Biophysical Journal, Volume 99

Supporting Material

Title: Statistical Mechanics of Integral Membrane Protein Assembly.

Authors: Karim Wahba, David Schwab, and Robijn Bruinsma

Supplementary Material

1) Supplementary Figures and Table

Figure S1.

Comparison of segment placements for bR with k the residue number. Row A: Measured segment locations by Jayasinghe et al. (*Protein Sci.* 10:455-458 (2001)); Row B: groundstate computed from the model for variable segment lengths for $\mu = 0.7$ kcal/mol and $L_\alpha = 21 - 26$; Row C: placement by sequential adsorption; Row D: linear sequential placement.

Figure S2.

Plot of the probability $\sigma(k)$ that residue *k* is part of a TM segment at room temperature for $\mu = 0.7$ kcal/mol. (A) Wildtype bR sequence. The sixth segment is subject to location and length fluctuations. There are no number fluctuations. (B) Shuffled bR sequence. In addition to strong location/size fluctuations there are also segment number fluctuations.

Figure S3.

Mean number of transmembrane segments ρ_{TM} for a randomly shuffled bR sequence as a function of μ. (A) k_BT = 0.01 kcal/mole. The plot of ρ_{SA} for sequential adsorption (dashed line) starts to deviate from ρ_{TM} at the arrow. (B) Room temperature. Structures with more than one segment suffer from strong number fluctuations (up to the 9-TM segment structure). The corresponding occupancy plot at the site of the arrow is shown in Figure 3 of the main text.

Figure S4.

Mean number of TM segments ρ_{TM} as a function of μ. (A) Diacylglycerol kinase , a 3-TM segment protein. (B) Cytochrome C oxidase, a 12-TM segment protein. Dashed lines show the mean number of TM segments ρ_{SA} computed by sequential adsorption.

Table Legend

Table 1.

The second column gives the number of wildtype TM segments of the eleven IMPs shown in column one with their corresponding PDB id. The third column gives the size of the stability interval $\Delta\mu$ for $P = P_w$ of the groundstate structure as computed from the model. The fourth column gives the average (over 100 runs) number of random single point mutations (SPMs) normalized by sequence length required to change the segment from its wildtype value for each protein, with the standard error in the fifth column. Note that 1PW4 has a high thermodynamic stability but a low mutation threshold. The sixth column gives the average (over 100 runs) mutation thresholds for each sequence after random shuffling (100 realizations) with the standard error in the seventh column.

