Supporting Information

S1. Identifying Contacting Strands

Contacts between neighboring strands are identified using a Voronoi diagram criteria [1, 2]. Briefly, we construct the Delaunay triangulation of the atom centers using a probe radius of 0.5 Å (see ref [3] for justification of this choice). If two atoms are in physical contact, by the criterion that their Voronoi cells intersect and that the intersection is partially contained within the body of the molecule, they will be connected by a dual Delaunay edge between two atoms from two neighboring strands. The contacting atoms are then identified from the Delaunay triangulation of the atomic centers, in which a subset of Delaunay edges are selected as the set of dual edges [2].

We declare that two β-strands in the transmembrane domain are in contact if ≥ 5 residues participate in strand interaction. If only 3 or 4 residues participate in strand interaction, we label these strands as being in contact if they have ≥ 30 atomic contacts across the strand interface. We identify the protein-protein interface as the largest stretch of consecutive contacting strands.

S2. Configuration Space

We assume a canonical model for the transmembrane strands in which each strand interacts with two neighboring strands [5–7]. In this model, the strands all have the same length $L + 1 = 16$. A strand in the canonical position with no offset is centrally located, with the first residue at the beginning of the strand's periplasmic side labeled as position 0.

A strand can deviate from this canonical position. The deviation is described by an offset value, that is, a displacement integer d, which represents the offset distance between the first residue in the strand and that in a canonical strand. We have $d \in \{-L, \cdots, 0, \ldots, L\}$. The conformation of the transmembrane region of a β-barrel protein with N strands will have N strand-pair interactions, and can be parameterized by the vector $\boldsymbol{d}$ of offset integers:

$$\boldsymbol{d} = (d_1, \cdots, d_N),$$

in which $d_i \in [-L, \cdots, 0, \cdots, L]$ denotes the offset position of strand i.

The full configuration space Ω for the transmembrane strands of this simplified model is:

$$\Omega = \{\boldsymbol{d} | \boldsymbol{d} = (d_1, \cdots, d_N) \in \mathbb{Z}^N\}.$$

Because each strand can have a displacement offset value from $-L$ to L, the size of the configuration space is $|\Omega| = (2L+1)^N$. In this study, we use a model with reduced configuration space, in which $d_i \in [-l, \cdots, 0, \cdots, l]$, with $l = 3$.

We use $\boldsymbol{d}_i$ to denote the possible configuration for the strand triplet of $(i-1, i, i+1)$. We allow strands $i-1$ and $i+1$ to take any of the positions $-l, \cdots, 0, \cdots, l$. When calculating

