

## Predictive energy landscapes for folding membrane protein assemblies

Ha H. Truong,<sup>1,2</sup> Bobby L. Kim,<sup>1,2</sup> Nicholas P. Schafer,<sup>2,3</sup> and Peter G. Wolynes<sup>1,2,4,a)</sup>

<sup>1</sup>Department of Chemistry, Rice University, Houston, Texas 77005, USA

<sup>2</sup>Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, USA

<sup>3</sup>Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Gustav Wieds Vej 10, 8000 Aarhus C, Denmark

<sup>4</sup>Departments of Physics and Astronomy, Rice University, Houston, Texas 77005, USA

(Received 11 May 2015; accepted 1 July 2015; published online 27 August 2015)

We study the energy landscapes for membrane protein oligomerization using the Associative memory, Water mediated, Structure and Energy Model with an implicit membrane potential (AWSEM-membrane), a coarse-grained molecular dynamics model previously optimized under the assumption that the energy landscapes for folding  $\alpha$ -helical membrane protein monomers are funneled once their native topology within the membrane is established. In this study we show that the AWSEM-membrane force field is able to sample near native binding interfaces of several oligomeric systems. By predicting candidate structures using simulated annealing, we further show that degeneracies in predicting structures of membrane protein monomers are generally resolved in the folding of the higher order assemblies as is the case in the assemblies of both nicotinic acetylcholine receptor and V-type  $Na^+$ -ATPase dimers. The physics of the phenomenon resembles domain swapping, which is consistent with the landscape following the principle of minimal frustration. We revisit also the classic Khorana study of the reconstitution of bacteriorhodopsin from its fragments, which is the close analogue of the early Anfinsen experiment on globular proteins. Here, we show the retinal cofactor likely plays a major role in selecting the final functional assembly. © 2015 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4929598>]

### I. INTRODUCTION

Membranes in the cell are packed with proteins that make up roughly 50% of their volume. In such a crowded environment, it is no surprise that most membrane proteins form parts of larger oligomeric assemblies. As in the case of globular proteins, the energy landscape for folding membrane proteins and the landscape for forming complexes from them must be intimately connected. Ultimately, folding and assembly of proteins within a membrane must rely on the same forces, being modulated by the membrane environment. The fidelity of these processes is the result of aeons of evolution. The mechanistic consequences of the landscapes also must be similar since, apart from the higher local concentration in folding, it is impossible to distinguish the docking events that lead to an oligomeric assembly from the motions needed to organize a fully covalently connected single chain. For globular proteins, the intimate relation between folding and binding landscapes had been well documented a decade ago.<sup>1</sup> Nevertheless, the investigation of the globular protein binding landscapes led to the realization that, in the aqueous environment, water mediated hydrophilic interactions are needed to augment the well-known hydrophobic forces if interfaces are to be predicted with accuracy.<sup>2-5</sup> This experience leads us to inquire whether the coarse-grained force field models found successfully to predict the tertiary folds of membrane proteins in their milieu

also will suffice to predict the structure of larger membrane protein assemblies. In this paper, we explore this issue by using Associative memory, Water mediated, Structure and Energy Model with an implicit membrane potential (AWSEM-membrane) simulations to carry out tertiary structure prediction on several membrane protein assemblies and by analyzing the major basins on the free energy landscapes of this model both for the complexes as a whole and for their constituent monomers in both the absence and presence of a binding partner.

We show here that a transferable, coarse-grained force field inferred using an energy landscape algorithm for folding monomeric membrane proteins also suffices to assemble larger complexes and can predict their tertiary structure at moderate resolution. This force field, called AWSEM-membrane, uses an energy function of the same form as has been used for folding globular proteins: the associative memory water-mediated structure and energy model.<sup>6</sup> The parameters, however, were re-optimized using a database of individual membrane protein domains.<sup>7</sup> Very often, in fact, we find the tertiary structures predicted in the context of the multimeric assembly using AWSEM-membrane are better than those that would be predicted when the monomer is studied in isolation. Degenerate free energy basins in the monomer free energy landscape often turn out to involve forming, internally, native quaternary contacts that have apparently evolved to put together the full multiunit protein assembly. Thus, the near degenerate mis-predictions of the monomer actually correspond to what may be termed internally domain swapped structures.

<sup>a)</sup> Author to whom correspondence should be addressed. Electronic mail: [pwolynes@rice.edu](mailto:pwolynes@rice.edu)

We also examine the question of the mechanism and landscapes of folding versus assembly in the context of the classic investigation carried out by the Khorana group of the assembly of functional bacteriorhodopsin from cleaved fragments. In that classic study, correct reconstitution and proper assembly were tested spectroscopically by whether retinal binding forms a purple product which shows retinal has found its proper protein environment.<sup>8</sup> Here, we show that the AWSEM-membrane energy landscape predicts the initial assembly involves a misfolded species but that, once constraints consistent with the retinal contacts are added, proper assembly follows. This suggests the reconstitution mechanism involves significant rearrangement after the initial fragment binding occurs. This is consistent with the relatively slow time scale of the reconstitution process.<sup>9</sup>

## II. METHODS

The AWSEM-membrane code was recently described in detail in Ref. 10. It is instantiated in the open source LAMMPS simulation code.<sup>11</sup> In some of the simulations carried out in this paper, we used the single memory (SM) AWSEM-membrane model, in which the associative memory term is determined by the structure of the monomer in the experimentally determined structure, to elucidate the role of oligomerization in eliminating the non-native packing of the individual subunit and to predict the structure of dimer complexes of various proteins. In other simulations, we used the fragment memory (FM) AWSEM-membrane model, in which local-in-sequence interactions are derived from a database of structures by performing homology searches using short fragments of the target sequence.<sup>6</sup> AWSEM-membrane employs a coarse-grained backbone description wherein the position and orientation of each amino acid residue are dictated by the positions of its  $C_\alpha$ ,  $C_\beta$ , and  $O$  atoms (except glycine, which lacks a  $C_\beta$  atom). Full details of the functional form of the potential can be found in the supplementary material of Ref. 6 and the re-optimized, membrane protein specific parameters can be found in the supplementary material of Ref. 10.

We only focus on studying the second step of the membrane protein folding process, which occurs after the protein conformations have already been restricted to have a proper topology within the membrane. The topology, by which we mean the “specification of the number of transmembrane helices and their in and/or out orientation across the membrane,”<sup>12,13</sup> was obtained directly from the three dimensional experimentally determined structure using the TMDet web server.<sup>14</sup> Similar results are expected if *a priori* predicted topologies<sup>15</sup> (which are generally quite good) would be used as input, as in our previous study.<sup>10</sup>

In order to sample along  $Q_w$ , the fraction of pairwise distances within 1 Å of their distances in the native structure (Equation (1)), and to construct free energy profiles, we ran umbrella sampling simulations in which a harmonic bias (given in Equation (2)) was added to the Hamiltonian. All free energy profiles and expectation values were calculated using the multi-state Bennett acceptance ratio (MBAR) method as implemented in the pyMBAR package.<sup>16</sup> Samples were

collected for a range of temperatures near the empirically determined binding temperature of each system,

$$Q = \frac{1}{N_p} \sum_i \sum_{j>i+2} \exp \left[ \frac{-(r_{ij} - r_{ij}^\mu)^2}{2\sigma_{ij}^2} \right], \quad (1)$$

$$V_{Q-bias} = \frac{1}{2} k_{Q-bias} (Q - Q_0)^2. \quad (2)$$

## III. RESULTS

AWSEM-membrane is an implicit force field model built on the hypothesis that funneled energy landscapes drive the folding of  $\alpha$ -helical membrane protein monomers within their native topological sector, which is that part of conformational space wherein all the transmembrane helices have adopted the same orientation with respect to the membrane as is seen in the final folded structure. The parameters of the coarse-grained force field have been optimized using an algorithm based on the minimal frustration principle.<sup>17</sup> This algorithm developed by Goldstein *et al.*<sup>18</sup> involves statistically optimizing a Z-score that is monotonically related to the ratio of the folding temperature over the glass transition temperature,  $T_f/T_g$ . The folding temperature is the temperature at which there are equal equilibrium populations of the low energy, low entropy native state and the high energy, high entropy denatured ensemble. The glass transition temperature is the temperature at which a sequence would be expected to become trapped in one of many degenerate low energy structures if that sequence has not been optimized by evolution to fold to a nearly unique native structure. The ratio of these two temperatures is a measure of the degree to which sequence evolution has led to a bias guiding the protein from all parts of the conformational space towards the native state during conformational search. High values of  $T_f/T_g$  indicate a large bias and high specificity in the folded structure. Energy landscapes that have large biases lead to rapid folding and are said to be funneled.<sup>19-23</sup>