Figures S1

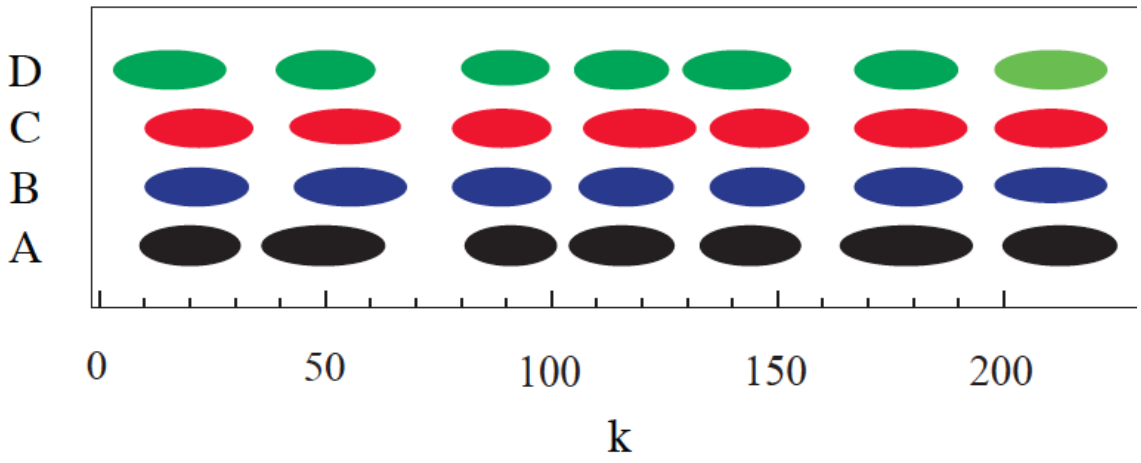
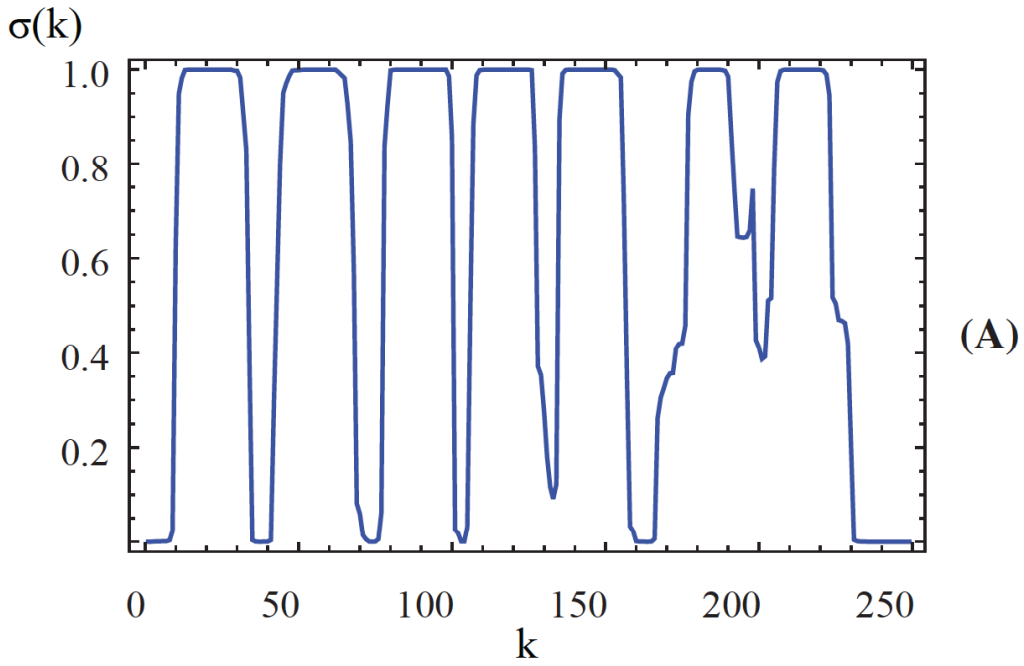
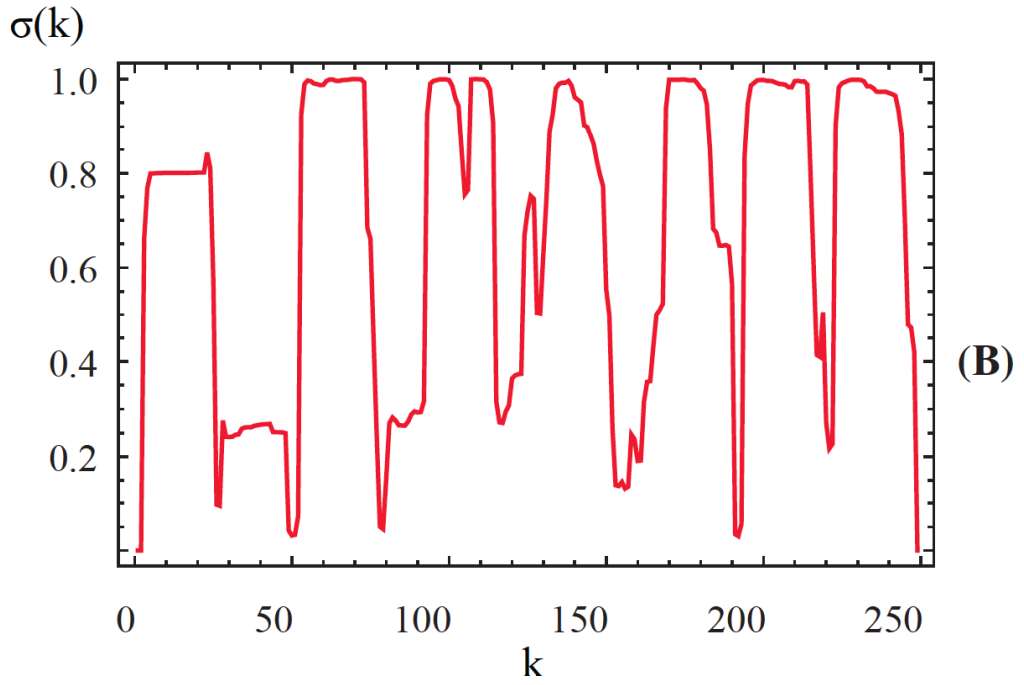