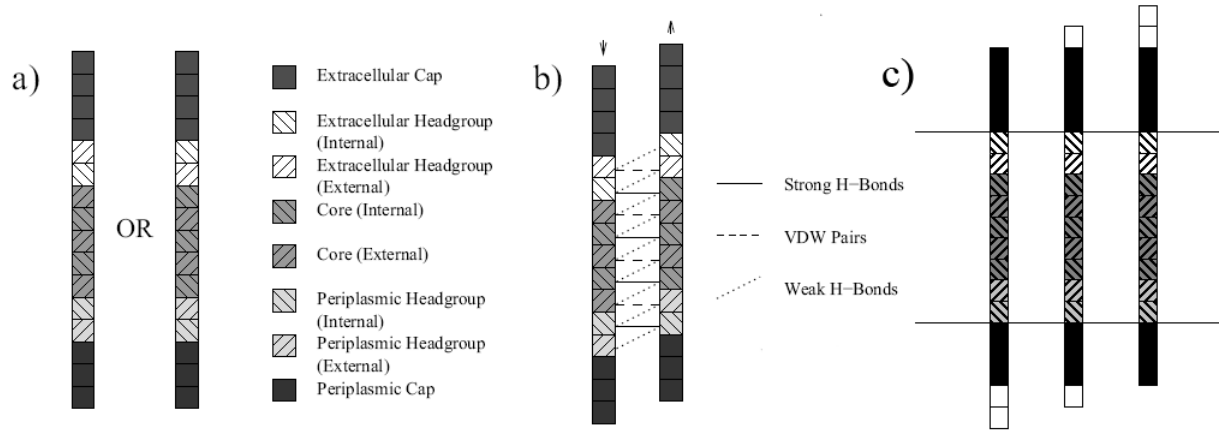

Fig. S1: The configuration space of transmembrane strands. a) The canonical position of a transmembrane strand. A strand contains 16 residues, and can have two orientations such that a position can face either the internal barrel space, or the external lipid space. In addition, the positions of the 16 residues are located in different regions: the periplasmic cap region, the periplasmic headgroup region, the core region, the extracellular headgroup region, and the extracellular cap region. The 1-body contribution of each of the 16 residue to the strand energy depends on the region in which it is located and its orientation. b) The 9 central residue pairs located on adjacent strands interact and contribute to the strand energy through 2-body terms. Contributions from a pair of interacting residues depend on the type of residues and the type of interaction (side chain, strong-H bond, and/or weak H-bond, Figure from [4]). c) The middle strand represents a canonical strand with an offset of 0. The left strand has an offset of -1 and the right strand has an offset of 1.

the strand energy for finding the protein-protein interface strands, the central strand of the triplet has a fixed position. When calculating the melting temperature, the central strand i is allowed to take any of the positions $-l, \cdots, 0, \cdots, l$.

S3. Energy Model of Strands and Interface Identification

In our model of β-strands, we assume that the energy associated with each transmembrane strand in a particular configuration has two components. First, each residue in strand i contributes to the single body strand energy $E_1(d_i)$ depending on which of the 8 regions it is located as well as the orientation of the strand (Fig S1a). Second, a strand interacts with two neighboring strands through strong backbone H-bond interaction (E_H), side-chain interactions (E_{SC}), and weak H-bond interactions (E_{WH}) [5–7] (see [7] for details). The strand energy $E(i, \boldsymbol{d}_i)$ for the strand i associated with a particular strand configuration $\boldsymbol{d}_i$

is calculated as:

$$
\begin{aligned}
E(i,\boldsymbol{d}_i) =& E_i(d_{i-1},d_i,d_{i+1}) = \sum_{k_i} E_1(k_i;d_i) + \sum_{k_{i-1}} E_1(k_{i-1};d_{i-1}) + \sum_{k_{i+1}} E_1(k_{i+1};d_{i+1}) + \\
& \alpha[\sum_{k_i}\sum_{k_{i-1}} E_H(k_i,k_{i-1};d_i,d_{i-1}) + \sum_{k_i}\sum_{k_{i+1}} E_H(k_i,k_{i+1};d_i,d_{i+1})] + \\
& \beta[\sum_{k_i}\sum_{k_{i-1}} E_{SC}(k_i,k_{i-1};d_i,d_{i-1}) + \sum_{k_i}\sum_{k_{i+1}} E_{SC}(k_i,k_{i+1};d_i,d_{i+1})] + \\
& \gamma[\sum_{k_i}\sum_{k_{i-1}} E_{WH}(k_i,k_{i-1};d_i,d_{i-1}) + \sum_{k_i}\sum_{k_{i+1}} E_{WH}(k_i,k_{i+1};d_i,d_{i+1})],
\end{aligned}
\tag{1}
$$

where k_i is the position of a residue in strand i, and the summations are over all residues in the strands, namely, $k_i, k_{i-1}, k_{i+1} \in \{0, \cdots, L\}$. The values for α, β and γ are empirically set to 19/26, 15/26 and 10/26. Our results are not sensitive to the specifics of these values.

For the purpose of identifying strands located in protein-protein interface, we define $E(i)$ as the average of strand i and the two neighboring strands:

$$E(i,\boldsymbol{d}_i) = (E(i-1,\boldsymbol{d}_{i-1}) + E(i,\boldsymbol{d}_i) + E(i+1,\boldsymbol{d}_{i+1}))/3. \tag{2}$$

where $\boldsymbol{d}_{i-1}, \boldsymbol{d}_i$ and $\boldsymbol{d}_{i+1}$ are the offsets of the three strands. This is to discount any spurious interfacial strand due to isolated stable low energy strands within the interface. These isolated strands are relatively stable compared to the rest of the interface strands although they are located within a larger continuous interface that is unstable. This model allows us to detect larger interfaces and avoid artifacts from which interfaces decompose into smaller disconnected surface patches.

