \mathbb{E}[f'(X,Y|k)] \pm 0.5}{\sqrt{\mathrm{var}[f'(X,Y|k)]}},$$

where the 0.5 factor is a correction for continuity. These *z*-values can be compared to a standard Gaussian distribution in order to estimate *p*-values. Because full proteins contain more than just TM sequences, the correction for single-body propensities (i.e. separating out "substrands") is not applied here.

Position-dependent null model

In order to determine whether position-dependent individual residue propensities have a confounding effect on pairwise propensities, we have developed a separate positional null model. Instead of using exhaustive permutation within individual strands to obtain $\mathbb{E}[f'(X,Y|k)]$, we permute residues at a specific position across all strands (with replacement). We first define the *positional residue frequency* x_p as the number of residues of type X occupying the *p-th* position of a strand in the database. Because not all strands have the same length, we must normalize p to be within the range of 1-10, to approximate the average strand length of ~10:

$$p = \lceil \frac{10(p_{obs} - 0.5)}{l} \rceil,$$

where p_{obs} is the actual position of the residue within its strand and l is the length of the strand. This ensures that $1 \le p \le 10$. According to our null model, the probability of an arbitrary residue pair k residues apart in a strand being an XYk pattern is the expected value of the probability at a specific position p taken over all 10–k possible pair positions:

$$\mathbb{P}_{X,Y|k}(1) = \sum_{p=1}^{10-k} \frac{x_p}{n_p} \cdot \frac{y_{p+k}}{n_{p+k}} / (10-k),$$

where n_p is the number of all residues of all types in position p on all strands. To obtain $\mathbb{E}[f'(X,Y|k)]$, we multiply $\mathbb{P}_{X,Y|k}(1)$ by the number of all pairs of all residue types k residues apart in the dataset. We then calculate propensity as in equation (1). However, no analytical probability distribution exists for this null model, and full enumeration for an entire dataset is impractical. Therefore, *p*-values for this new null model are not available.

Acknowledgements

We thank Dr. Bosco Ho for insightful comments. We thank Xiang Li, Drs. William Wimley and Jinfeng Zhang for helpful discussions. This work is supported by grants from the National Science Foundation (CAREER DBI0133856), National Institute of Health (GM68958), and Office of Naval Research (N000140310329).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.07.095

References

1. Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.* **7**, 1029–1038.
2. Arkin, I. & Brunger, A. (1998). Statistical analysis of predicted transmembrane alpha-helices. *Biochim. Biophys. Acta.* **1429**, 113–128.
3. Adamian, L. & Liang, J. (2001). Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* **311**, 891–907.
4. Adamian, L. & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: Polar clamps and serine zippers. *Proteins*, **47**, 209–218.
5. Bowie, J. (1997). Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780–789.
6. Schulz, G. E. (2002). The structure of bacterial outer membrane proteins. *Biochim. Biophys. Acta*, **1565**, 308–317.
7. Jackups, R. & Liang, J. (2005). Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. *J. Mol. Biol.* **354**, 979–993.
8. Wimley, W. C. (2002). Toward genomic identification of β-barrel membrane proteins: composition and architecture of known structures. *Protein Sci.* **11**, 301–312.
9. Gromiha, M. M., Ahmad, S. & Suwa, M. (2004). Neural network-based prediction of transmembrane β-strand segments in outer membrane proteins. *J. Comput. Chem.* **25**, 762–767.