The optimization of the AWSEM-membrane model parameters was based on a non-redundant database of  $\alpha$ -helical membrane protein monomer structures.<sup>10</sup> Following the earlier study of the folding landscapes of  $\alpha$ -helical membrane monomers in isolation, we now in this paper examine whether the AWSEM-membrane code can predict the binding interfaces of oligomeric systems. The calculations resolve the issue of whether degenerate tertiary packings that are found in the free energy landscapes of nicotinic acetylcholine receptor subdomain (2BG9) and V-type  $Na^+$ -ATPase subdomain (2BL2) monomers are the result of domain swapping.<sup>1,24</sup> We also revisit the Khorana study of re-association of fragments of bacteriorhodopsin<sup>8</sup> which historically plays the role for membrane proteins that the work of Anfinsen did for globular proteins.<sup>25</sup>

### A. Prediction of the structures and binding interfaces of membrane protein complexes

We carried out structure prediction studies aimed toward studying the binding interfaces of chain A and chain B of nicotinic acetylcholine receptor subdomain (2BG9),<sup>26</sup> V-type

$\text{Na}^+$ -ATPase (2BL2),<sup>27</sup> and bacteriorhodopsin (1BRR)<sup>28</sup> using simulated annealing of the AWSEM-membrane force field, which is implemented in the open-source LAMMPS simulation package.<sup>11</sup> The molecular dynamics simulations begin with two folded monomers separated by approximately 80 Å. A single, weak and non-specific spring potential was used to ensure that the centers of mass of the monomers would be brought together during the course of the annealing. These simulations start at such a high temperature that the flexible monomers are allowed to explore many possible internal conformations and binding interfaces. Following this, the thermostat's temperature was slowly reduced to a quenching temperature. We evaluate the quality of the structures using both  $Q_i$ , the fraction of native interface contacts formed, and the  $C_\alpha$  root-mean-square deviation (RMSD).

Figure 1 shows the predicted quenched structures which have the highest final  $Q_i$  of such simulations for nicotinic acetylcholine receptor subdomain (2BG9) dimers, V-type  $\text{Na}^+$ -ATPase (2BL2) dimers, and bacteriorhodopsin (1BRR) dimers. These results show that the single memory AWSEM-membrane energy landscape gives accurate predictions for the interfaces of the nicotinic acetylcholine receptor subdomain (2BG9) and the V-type  $\text{Na}^+$ -ATPase (2BL2), with  $Q_i$  equal to 0.681 and 0.782, respectively. The predicted  $C_\alpha$  RMSD values are both less than 3.5 Å. The dimer interface of V-type  $\text{Na}^+$ -ATPase (2BL2) was better predicted, according to both its larger  $Q_i$  and its smaller overall  $C_\alpha$  RMSD, than the nicotinic acetylcholine receptor subdomain dimer (2BG9) despite the larger size of the ATPase domain, likely due to its simpler free energy landscape (Figure 4) and larger interface. The bacteriorhodopsin dimer complex turns out to be harder computationally to sample during the simulation because it is a significantly larger system (460 residues) than either nicotinic acetylcholine receptor subdomain (2BG9) (182 residues) or V-type  $\text{Na}^+$ -ATPase (2BL2) (312 residues). Nonetheless, the predicted structure of the bacteriorhodopsin dimer shows that a significant fraction of native interface contacts is formed ( $Q_i = 0.415$ ), and the overall structure is quite native like (RMSD

= 5.732 Å). It should be noted that the experimentally determined structure of bacteriorhodopsin to which we are comparing has a retinal molecule situated in the core of each monomer. This cofactor is omitted from the AWSEM-membrane prediction simulations just described. The absence of the cofactor allows distortions in helical packing which would likely be prohibited due to excluded volume when the retinal cofactor is present.

## B. Non-native helical packings found for the monomers in isolation are disfavored in the presence of their binding partners

In the previous study of individual monomers, the free energy landscape analysis for nicotinic acetylcholine receptor subdomain (2BG9) revealed the presence of two nearly degenerate free energy basins. These basins have similar contact maps, but only one corresponds to structures with the native helical packing, while the other basin centers on a structure that is a pseudo-mirror image of the native structure. The actual contacts that are made in both basins are nearly the same.<sup>10</sup> Since both structural ensembles possess a high fraction of native contacts, they are essentially degenerate according to the contact energy term,  $E_{\text{contact}}$ . Free energy landscape analysis of V-type  $\text{Na}^+$ -ATPase subdomain (2BL2) also revealed the presence of two degenerate free energy basins for the AWSEM-membrane force field. Again, one basin corresponds to the native helical packing but the other basin was characterized as having a non-native helical packing. These two structural ensembles also could not be distinguished by their contact energies. In the earlier study, these proteins were simulated as monomers, but in nature, both proteins are part of larger multimeric assemblies.<sup>26,27</sup> Is this degeneracy of alternative tertiary packings resolved when simulating a monomer in the presence of one of its binding partners? We now answer this question by carrying out free energy landscape analysis for both systems using two instantiations of the AWSEM-membrane prediction scheme. In the first scheme, the local-in-sequence forces determined

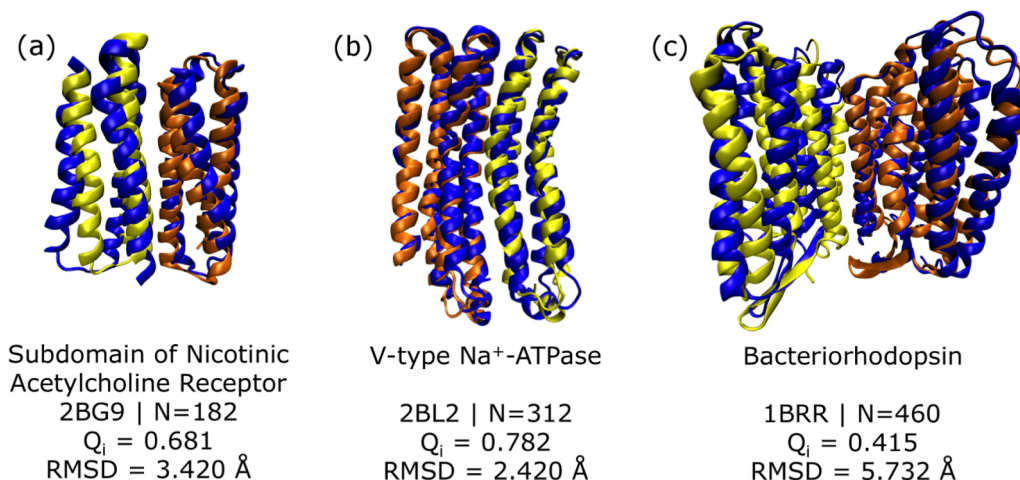


FIG. 1. Structures with the highest final  $Q_i$  values from ten simulated annealing runs using the AWSEM-membrane force field for (a) nicotinic acetylcholine receptor subdomain, (b) V-type  $\text{Na}^+$ -ATPase, and (c) bacteriorhodopsin dimers. In all cases, one of the chains, chain A, is colored in yellow, and the other, chain B, is colored in orange, and the experimental structure of the complex, obtained from the Protein Data Bank,<sup>29</sup> is colored in blue. The names of the proteins, their four-character unique identifier in the Protein Data Bank (PDB ID) and the number of residues are shown below each structure. The fraction of native interface contacts,  $Q_i$  and the  $C_\alpha$  RMSD of the complex compared with the experimental structure indicate the quality of the AWSEM-membrane predictions.