The approach of deriving energy parameters as described in reference [7] is general. It is possible to generate potential functions that exclude either all oligomers, or all monomers. Although the reduced number of structures limits the accuracy of the estimated potential function, significant differences in these functions are found. For example, the estimated E_1 for interface strands favors charged residues (K, propensity 1.67; R, 1.58; D, 1.73; E, 1.73) and M (1.87), whereas the E_1 for non-interface strands does not favor these residues (K, 0.89; R, 0.90; D, 0.88; E, 0.88; and M, 0.85). These differences suggest that the findings presented in this study are general, and are not consequences of chance occurrences of the particular chosen protein systems described in the main text. Another model is to develop potential function specifically for the inside half and for the outside half of the barrel, respectively, with the goal of discriminating different non-barrel elements such as in-plugs and out-clamps. However, this would require substantial amount of structural data to discriminate the in-plugs from the out-clamps that currently does not exist. We expect that the effectiveness of the potential function depends on the choice and availability of the structural data used for

parameter derivation, which needs to be consistent with the physical model and the question to be addressed.

S4. Results on Strand Energy Evaluations

Fig S2 shows the energy difference between a strand and the protein-specific strand average evaluated using Eqn 2 for all 116 strands in the 7 oligomeric proteins. In general, we find that strands located in protein-protein interfaces have higher than average energy.

Fig S3 shows the overall distributions of energy differences between strands and the protein-specific strand averages for all strands with higher than protein-specific averages in (a) monomers and in (b) oligomers. Overall, oligomers have larger deviations and variances.

In OmpF, strand 12 has energy similar to that of strand 2 (Fig 1 in the main text). However, it is not predicted to be located in the protein-protein interface, as the protein-protein interaction interface is defined as the largest set of consecutive strands with $E_i \geq \mathbb{E}(E)$. The higher energy of strand 12 may be due to steric factors, similar to strands F and G in PagP [8].

S5. Energy Model for Melting Temperature T_m Calculations

When computing the total energy of a specific configuration for melting point calculations, we use the following formula to avoid double counting of strand-pair interactions. For the i-th strand, we have the interaction energy as:

$$\begin{aligned} E(i, \boldsymbol{d}_i) =& E_i(d_{i-1}, d_i, d_{i+1}) = \sum_{k_i} E_1(k_i; d_i) + \\ & \frac{\alpha}{2}[\sum_{k_i}\sum_{k_{i-1}} E_H(k_i, k_{i-1}; d_i, d_{i-1}) + \sum_{k_i}\sum_{k_{i+1}} E_H(k_i, k_{i+1}; d_i, d_{i+1})] + \\ & \frac{\beta}{2}[\sum_{k_i}\sum_{k_{i-1}} E_{SC}(k_i, k_{i-1}; d_i, d_{i-1}) + \sum_{k_i}\sum_{k_{i+1}} E_{SC}(k_i, k_{i+1}; d_i, d_{i+1})] + \\ & \frac{\gamma}{2}[\sum_{k_i}\sum_{k_{i-1}} E_{WH}(k_i, k_{i-1}; d_i, d_{i-1}) + \sum_{k_i}\sum_{k_{i+1}} E_{WH}(k_i, k_{i+1}; d_i, d_{i+1})], \end{aligned} \tag{3}$$

where the summations are over all residues in the strand, namely, $k_i, k_{i-1}, k_{i+1} \in \{1, \cdots, L\}$.

The overall energy for a specific conformation $\boldsymbol{d} = (d_1, \cdots, d_N)$ of a protein with N transmembrane strands is:

$$E(\boldsymbol{d}) = \sum_{i=1}^{N} E(i, \boldsymbol{d}_i)$$

The values for α, β, and γ are set to the same values as before.