10. Bigelow, H. R., Petrey, D. S., Liu, J., Przybylski, D. & Rost, B. (2004). Predicting transmembrane β-barrels in proteomes. *Nucleic Acids Res.* **32**, 2566–2577.
11. Seshadri, K., Garemyr, R., Wallin, E., von Heijne, G. & Elofsson, A. (1998). Architecture of β-barrel membrane proteins: analysis of trimeric porins. *Protein Sci.* **7**, 2026–2032.
12. Ulmschneider, M. B. & Sansom, M. S. (2001). Amino acid distributions in integral membrane protein structures. *Biochim. Biophys. Acta*, **1512**, 1–14.
13. Chamberlain, A. K. & Bowie, J. U. (2004). Asymmetric amino acid compositions of transmembrane β-strands. *Protein Sci.* **13**, 2270–2274.
14. Chamberlain, A. K., Lee, Y., Kim, S. & Bowie, J. U. (2004). Snorkeling preferences foster an amino acid composition bias in transmembrane helices. *J. Mol. Biol.* **339**, 471–479.
15. Chamberlain, A. K. & Bowie, J. U. (2004). Analysis of side-chain rotamers in transmembrane proteins. *Biophys. J.* **87**, 3460–3469.
16. Senes, A., Gerstein, M. & Engelman, D. M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with β-branched residues at neighboring positions. *J. Mol. Biol.* **296**, 921–936.
17. Senes, A., Engel, D. E. & DeGrado, W. F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. *Curr. Opin. Struct. Biol.* **14**, 465–479.
18. Bitto, E. & McKay, D. B. (2003). The periplasmic molecular chaperone protein SurA binds a peptide motif that is characteristic of integral outer membrane proteins. *J. Biol. Chem.* **278**, 49316–49322.
19. Hart, R., Royyuru, A., Stolovitzky, G. & Califano, A. (2000). Systematic and fully automated identification of protein sequence patterns. *J. Comput. Biol.* **7**, 585–600.
20. Wouters, M. A. & Curmi, P. M. (1995). An analysis of side chain interactions and pair correlations within antiparallel β-sheets: the differences between backbone hydrogenbonded and non-hydrogen-bonded residue pairs. *Proteins*, **22**, 119–131.
21. Jackups, R. & Liang, J. (2006). Combinatorial Model for Sequence and Spatial Motif Discovery in Short Sequence Fragments: Examples of Beta-Barrel Membrane Proteins. *IEEE EMBS*. In press.
22. Struyvé, M., Moons, M. & Tommassen, J. (1991). Carboxy-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein. *J. Mol. Biol.* **218**, 141–148.
23. Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze **8**, 3–62.
24. Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* **57**, 289–300.
25. Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad. Sci. USA*, **98**, 5116–5121.
26. White, S. H. & Wimley, W. C. (1999). Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 319–365.
27. Curran, A. R. & Engelman, D. M. (2003). Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struct. Biol.* **13**, 412–417.
28. Dawson, J. P., Weinger, J. S. & Engelman, D. M. (2002). Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.* **316**, 799–805.
29. Ng, P. C., Henikoff, J. G. & Henikoff, S. (2000). PHAT: a transmembrane-specific substitution matrix. *Bioinformatics*, **9**, 760–766.
30. Pautsch, A. & Schulz, G. E. (1998). Structure of the outer membrane protein A transmembrane domain. *Nat. Struct. Biol.* **5**, 1013–1017.
31. Vogt, J. & Schulz, G. E. (1999). The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. *Structure Fold Des.* **7**, 1301–1309.
32. Vandeputte-Rutten, L., Bos, M. P., Tommassen, J. & Gros, P. (2003). Crystal structure of Neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential. *J. Biol. Chem.* **278**, 24825–24830.
33. Prince, S. M., Achtman, M. & Derrick, J. P. (2002). Crystal structure of the OpcA integral membrane adhesin from *Neisseria meningitidis*. *Proc Natl Acad. Sci. USA*, **99**, 3417–3421.
34. Vandeputte-Rutten, L., Kramer, R. A., Kroon, J., Dekker, N., Egmond, M. R. & Gros, P. (2001). Crystal structure of the outer membrane protease OmpT from *Escherichia coli* suggests a novel catalytic site. *EMBO J.* **20**, 5033–5039.
35. Snijder, H. J., Ubarretxena-Belandia, I., Blaauw, M., Kalk, K. H., Verheij, H. M., Egmond, M. R. *et al.* (1999). Structural evidence for dimerization-regulated activation of an integral membrane phospholipase. *Nature*, **401**, 717–721.
36. Oomen, C. J., Van Ulsen, P., Van Gelder, P., Feijen, M., Tommassen, J. & Gros, P. (2004). Structure of the translocator domain of a bacterial autotransporter. *EMBO J.* **23**, 1257–1266.
37. Weiss, M. S. & Schulz, G. E. (1992). Structure of porin refined at 1.8 A resolution. *J. Mol. Biol.* **227**, 493–509.
38. Kreusch, A. & 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.
39. Cowan, S., Garavito, R. M., Jansonius, J. N., Jenkins, J. A., Karlsson, R., Konig, N. *et al.* (1995). The structure of OmpF porin in a tetragonal crystal form. *Structure*, **3**, 1041–1050.
40. Zeth, K., Diederichs, K., Welte, W. & Engelhardt, H. (2000). Crystal structure of Omp32, the anion-selective porin from *Comamonas acidovorans*, in complex with a periplasmic peptide at 2.1 A resolution. *Structure Fold Des.* **8**, 981–992.
41. Meyer, J. E., Hofnung, M. & Schulz, G. E. (1997). Structure of maltoporin from *Salmonella typhimurium* ligated with a nitrophenyl-maltotrioside. *J. Mol. Biol.* **266**, 761–775.
42. Forst, D., Welte, W., Wacker, T. & Diederichs, K. (1998). Structure of the sucrose-specific porin ScrY from *Salmonella typhimurium* and its complex with sucrose. *Nat. Struct. Biol.* **5**, 37–46.
43. Buchanan, S. K., Smith, B. S., Venkatramani, L., Xia, D., Esser, L., Palnitkar, M. *et al.* (1999). Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. *Nat. Struct. Biol.* **6**, 56–63.
44. 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.

45. Ferguson, A. D., Chakraborty, R., Smith, B. S., Esser, L., Van Der Helm, D. & Deisenhofer, J. (2002). Structural basis of gating by the outer membrane transporter FecA. *Science*, **295**, 1715–1719.
46. Chimento, D. P., Mohanty, A. K., Kadner, R. J. & Wiener, M. C. (2003). Substrate-induced transmembrane signaling in the cobalamin transporter BtuB. *Nat. Struct. Biol.* **10**, 394–401.
47. Koronakis, V., Sharff, A. J., Koronakis, E., Luisi, B. & Hughes, C. (2000). Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature*, **405**, 914–919.
48. Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H. & Gouaux, J. E. (1996). Structure of staphylococcal β-hemolysin, a heptameric transmembrane pore. *Science*, **274**, 1859–1866.

Edited by J. Bowie

(Received 9 May 2006; received in revised form 29 July 2006; accepted 31 July 2006)
Available online 15 August 2006