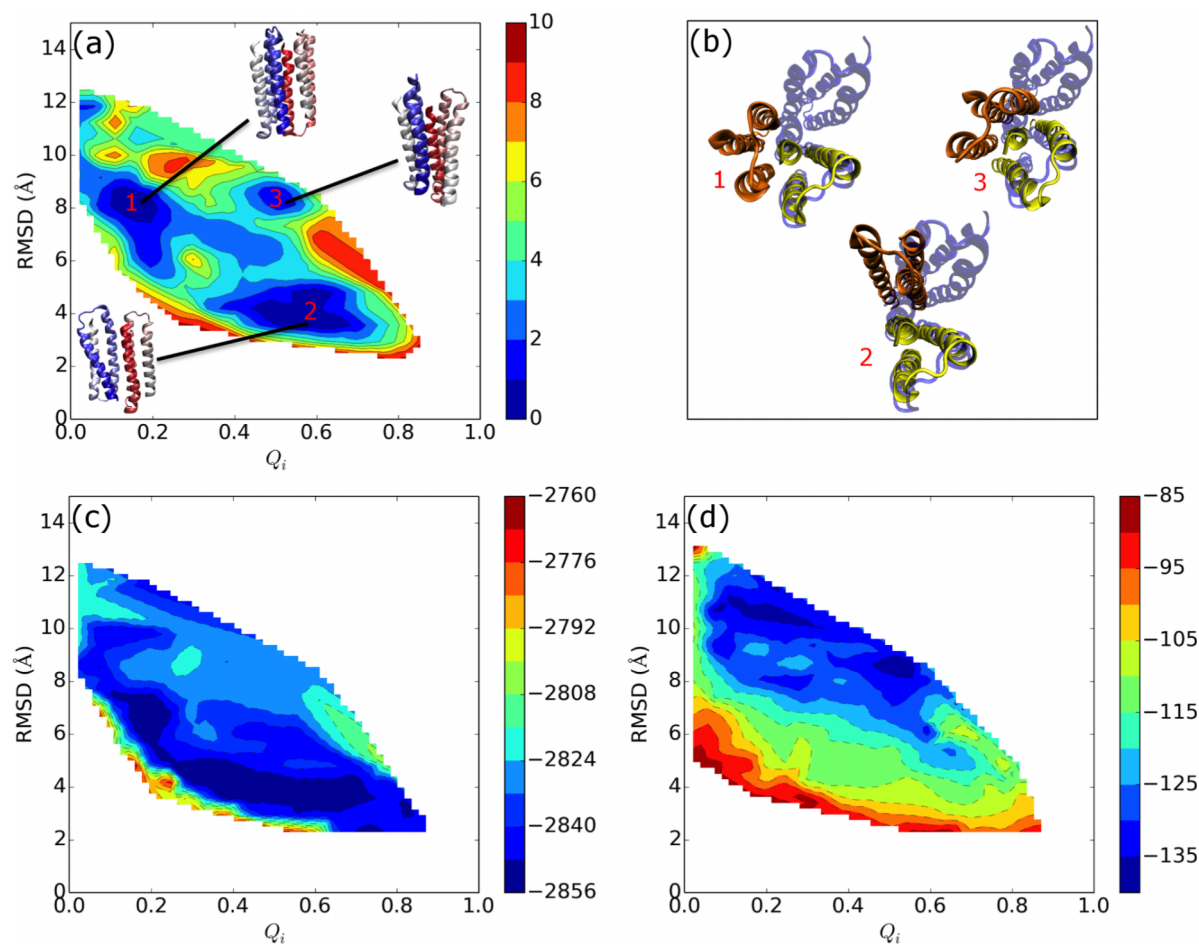


FIG. 2. (a) Free energy profile of the nicotinic acetylcholine receptor subdomain (2BG9) dimer complex obtained using single memory AWSEM-membrane. The free energy is plotted versus  $Q_i$ , the fraction of native interface contacts (x-axis) and the RMSD (y-axis). Representative structures are shown from the three free energy basins. (b) Top views of representative structures from each low free energy basin (chain A is colored in yellow, chain B is colored in orange), with the native structure (shown in transparent blue). Expectation values of (c) the total potential energy,  $PE$ , and (d) the contact energy,  $E_{contacts}$ , are plotted versus the same order parameters.

by the fragment memory term were chosen by performing a homology search of short sequence segments in a database of sequences with known structures and using these as the input associative memories. This approach mimics what must be done when using AWSEM models for fully *de novo* structure prediction. In the second scheme, only a single memory is used for the short range interactions such that only proper native secondary structure information is incorporated. Thus, this landscape has less secondary structure frustration than does the landscape that uses multiple inputs in the fragment memory term.

To be certain to sample a wide range of configurations efficiently, we used umbrella sampling along the collective coordinate  $Q_w$ , a similarity measure based on the fraction of native pairwise distances, and use these sampled structures to construct two-dimensional free energy profiles  $F(Q_i, RMSD)$  for the nicotinic acetylcholine receptor subdomain (2BG9) dimer and for the V-type  $Na^+$ -ATPase (2BL2) dimer. We employ the MBAR method<sup>16</sup> to calculate free energy profiles and expectation values. Free energy is in  $k_B T$ , in which  $T$  was chosen to be below the folding temperature of the monomer but high enough to sample multiple bound configurations. The two order parameters used in the profile are  $Q_i$ , the fraction of native interface contacts, and RMSD, the  $C_\alpha$  root-mean-

square deviation. We also computed the expectation value of the potential energy ( $PE$ ) and the contact energy ( $E_{contact}$ ) for each system and displayed these also as two dimensional profiles with respect to  $Q_i$  and RMSD.

The experimentally determined native binding interface of the first two chains (A and B) of the nicotinic acetylcholine receptor subdomain (2BG9) complex consists of two specific helix-helix interactions as shown in Figure 3(d): one of these interactions involves the association of the first helix of chain A and the third helix of chain B (A1, B3), and the other involves the docking of the second helix of chain A to the second helix of chain B (A2, B2). The free energy profile of the nicotinic acetylcholine receptor subdomain (2BG9), shown in Figure 2(a), has three low free energy basins. We characterize the contact maps of representative structures from each of these basins in Figure 3. Structures in basin 2 are the most native-like in structure. Structures in this basin have well-formed intra-monomer contacts and both native binding interface helix-helix interactions (A1, B3) and (A2, B2), as shown in Figure 3(b). Structures in basin 3 also contain the native binding interface contacts (A1, B3) and (A2, B2), but in this basin, the intra-monomer contacts between helices (A2, A3) and (B2, B3) are disrupted (Figure 3(c)). Structures in basin 1, while successfully forming interface contacts between helices