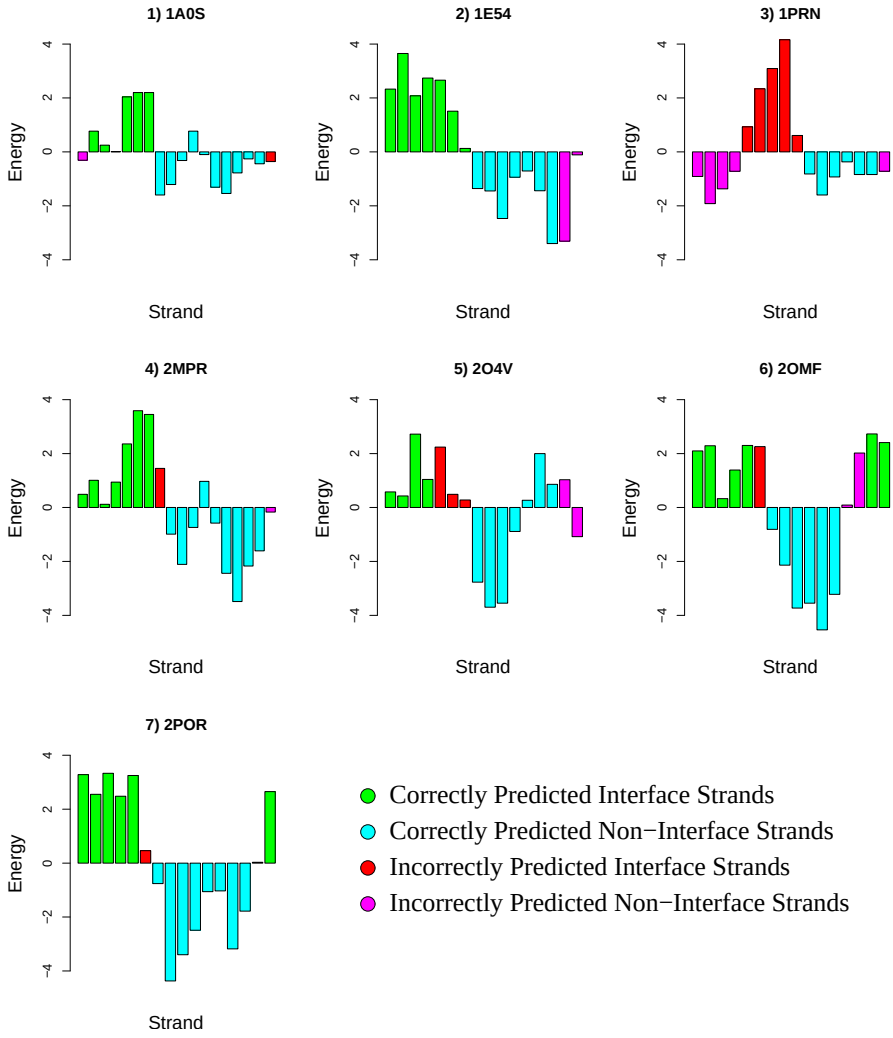

Fig. S2: The energy differences between individual transmembrane strand and protein-specific average for 7 oligomeric β-barrel membrane proteins. Overall, strands with above average energy are dominated by those that are located in interface, and stable strands with lower than average energy are mostly non-interfacial strands. The largest contiguous region containing high energy strands as described in the main text can identify the interface of protein-protein interactions with improved accuracy. The colors are encoded to reflect correct/incorrect predictions.