Figure S2

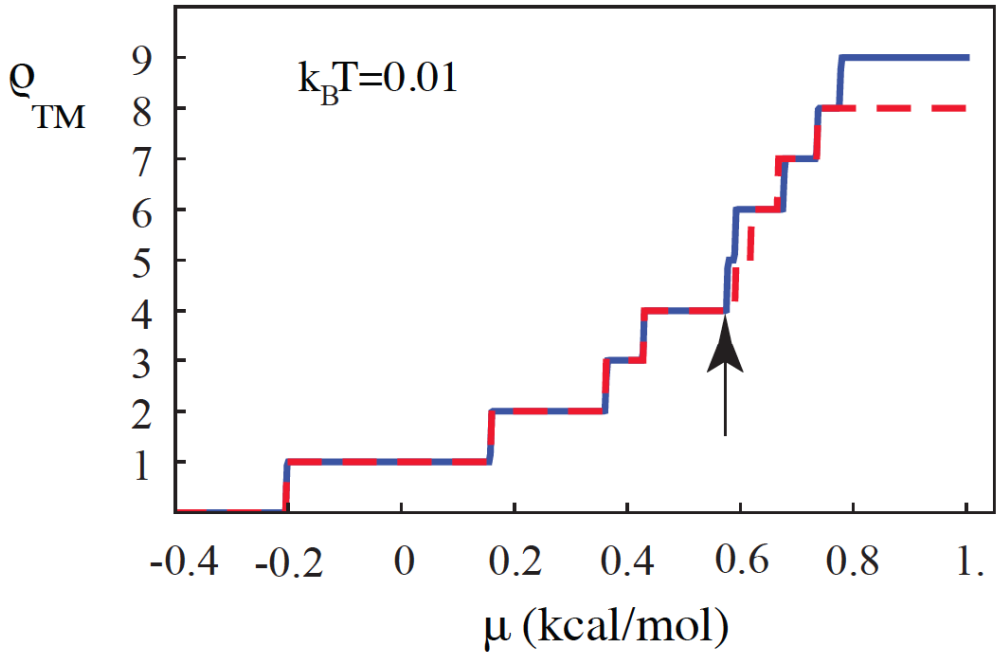

σ(k)
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
k

(A)