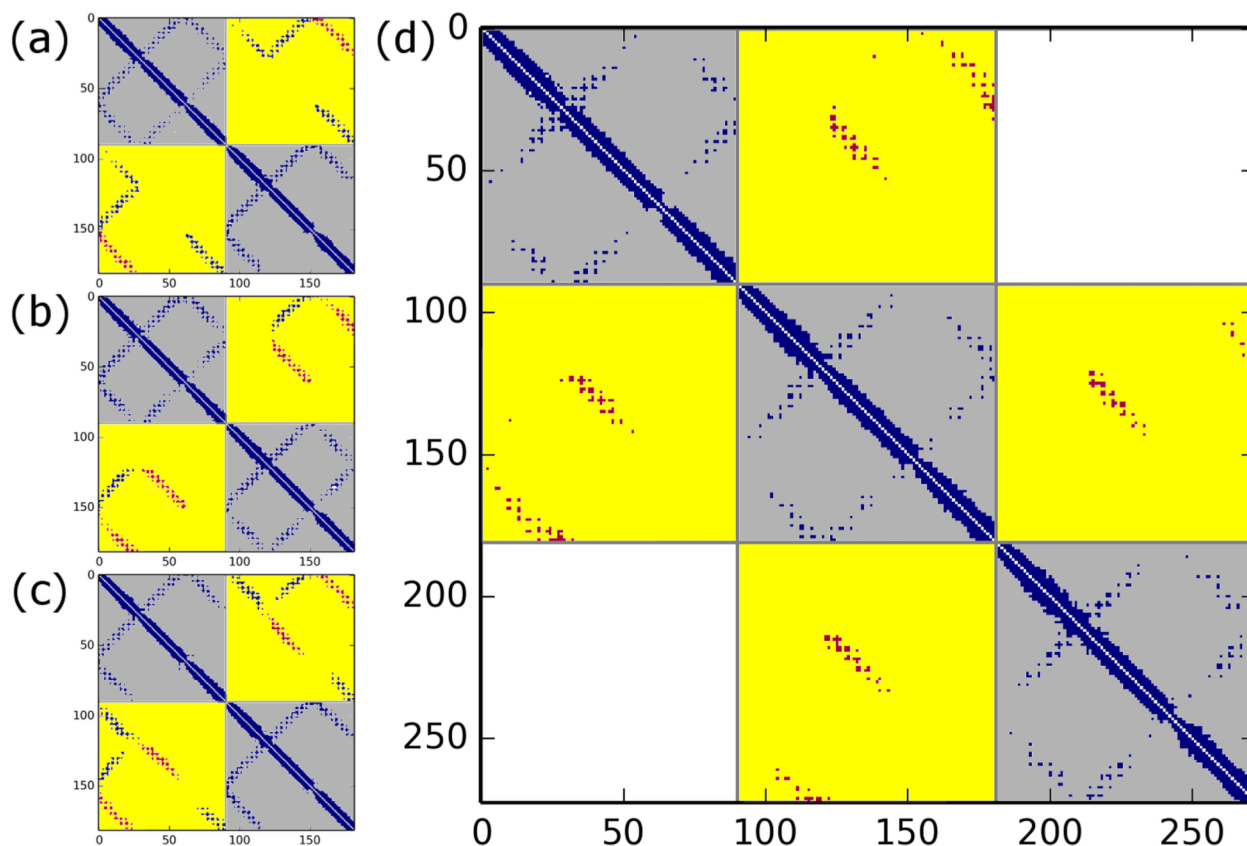


FIG. 3. Contact maps of representative structures obtained from basins of the free energy profile of nicotinic acetylcholine receptor subdomain (2BG9) dimer complex (contact maps (a), (b), and (c) correspond to structures 1, 2, and 3 in Figure 2, respectively) and (d) contact map of nicotinic acetylcholine receptor subdomain (2BG9) trimer including chain A, chain B, and chain E of the full pentamer complex. Sections of the maps that show intra-monomer contacts are colored in gray. Sections of the maps that show inter-monomer contacts are colored in yellow. Inter-monomer contacts found in both the low free energy structures and in the trimer complex are colored in red.

(A1, B3), lack stable native interface interactions between helices (A2, B2) and intra-monomer helix-helix interactions between (B2, B3) (Figure 3(a)). Structures in all three basins show some degree of over-collapse resulting from formation of non-native interface contacts. Nicotinic acetylcholine receptor is a pentamer in its crystal structure, and this over-collapse will likely be resolved when folding the complete multimeric assembly. The observed over-collapse may also be a consequence of the generic cylindrical radius of gyration bias ( $R_g$ ) employed in the AWSEM-membrane code. This bias is similar to typical  $R_g$  bias used for globular proteins but is applied only to coordinates in the membrane plane. This constraint was originally implemented in order to mimic the lateral pressure of the membrane lipid molecules on the protein that gives rise to the liquid crystalline-like ordering of helices in membrane proteins.

Figures 2(c) and 2(d) show two dimensional profiles of the expectation values of the potential energy,  $PE$ , and of the contact energy,  $E_{contacts}$ , for the nicotinic acetylcholine receptor subdomain (2BG9) dimer, respectively. The full potential energy which includes both contact terms and associative memory terms appears to favor basins 1 and 2, while the contact energy landscape by itself dominantly favors basin 3 and moderately favors basin 2. Basin 3 is disfavored in the full potential energy landscape primarily because of the distortion of the intra-monomer structure of both chains in structures found in this basin. These configurations are energetically penalized by  $E_{SM}$ ,

the single memory term which favors proper secondary structure. Conversely, structures in basin 2 have fully native-like intra-monomer structure of both chains. Basin 1 is slightly less favored in the full potential energy landscape than is basin 2 due to the distortion observed in the intra-monomer structure of chain B as discussed above. Why are basins 1 and 3 favored when only the contact energy is considered? The over-collapse of both states allowed by the local distortion simply leads to a larger gross number of contacts formed when compared to basin 2, as is evident in the contact maps in Figure 3.

In Figure 4, we show the free energy profiles for the V-type  $Na^+$ -ATPase (2BL2) dimer. Here, the properly folded structure is strongly favored; the landscape has only one basin at low RMSD (between 2 Å and 3.5 Å) and high  $Q_i$  ( $Q_i \geq 0.7$ ), which corresponds to the native structure. The expectation values of the total potential energy,  $PE$ , and contact energy,  $E_{contacts}$ , both show the native conformation to be the most stable.

Figure 5 shows the results of the free energy landscape analysis for the nicotinic acetylcholine receptor subdomain (2BG9) and the V-type  $Na^+$ -ATPase (2BL2) dimer complexes using fragment memory AWSEM-membrane. The two dimensional free energy profile of nicotinic acetylcholine receptor subdomain (2BG9) dimer complex (Figure 5(a)) reveals two low free energy basins. Representative structures from the low RMSD basin are significantly native-like, forming native binding interface helix-helix interactions (A1, B3) and (A2, B2), with moderate helix distortion. The representative