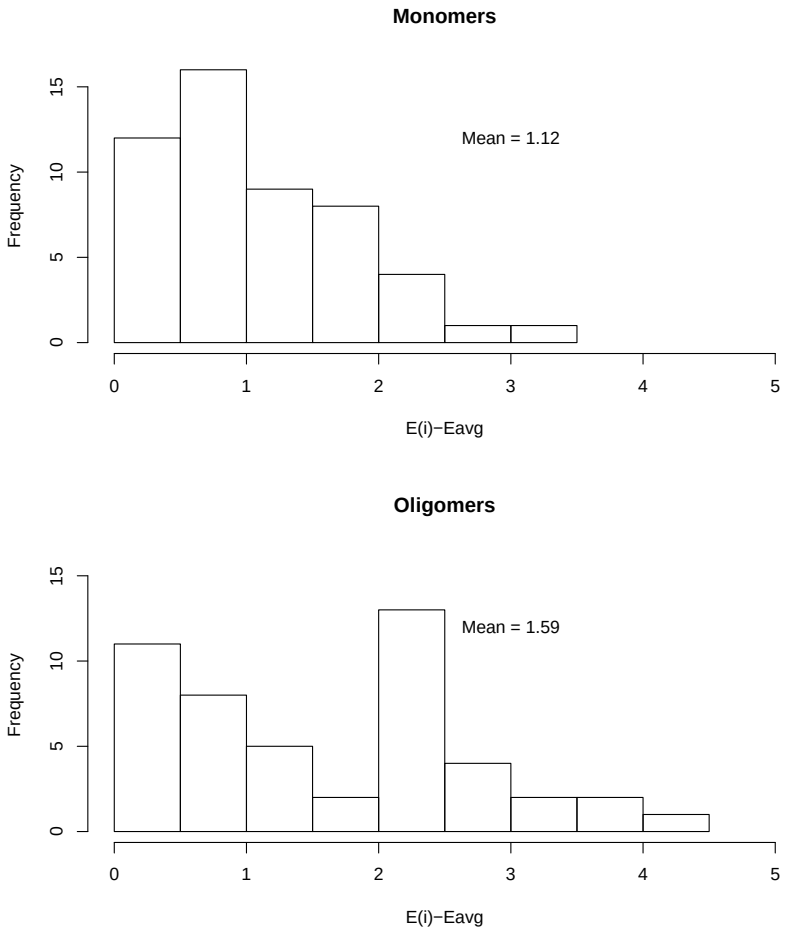

Fig. S3: The distributions of energy differences between individual transmembrane strand and protein-specific average for (a) 16 monomeric and (b) 7 oligomeric β-barrel membrane proteins. Overall, high energy strands in oligomers have larger deviations.

S6. Enumerating Conformation and Computing Partition Function

When the number of strands is small (≤ 14), we can enumerate all states in the configuration space and exactly compute the thermodynamic properties in this reduced conformational space.

When the number of strands is large, we divide the set of strands into four parts. Part 1 consists of strand 1 alone, part 2 includes strands from $i = 2$ to $i = \lfloor \frac{n}{2} \rfloor - 1$, part 3 consists of strand $i = \lfloor \frac{n}{2} \rfloor$, and part 4 includes strands from $i = \lfloor \frac{n}{2} \rfloor + 1$, to n. We can enumerate all configurations of parts 2 and 4 exactly, with the exception that the constants α, β and γ are not divided by 2 for the terminal strands, *i.e.*, strand 2 and strand $i = \lfloor \frac{n}{2} \rfloor - 1$ for part 2, strand $i = \lfloor \frac{n}{2} \rfloor + 1$, and strand $i = n$ for part 4. To avoid double counting, we only include 1-body interaction terms for parts 1 and 3.

This approximation gives accurate results, as errors are bounded by that of only 6 residue-pair interactions per configuration. If the first residue in the extracellular cap region faces inside the barrel, we miss 4 weak hydrogen bond and 2 side chain interactions. Otherwise, we miss 4 weak hydrogen bond and 2 strong hydrogen bond interactions. This error is small as there are 256 single body interactions and 232 pairwise interactions in 16 strand β-barrel of which we miss 1.2% of the interactions. This error reduces further as the number of strands in the beta barrel increase. For a 22 strand beta-barrel, we have 671 interactions (352 single body and 319 pairwise), and only about 0.9% of the residue-pair interactions are not accounted for.

S7. Melting Temperature

We can compute the relative melting temperature T_m of β-barrel strands as:

$$T_m = \arg\max_T C = \arg\max_T \left[\{ \mathbb{E}(E^2) - \mathbb{E}(E)^2 \} / T^2 \right]$$

Here C is the calculated heat capacity, and the expected values of the energy $\mathbb{E}(E)$ and the second moment of energy $\mathbb{E}(E^2)$ are calculated as:

$$\mathbb{E}(E) = \frac{\sum_{\boldsymbol{d}} E(\boldsymbol{d}) e^{-E(\boldsymbol{d})/T}}{\sum_{\boldsymbol{d}} e^{-E(\boldsymbol{d})/T}},$$

and

$$\mathbb{E}(E^2) = \frac{\sum_{\boldsymbol{d}} E(\boldsymbol{d})^2 e^{-E(\boldsymbol{d})/T}}{\sum_{\boldsymbol{d}} e^{-E(\boldsymbol{d})/T}}.$$

T_m is identified by locating the temperature at which C reaches maximum value.