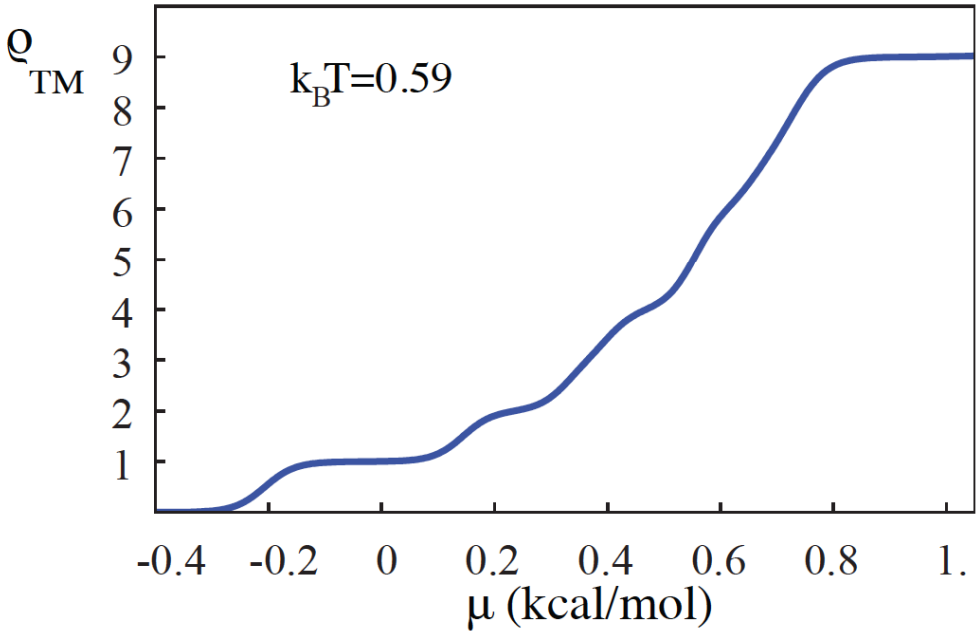

σ(k)
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
k

(B)