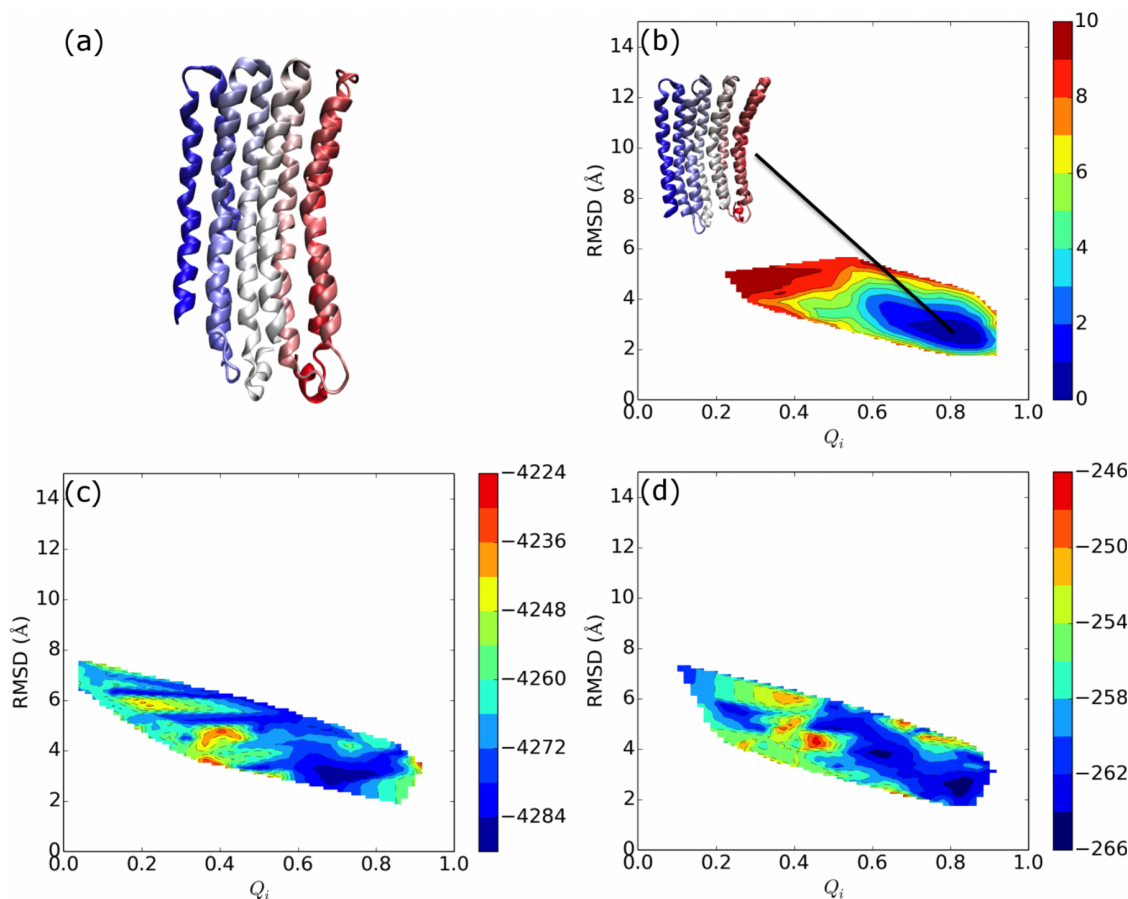


FIG. 4. (a) Native configuration of the V-type  $Na^+$ -ATPase (2BL2) dimer complex visualized using Visual Molecular Dynamics (VMD).<sup>30</sup> (b) Free energy profile of the V-type  $Na^+$ -ATPase (2BL2) dimer complex obtained using single memory AWSEM-membrane. The free energy is plotted versus  $Q_i$ , the fraction of native interface contacts (x-axis) and the RMSD (y-axis). A representative structure from the free energy basin is shown. The expectation values of (c) the total potential energy,  $PE$ , and (d) the contact energy,  $E_{contact}$ , are plotted versus the same order parameters.

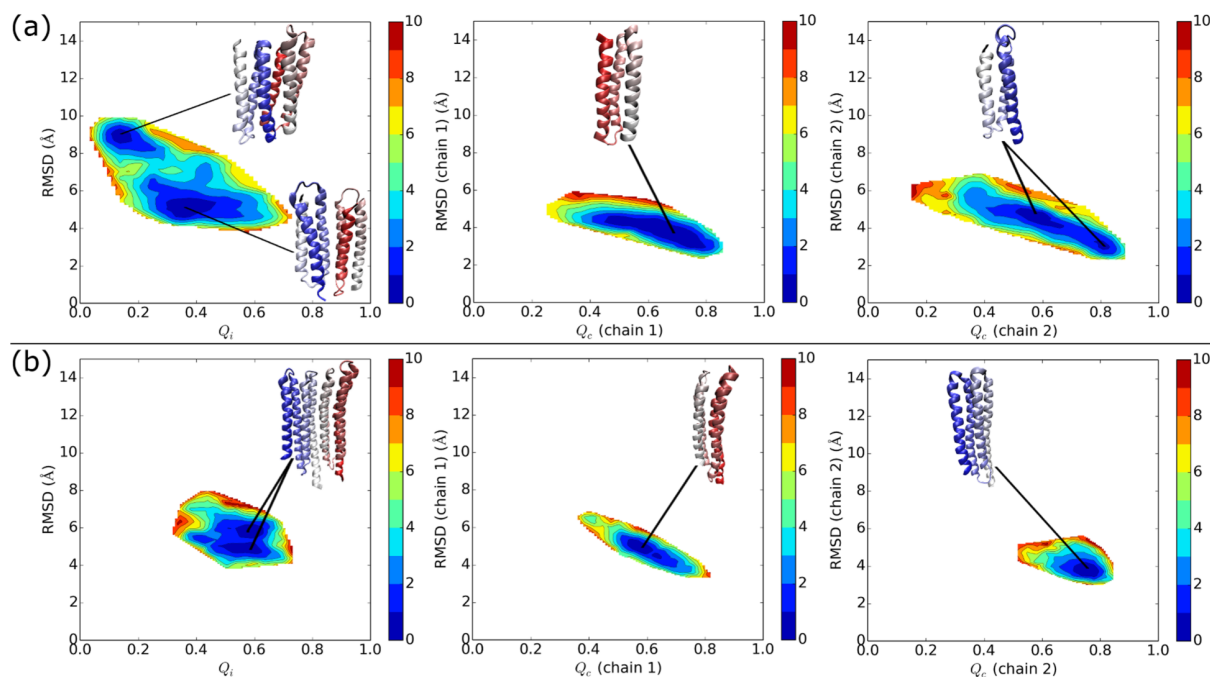


FIG. 5. Free energy profile for (a) the nicotinic acetylcholine receptor subdomain (2BG9) and (b) the V-type  $Na^+$ -ATPase (2BL2) dimer complexes obtained by using the fragment memory AWSEM-membrane code. From left to right, the free energy is plotted versus  $Q_i$ , the fraction of native interface contacts (x-axis) and the RMSD aligned to the dimer complex (y-axis), and versus  $Q_c$ , the fraction of native contacts of each subunit (x-axis) and the RMSD of structures aligned to each subunit (y-axis), respectively. Representative structures from the low free energy basins are shown. Free energy profiles for the nicotinic acetylcholine receptor subdomain (2BG9) monomer and the V-type  $Na^+$ -ATPase (2BL2) monomer are shown in Figures 5 and 6 of Ref. 10.

structure from the high RMSD basin however exhibits a non-native association of the monomers such that the non-native helix-helix interactions (A1, B2) and (A3, B3) are formed. However, both subunits maintain a native helical packing in both basins, as shown in the free energy profile of chain 1 and chain 2 plotted versus  $Q_c$ , the fraction of intra-monomer native contacts, and RMSD for each chain, respectively (the second and third panels of Figure 5(a)). There is only one low free energy basin at high  $Q_c$  (above 0.5) and low RMSD (below 5 Å), which corresponds to the native monomeric helical packing.

The two dimensional free energy profile of the V-type  $Na^+$ -ATPase (2BL2) dimer complex (Figure 5(b)) exhibits two low free energy basins at high  $Q_i$  (from 0.5 to 0.65) and relatively low RMSD (at  $\approx 6$  Å and  $\approx 5$  Å). These basins are separated by a low free energy barrier. Both of these two basins contain representative structures which are nearly native having more than 55% of the native interface contacts formed ( $Q_i \approx 0.55$ ). The difference in RMSD ( $\approx 1$  Å) is mostly the result of a small helix distortion and the overall over-collapse of the structure. As for the nicotinic acetylcholine receptor subdomain (2BG9), we did not observe any stable non-native packings for the monomers. The absence of non-native contacts in the dimer contrasts with our previous study of the monomer energy landscape which did display non-native contacts. For the monomer, both the native and a particular non-native helical packing were found to be nearly degenerate in free energy.<sup>10</sup> When simulated in the presence of another monomer, however, as shown in the second and third panels of Figure 5(b), only one low free energy basin is found in the free energy profile for folding each individual chain. This basin corresponds to a structure which has fully native helical packing. In other words, the non-native helical packing basins of the monomeric form of nicotinic acetylcholine receptor subdomain (2BG9) and V-type  $Na^+$ -ATPase (2BL2) are resolved when multiple chains are present and are allowed to interact in the simulation. In the case of the nicotinic acetylcholine receptor subdomain (2BG9), the native helical packing within each monomer is maintained whether the multimeric complex has a near-native binding interface or an alternative, non-native binding interface.

### C. Re-association of fragments of bacteriorhodopsin monomer and the role of cofactor and fragment rigidity

One of the classic and indeed heroic early experimental studies of membrane protein folding was undertaken by Khorana's group in 1980s. They showed that fragments of bacteriorhodopsin could reassociate in the presence of retinal to form a functional molecule.<sup>8</sup> We re-visit computationally their laboratory study of the re-association of the cleaved bacteriorhodopsin monomer from two of its fragments,  $C_2$  consisting of the first and second helices of bacteriorhodopsin and  $C_1$  consisting of the remaining five helices.