Table S1: **Data Set.The Melting Temperature and expected energy value of transmembrane strands of 25 non-homologous β-barrel membrane proteins.**

Proteins/pdb	Organism	Strands	Structure	Melting Temperature	Energy	Prediction1
TolC/1ek9	*E. coli*	12	M[1]	3.23	-0.74	T
OmpG/2f1c	*E. coli*	12	M	2.91	-1.19	T
NspA/1p4t	*N. meningitidis*	8	M	2.74	0.01	T
OmpA/1bxw	*E. coli*	8	M	2.66	-1.59	T
OpcA/1k24	*N. meningitidis*	10	M	2.66	-0.10	T
OmpX/1qj8	*E. coli*	8	M	2.62	0.00	T
OmpLA/1qd6	*E. coli*	12	M	2.62	0.03	T
FadL/1t16	*E. coli*	14	M/I	2.36	-1.37	T
OmpW/2fit	*E. coli*	8	M	2.15	-1.59	T
OmpT/1i78	*E. coli*	10	M	1.91	0.48	T
Porin/1prn	*R. balistica*	16	T/I	1.74	-0.41	T
NalP/1uyn	*N. meningitidis*	10	M/I	1.72	-0.48	T
Omp32/1e54	*C. acidovorans*	16	T/I	1.57	0.24	T
PagP/1thq	*E. coli*	8	M/O	1.55	0.84	T
Porin P/2o4v	*P. aeruginosa*	16	T/I	1.43	0.44	T
LamB/2mpr	*S. typhimurium*	18	T/I	1.39	1.16	T
FecA/1kmo	*E. coli*	22	M/I	1.25	1.10	T
BtuB/1nqe	*E. coli*	22	M/I	1.25	-0.64	T
Scr Y/1a0s	*S. typhimurium*	18	T/I	1.13	0.26	T
FepA/1fep	*E. coli*	22	M/I	1.06	0.65	T
α-hemolysin/7AHL	*S. aureus*	14	M/O[2]	1.03	0.16	T
OmpF/2omf	*E. coli*	16	T/I	1.01	0.10	T
FptA/1xkw	*P. aeruginosa*	22	M/I	1.01	-0.88	T
FhuA/2fcp	*E. coli*	22	M/I	0.85	1.97	F
Porin/2por	*R. capsulatus*	16	T/I	0.06	0.10	T

S8. Melting Temperature, Heat Capacity Calculation Results

The set of 25 β-barrel membrane protein structure are listed in Table S1, along with their PDB codes, their associated organisms, the number of transmembrane strands, and based on the known structure, whether each protein is a monomer (M), a monomer with an in-plug (M/I), a monomer with an out-clamp (M/O), or an oligomer with an in-plug (T/I). The computed melting temperature T_m and the expected energy value for the transmembrane strands are also listed in this Table. Whether our prediction of the oligomerization state is correct is indicated in the last column (T for true prediction, and F for false prediction).

Figure S4 shows the computed heat capacity at different temperatures for the 25 proteins. The temperature at which the heat capacity reaches maximum is the melting temperature.