Figures S3

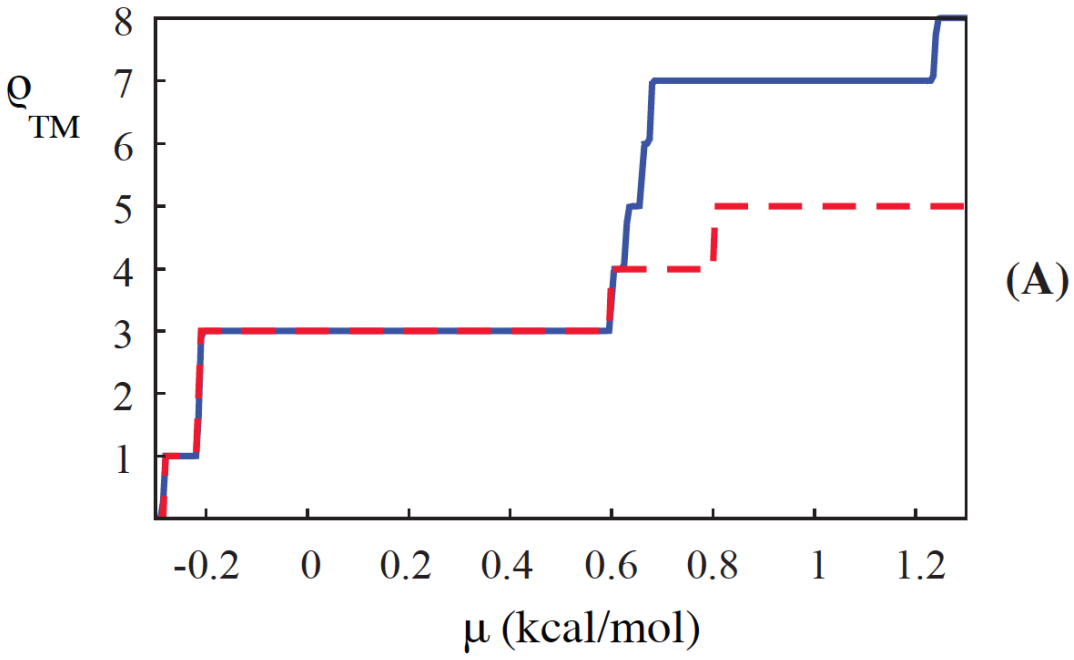


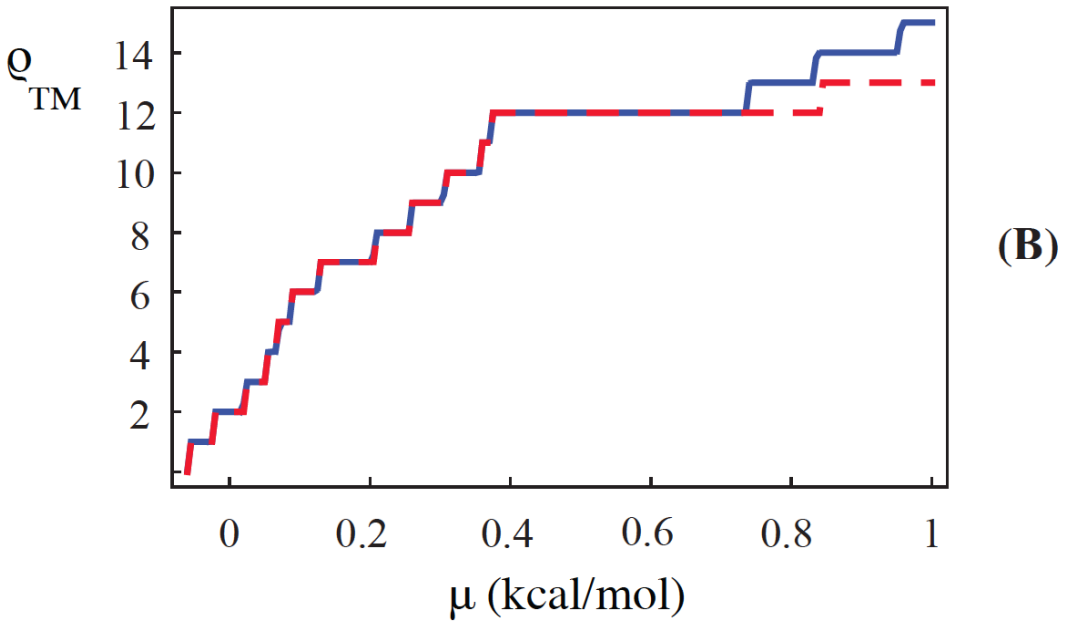
(A)



(B)

Figure S4


Q
TM
8
7
6
5
4
3
2
1
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
μ (kcal/mol)
(A)


Q
TM
14
12
10
8
6
4
2
0
0.2
0.4
0.6
0.8
1
μ (kcal/mol)
(B)

Protein (PDB #)	Known # of TM helices	Relative gap size $r = \frac{\Delta\mu}{\mu_{N-1} + \Delta\mu}$	Average # of random point mutations (wild type sequence)	Standard deviation (wild type sequence)	Average # of random point mutations (shuffled sequence)	Standard deviation (shuffled sequence)
Bacteriorhodopsin (1BRD)	7	0.53	25.31	8.71	6.09	3.08
Sensory rhodopsin – *Anabaena* (1XIO)	7	0.60	16.47	5.95	3.7	1.87
Halorhodopsin (1E12)	7	0.51	13.38	6.85	4.78	2.45
Bovine rhodopsin (1F88)	7	0.55	32.55	11.21	8.87	4.14
Sensory rhodopsin II (1H68)	7	0.71	24.78	7.02	10.8	5.03
Bovine Cyto. C. Oxidase – subunit III (1OCC)	7	0.51	31.65	11.37	9.49	4.42
Bovine Cyto. C. Oxidase – subunit I (1OCC)	12	0.47	35.44	13.59	3.36	1.73
Glycerol -3 Phosphate Transporter – *E. coli* (1PW4)	12	0.50	25.01	12.69	3.50	1.76
(*P. denitrificans*) Cyto. C. Oxidase – subunit I (1QLE)	12	0.34	13.11	6.49	2.73	1.25
Lactose Permease – *E. coli* (1PV6)	12	0.32	8.35	4.36	4.19	2.16
Multi drug resistance protein EmrD – *E. coli* (2GFP)	12	0.36	8.58	4.34	4.08	2.12