The experimentally determined structure of the bacteriorhodopsin monomer and simulated structures from two example structure predictions using single memory AWSEM-membrane are shown in Figure 6. The experimentally deter-

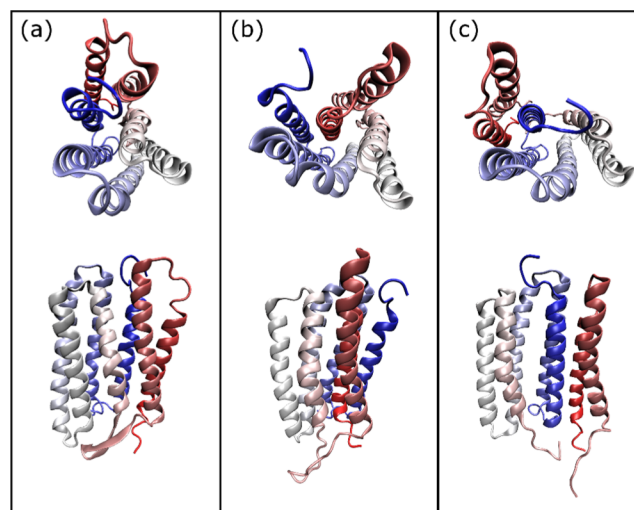


FIG. 6. The top view and side view of (a) the experimentally determined bacteriorhodopsin monomer structure (1BRR), (b) a predicted structure using single memory AWSEM-membrane for the intact bacteriorhodopsin monomer, and (c) a predicted structure using single memory AWSEM-membrane for the cleaved bacteriorhodopsin monomer. Figures were generated using VMD.<sup>30</sup>

mined structure of bacteriorhodopsin (Figure 6(a)) has a retinal molecule situated in its core. This cofactor supports a configuration in which the seven helices pack around it in an overall elliptic cylinder shape. When the retinal cofactor is omitted from the simulations, we observed over collapsed configurations and distortions in helical packing in both the intact and the cleaved bacteriorhodopsin systems. A predicted structure of the intact bacteriorhodopsin monomer, despite having more than 66% of the native contacts formed ( $Q_c = 0.663$ ) and  $RMSD = 6.459$  Å, is over collapsed. The first helix is buried and is surrounded by the other six helices as shown in Figure 6(b). The predicted structure of the cleaved bacteriorhodopsin monomer (Figure 6(c)) has less than 60% of the overall native contacts formed ( $Q_c = 0.572$ ) and a still larger  $RMSD (=7.516$  Å). The binding interface of fragments  $C_1$  and  $C_2$  is also incorrectly predicted in the cleaved system. We observed non-native helix-helix interactions between the first helix and the sixth helix, ( $C_2-1, C_1-6$ ), and between the second helix and the seventh helix, ( $C_2-2, C_1-7$ ). Based on these observations, we infer that the retinal cofactor likely plays an important role in the reconstitution of cleaved bacteriorhodopsin and its effects must be taken into account in the simulations.

To mimic the effects of retinal, we applied three pairwise distance constraints to residue pairs in fragment  $C_1$  alone that are in contact with the retinal molecule in the crystal structure. Note that there are no constraints that connect the two fragments together. These constraints, internal to fragment  $C_1$ , partially compensate for the lack of an explicit representation of the retinal molecule in our simulations. We also increased the strength of the memory term in order to rigidify the secondary structure where flexibility may also contribute to over-collapse. Figure 7(a) summarizes the results from ten simulated annealing runs. The  $Q_i$  value of the final snapshot in the runs is plotted against the total potential energy ( $PE$ ). The structure which has the highest  $Q_i$  ( $Q_i = 0.764$ ) is a correctly bound structure in which the two fragments have re-associated

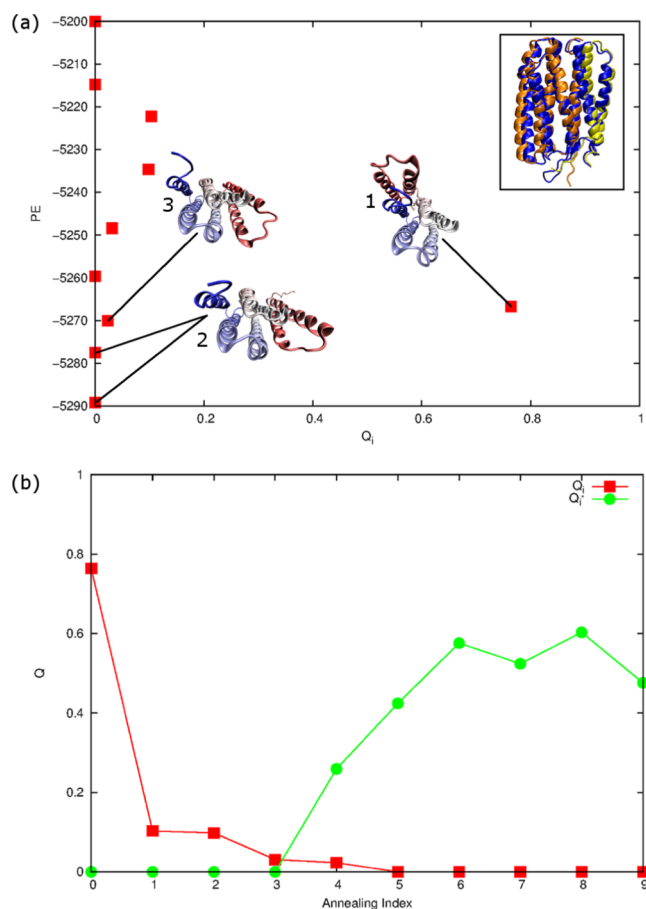


FIG. 7. (a) The final snapshots of simulated annealing runs are plotted as a function of the fraction of native interface contacts,  $Q_i$ , and the total potential energy,  $PE$ . Top view of the snapshots of the predicted structures is shown and colored according to residue index starting at the N-terminus (red) and going to the C-terminus (blue): Structure (1) is a correctly bound structure, which is shown superimposed on the native structure in the inset (yellow: fragment  $C_2$ , orange: fragment  $C_1$ , blue: native structure). Structures (2) and (3) are competitive low energy predicted structures. (b) Final  $Q$  versus annealing index of dimer interface predictions of the cleaved bacteriorhodopsin monomer. Ten independent simulated annealing simulations were conducted and their final  $Q_i$  values, the fractions of native interface contacts formed, were plotted in the order of decreasing  $Q_i$  from left to right.  $Q_i$ , the fraction of native interface contacts formed, and  $Q_i^*$ , the fraction of dimerization interfacial contacts formed, are plotted in red and green, respectively. Note that the “annealing index” does not refer to the actual order in which the simulated annealing simulations were carried out.

to form the native structure (structure (1) in Figure 7(a)). This structure agrees well with the native crystal structure of the monomer and has a very high fraction of native contacts formed ( $Q_c = 0.877$ ) and a very low RMSD for such a large system ( $= 2.23 \text{ \AA}$ ). Three other energetically competitive structures were also observed in this set of simulated annealing runs. These are given labels (2) and (3) in Figure 7. These structures all share helix-helix interactions between the second helix that belongs to the  $C_2$  fragment and the fourth helix that belongs to  $C_1$  fragment ( $C_2$ -2,  $C_1$ -4). Although not present in the monomer crystal structure, these strong helix-helix interactions are found on the binding interfaces in the complete bacteriorhodopsin trimer complex. Thus, we can view these structures as resulting from a kind of domain swapping. To further investigate these ( $C_2$ -2,  $C_1$ -4) helix-helix interactions, we created a modeled domain swapped structure which con-

sists of the first and second helices (fragment  $C_2$ ) of chain A of the experimentally determined bacteriorhodopsin trimer complex and the last five helices (fragment  $C_1$ ) of chain B of the experimentally determined bacteriorhodopsin trimer complex. This modeled domain swapped structure has the ( $C_2$ -2,  $C_1$ -4) interfacial helix-helix interactions mentioned above and is used as reference structure for calculating  $Q_i^*$ , the fractions of dimerization interfacial contacts formed, and  $RMSD_{swapped}$ . The ( $C_2$ -2,  $C_1$ -4) dimerization interfacial helix-helix interactions were observed to some degree in six out of the ten simulated annealing simulations that we carried out, as shown in Figure 7(b). Five of the simulated annealing runs produced structures with  $Q_i^* > 0.4$ , indicating that half of the structures have 40% or more of the dimerization interfacial contacts formed.