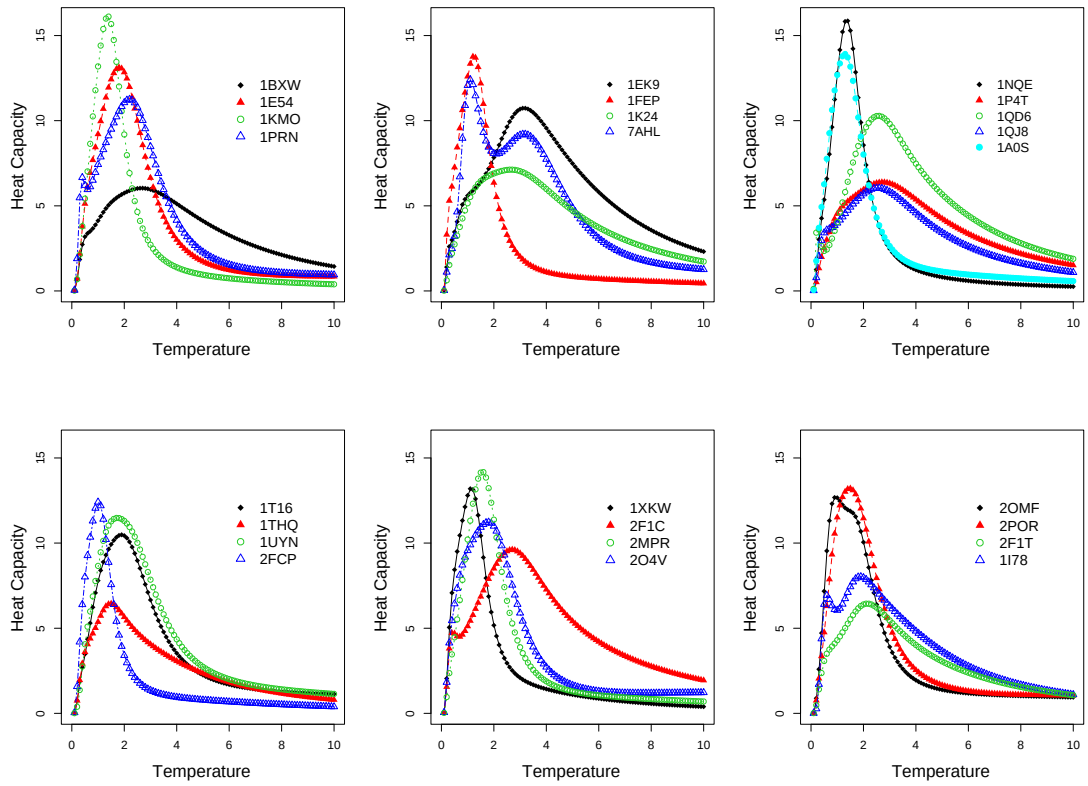

Fig. S4: Calculated heat capacity of 25 non-homologous proteins at different temperature. These values are for the strands in the transmembrane domain and are obtained based on accurate calculation of the partition function of the barrel, assuming the simplified conformational state model. The temperature at which the heat capacity reaches its maximum corresponds to the melting temperature T_m.

S9. Prediction Performance Evaluation

We calculate accuracy, sensitivity, and specificity to evaluate the performance of our prediction of protein-protein interaction interface strands. We denote the numbers of true strands inside and outside the interface by the Voronoi criterion as P and N, respectively, and the number of correctly predicted interfacial and all other strands as TP and TN, respectively. We have accuracy $= (TP + TN)/(P + N)$, sensitivity $= TP/P$, and specificity $= TN/N$.

References

[1] Liang J, Dill K A (2001) Are proteins well-packed? *Biophys J* 81(2):751–766.

[2] Li X, Hu C, Liang J (2003) Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins* 53(4):792–805.

[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] Ronald Jakcups J (2006) *Prediction of transmembrane β-barrel protein structures* Ph.D. thesis University of Illinois at Chicago.

[5] Wouters M A, Curmi P M (1995) An analysis of side chain interactions and pair correlations within antiparallel β-sheets: the differences between backbone hydrogen-bonded and non-hydrogen-bonded residue pairs *Proteins* 22:119–131.

[6] Ho B, Curmi P (2002) Twist and shear in β-sheets and β-ribbons. *J Mol Biol* 317(2):291–308.

[7] Jackups Jr 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(4):979–93.

[8] Evanics F, Hwang P, Cheng Y, Kay L, Prosser R (2006) Topology of an outer-membrane enzyme: Measuring oxygen and water contacts in solution NMR studies of PagP. *J Am Chem Soc* 128(25):8256–64.