Table 1

2) Recursion Relation Method

Single Species

We first construct the recursion relations for the simple case where all TM segments have the same length. We will work in the grand canonical ensemble, since the number of the TM segments is allowed to fluctuate A one-dimensional lattice of L sites determines the possible segment start locations, where L is the total length of the protein minus the length of one TM helix. Define $Z_F(l)$ to be the (forward) partition function of the first l sites subject to the restriction that one TM segment begins at site 1 and another at site l. An on-site potential energy *u(l)* is included obtained by summing the hydropathy values of the amino acids in the TM segment starting at l, as in equation [1] (for simplicity, we absorbed here the chemical potential into the potential energy). Define the nearest-neighbor interaction energy as $-\beta^{-1}\log N(x)$, where x is the distance between the starting sites of the two neighboring segments. The interaction can, in principle, be arbitrary but we focus on the hard-core repulsion given by [2] of the main text. Note that *N(x)* depends on the fixed length of the segment. Moving the fixed segment starting at l by one step to the right produces a recursion relation for $Z_F(l)$:

$$Z_F(l+1) = e^{-\beta u(l+1)} \sum_{j=1}^{l} Z_F(j) N(l+1-j).$$

This equation must be supplemented with the boundary condition $Z_F(1) = e^{-\beta u(1)}$. One can construct a similar recursion relation for the backwards partition function, $Z_B(L-l)$, i.e. the partition function with TM segments fixed to begin at sites L and $L-l$:

$$Z_B(L-l) = e^{-\beta u(L-l)} \sum_{j=L-l+1}^{L} Z_B(j) N(L-l-j)$$

along with the boundary condition $Z_B(L) = e^{-\beta u(L)}$. The probability for a TM to start at site i is then given by:

$$\rho(i) = \frac{e^{\beta u(i)} Z_F(i) Z_B(i)}{Z_B(1)}.$$

The extra Boltzmann factor avoids double-counting of the factors in the forward and backward recursion relations. Here, $Z_B(1)$ is the entire partition function (alternatively given by $Z_F(L)$).

Multiple Species

We now generalize the recursion relations to allow for *s* different segment lengths. Each segment species has its own external potential function $u_\alpha(i)$, as in Eq. [1] of the main text, through the varying lengths of the residue hydrophobicity sum. The hard-core nearest-neighbor interaction between species is given by $-\beta^{-1}\log N_{\alpha,\delta}(x)$, where the Greek indices represent the two species of the neighboring segments. $N_{\alpha,\delta}$ actually only depends on the length of the left segment, corresponding to the index α, and is given by equation [2] of the main text. We need to define the forward partition function $Z_F^{(\alpha)}(l)$ of the first *l* sites, subject to the restriction that a segment of species α is present at position *l* (we don't need an index for the particle sitting at position 1 because we will assume that this particle is of fixed species "1"). Then this partition function has boundary condition $Z_F^{(1)}(1) = e^{-\beta u_1(1)}$ for species "1" and equals zero for all other species. We can construct recursion relations for the forward and backward partition functions $Z_F^{(\alpha)}$ and $Z_B^{(\alpha)}$ for each species:

$$Z_F^{(\alpha)}(l+1) = e^{-\beta u_\alpha(l+1)}\sum_{\delta=1}^{s}\sum_{j=1}^{l} Z_F^{(\delta)}(j)N_{\alpha,\delta}(l+1-j)$$

$$Z_B^{(\alpha)}(L-l) = e^{-\beta u_\alpha(L-l)}\sum_{\delta=1}^{s}\sum_{j=1}^{l} Z_F^{(\delta)}(j)N_{\alpha,\delta}(L-l-j)$$

again with the boundary condition $Z_B^{(1)}(L) = e^{-\beta u_1(L)}$ for species "1" and zero for others. The probability to find a segment of species α starting at i is given by:

$$\rho_\alpha(i) = \frac{e^{\beta u_\alpha(i)} Z_F^{(\alpha)}(i) Z_B^{(\alpha)}(i)}{Z_B^{(1)}(1)}$$

Here, we required that the backward partition function have species "1" at site 1 as was assumed for the boundary conditions.

So far we have forced a segment of type "1" to sit at the first and last sites. We would like to relax this constraint and allow segments of any length (or none) at the boundaries. To achieve this, we append an *extra site* to both ends of the system, and treat

particle type “1” as an extra species of hard-core monomer that only binds to the boundary sites. Then the true boundaries of the system are free to fluctuate.