An analysis of the contacts found in the representative predicted structures is shown in Figure 8. All of the native intra-fragment contacts in the two-helix fragment  $C_2$  and five-helix fragment  $C_1$  are present in all three structures (Figures 8(a)–8(c)). The native binding interface of the cleaved bacteriorhodopsin monomer involves helix-helix interactions between the second helix and third helix ( $C_2$ -2,  $C_1$ -3), between the first helix and seventh helix ( $C_2$ -1,  $C_1$ -7), and between the second helix and seventh helix ( $C_2$ -2,  $C_1$ -7). The native-like structure (1) has all of the native monomer interface contacts ( $C_2$ -2,  $C_1$ -3), ( $C_2$ -1,  $C_1$ -7), and ( $C_2$ -2,  $C_1$ -7) formed, as is shown in Figure 8(a). Therefore, this contact map looks very similar to the contact map of the bacteriorhodopsin monomer subunit (colored gray in Figure 8(d)). Structure (2) and structure (3) do not have the native monomer interface contacts formed, but instead both have the dimerization interface helix-helix interactions ( $C_2$ -2,  $C_1$ -4) formed, as we mentioned previously. These ( $C_2$ -2,  $C_1$ -4) helix-helix interactions are the same as the contacts that are made at the protein-protein interfaces of the larger bacteriorhodopsin trimer (colored in red and shown in the yellow region of the contact maps). The other, non-native, helix-helix interactions ( $C_2$ -1,  $C_1$ -4) and ( $C_2$ -1,  $C_1$ -5) found at the interface of structure (2), and the non-native, helix-helix interactions ( $C_2$ -1,  $C_1$ -4) and ( $C_2$ -2,  $C_1$ -5) of structure (3), are not shared with the bacteriorhodopsin trimer.

Figure 9 shows the results of the free energy landscape analysis of the association of the cleaved bacteriorhodopsin complex using the single memory AWSEM-membrane force field. The two dimensional free energy profile with respect to  $RMSD_{swapped}$ , aligned to the modeled domain swapped cleaved bacteriorhodopsin structure, and  $RMSD_{native}$ , aligned to the native cleaved bacteriorhodopsin structure, contains three low free energy basins (labeled 1, 2, and 3). The profile also exhibits two other somewhat energetically competitive basins with higher free energy ( $\approx 3k_B T$ ) (labeled 4 and 5). Representative structures from both of the low  $RMSD_{native}$  and high  $RMSD_{swapped}$  basin (basin 1) are significantly native-like. The high  $RMSD_{native}$  and high  $RMSD_{swapped}$  basin (basin 5) contains nonspecifically bound structures. Structures in this basin have neither the binding interface contacts of the native structure nor the proper dimerization contacts. Representative structures from the three other low free energy basins (basins 2, 3, and 4) all have the ( $C_2$ -2,  $C_1$ -4) dimerization helix-helix interactions,



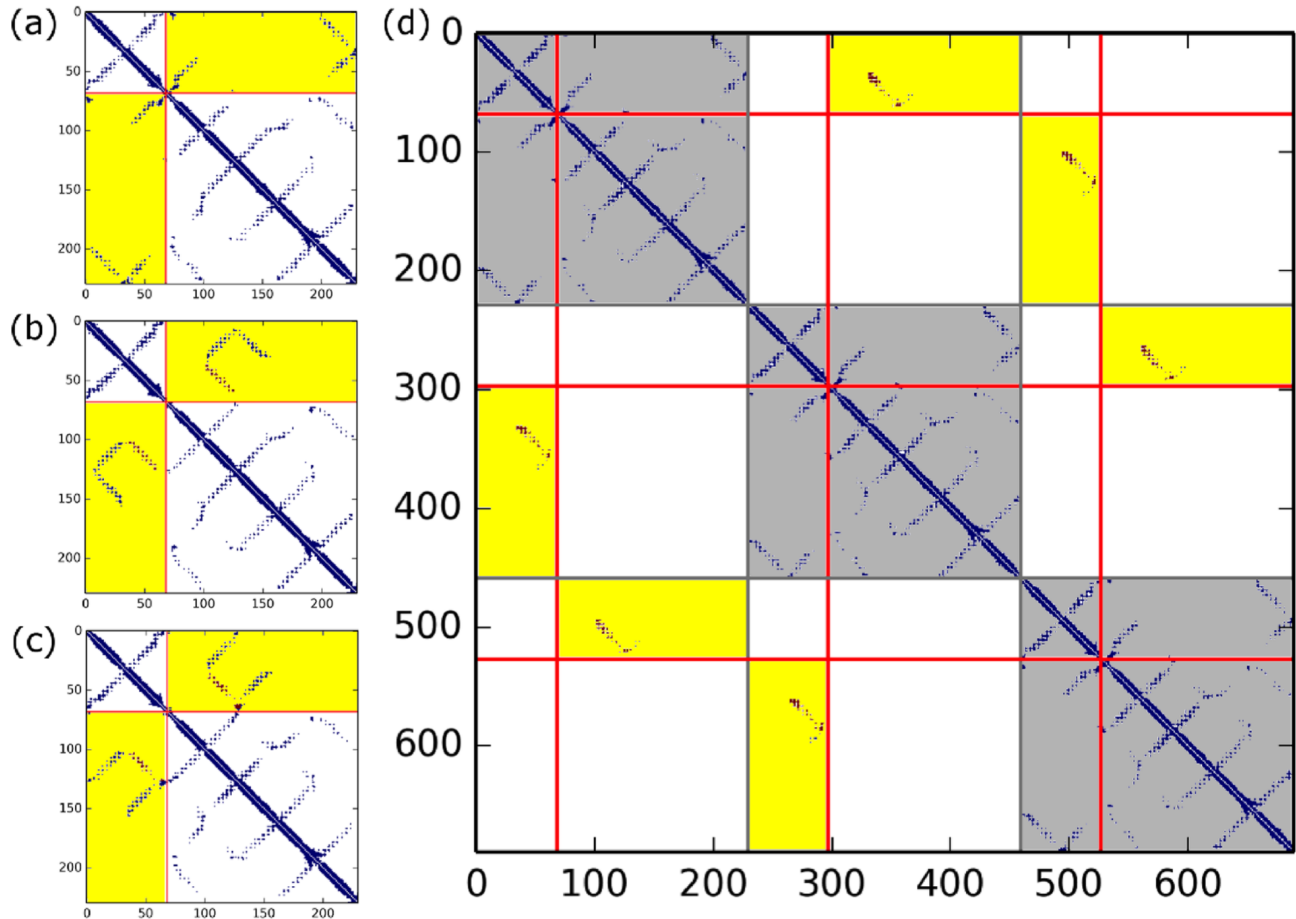


FIG. 8. Contact maps of representative structures obtained from simulated annealing simulations of the cleaved bacteriorhodopsin monomer (contact maps (a), (b), and (c) correspond to structure (1), (2), and (3) in Figure 7, respectively) and the contact map of the bacteriorhodopsin trimer complex (d). Sections of the maps that show intra-monomer contacts are colored in gray. Sections of the maps that show inter-subunit contacts are colored in yellow, in which conserved contacts found in simulated annealing structures and the inter-subunit contacts in the trimer complex are colored in red. Red lines separate the  $C_1$  fragment and  $C_2$  fragment of the cleaved bacteriorhodopsin molecule.

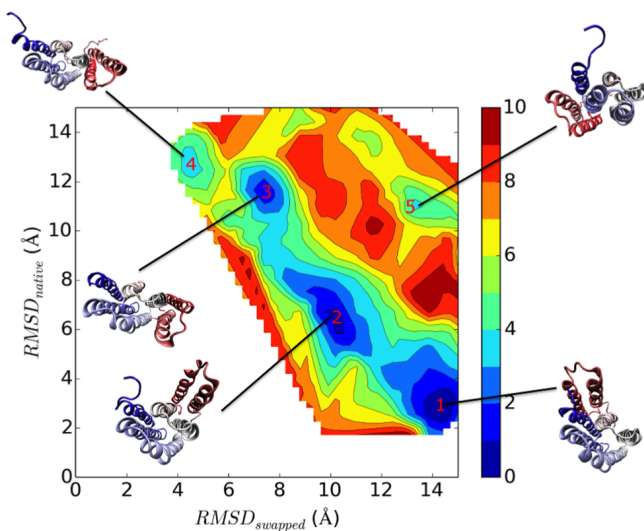


FIG. 9. Free energy profile of the cleaved bacteriorhodopsin assembly obtained using the single memory AWSEM-membrane code. Free energy is in  $k_B T$ , in which  $T$  was chosen to be below the folding temperature of the monomer but high enough to sample multiple bound configurations. The free energy is plotted versus  $RMSD_{swapped}$  (x-axis) and  $RMSD_{native}$  (y-axis). Representative structures from the low free energy basins are shown.

and become more similar to the modeled reference dimer structure as  $RMSD_{swapped}$  decreases. Figure 9 together with Figure 7 shows that there are two dominant states of the cleaved bacteriorhodopsin monomer: the natively bound state and the state that contains dimerization contacts that would form in the higher order assembly. All other, nonspecifically bound states are higher in free energy.

#### IV. DISCUSSION

We were able to predict native-like binding interfaces for dimeric complexes with various topologies (three helix bundle to seven helix bundle) and sizes (up to 460 residues) using the AWSEM-membrane force field. Using both the single memory and fragment memory flavors of the force field, we showed that oligomerization of the domains as occurs in the full *in vivo* assembly eliminates the non-native helical packing basins that were previously observed in the energy landscapes of the monomers of the nicotinic acetylcholine receptor subdomain (2BG9) and the V-type  $Na^+$ -ATPase (2BL2) monomers when they were simulated by themselves.

The retinal cofactor plays an important role in the folding process of bacteriorhodopsin. We found over-collapsed configurations and distortions in helical packing occur in

both the intact and the cleaved bacteriorhodopsin simulations when the retinal cofactor is omitted from the simulations. We observed, however, proper association of the bacteriorhodopsin monomer fragments, the two-helix fragment,  $C_2$ , and the five-helix fragment,  $C_1$ , when the force field is augmented through the aid of three pairwise distance constraints to residue pairs that make heavy atom contact with the retinal cofactor in the experimental structure and the aid of the rigidified secondary structure. The key role of the cofactor is consistent with the observations made in Khorana's experiments in which the protein was only shown to refold into a functional form after retinal was added.<sup>8</sup> We also observed that simulated annealing of the cleaved bacteriorhodopsin fragments often resulted in structures that contain dimerization helix-helix interactions instead of the helix-helix interactions that would lead to the proper monomer structure. In a certain sense, these structures are not misfolded but in fact can be viewed as resulting from a kind of internal domain swapping. This view would be consistent with the principle of minimal frustration applying in full force to membrane proteins much as it does for globular proteins. These domain swapped states are competitive in free energy terms with the native state when the constraints normally imposed by chain connectivity are relaxed by cleavage of the monomer into two fragments.

## ACKNOWLEDGMENTS

H.H.T. thanks Weihua Zheng for helpful discussions. The project described was supported by Grant No. R01 GM44557 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute of General Medical Sciences or the National Institutes of Health. Additional support was also provided by the D.R. Bullard-Welch Chair at Rice University, Grant No. C-0016. This work was also supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by NSF under Grant No. OCI-0959097.

- <sup>1</sup>Y. Levy, S. S. Cho, J. N. Onuchic, and P. G. Wolynes, *J. Mol. Biol.* **346**, 1121 (2005).
- <sup>2</sup>G. A. Papoian, J. Ulander, and P. G. Wolynes, *J. Am. Chem. Soc.* **125**, 9170 (2003).
- <sup>3</sup>G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3352 (2004).
- <sup>4</sup>C. Zong, G. A. Papoian, J. Ulander, and P. G. Wolynes, *J. Am. Chem. Soc.* **128**, 5168 (2006).
- <sup>5</sup>W. Zheng, N. P. Schafer, A. Davtyan, G. A. Papoian, and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19244 (2012).
- <sup>6</sup>A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, *J. Phys. Chem. B* **116**, 8494 (2012).
- <sup>7</sup>C. A. Schramm, B. T. Hannigan, J. E. Donald, C. Keasar, J. G. Saven, W. F. DeGrado, and I. Samish, *Structure* **20**, 924 (2012).
- <sup>8</sup>K. S. Huang, H. Bayley, M. J. Liao, E. London, and H. G. Khorana, *J. Biol. Chem.* **256**, 3802 (1981).
- <sup>9</sup>J.-L. Popot, S.-E. Gerchman, and D. M. Engelman, *J. Mol. Biol.* **198**, 655 (1987).
- <sup>10</sup>B. L. Kim, N. P. Schafer, and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.* **111**, 11031 (2014).
- <sup>11</sup>S. Plimpton, *J. Comp. Phys.* **117**, 1 (1995).
- <sup>12</sup>G. Von Heijne, *J. Mol. Biol.* **225**, 487 (1992).
- <sup>13</sup>G. von Heijne, *Nat. Rev. Mol. Cell Biol.* **7**, 909 (2006).
- <sup>14</sup>G. E. Tusnady, Z. Dosztanyi, and I. Simon, *Bioinformatics* **21**, 1276 (2005).
- <sup>15</sup>T. Nugent and D. T. Jones, *BMC Bioinf.* **10**, 159 (2009).
- <sup>16</sup>M. R. Shirts and J. D. Chodera, *J. Chem. Phys.* **129**, 124105 (2008).
- <sup>17</sup>J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.* **84**, 7524 (1987).
- <sup>18</sup>R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.* **89**, 9029 (1992).
- <sup>19</sup>P. E. Leopold, M. Montal, and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **89**, 8721 (1992).
- <sup>20</sup>J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins: Struct., Funct., Bioinf.* **21**, 167 (1995).
- <sup>21</sup>J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *Annu. Rev. Phys. Chem.* **48**, 545 (1997).
- <sup>22</sup>P. G. Wolynes, *Philos. Trans. R. Soc. London, Ser. A* **363**, 453 (2005).
- <sup>23</sup>N. P. Schafer, B. L. Kim, W. Zheng, and P. G. Wolynes, *Isr. J. Chem.* **54**, 1311 (2014).
- <sup>24</sup>S. Yang, S. S. Cho, Y. Levy, M. S. Cheung, H. Levine, P. G. Wolynes, and J. N. Onuchic, *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13786 (2004).
- <sup>25</sup>C. B. Anfinsen *et al.*, *Science* **181**, 223 (1973).
- <sup>26</sup>N. Unwin, *J. Mol. Biol.* **346**, 967 (2005).
- <sup>27</sup>T. Murata, I. Yamato, Y. Kakinuma, A. G. Leslie, and J. E. Walker, *Science* **308**, 654 (2005).
- <sup>28</sup>L.-O. Essen, R. Siebert, W. D. Lehmann, and D. Oesterhelt, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 11673 (1998).
- <sup>29</sup>H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res.* **28**, 235 (2000).
- <sup>30</sup>W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* **14**, 33 (1996).