Multi-Scale Hierarchical Structure Prediction of Helical Transmembrane Proteins

Zhong Chen
Dept. of Biochemistry and Molecular Biology
University of Georgia, Athens, GA 30602
Email: zc@csbl.bmb.uga.edu

Ying Xu
Dept. of Biochemistry and Molecular Biology
University of Georgia, Athens, GA 30602.
Email: xyn@bmb.uga.edu

Abstract

As the first step toward a multi-scale, hierarchical computational approach for membrane protein structure prediction, the packing of transmembrane helices was modeled at the residual and atomistic levels, respectively. For predictions at the residual level, the helix-helix and helix-lipid interactions were described by a set of knowledge-based energy functions. For predictions at the atomistic level, CHARMM19 force field was employed. To facilitate the system to overcome energy barriers, Wang-Landau sampling was carried out by performing a random walk in the energy and conformational spaces. Native-like structures were predicted at both levels for 2- and 7-helix systems. Interestingly, consistent results were obtained from simulations at residual and atomistic levels for the same system, strongly suggesting the feasibility of a hierarchical approach for membrane structure prediction.

1. Introduction

Membrane proteins play important roles in many cellular processes such as signal transduction, respiration, photosynthesis, cell trafficking, and transport of compounds and ions. The clinical importance of membrane proteins cannot be overstated as they are the leading targets for drug discovery. However our understanding of their 3-dimensional structures and their functions is still limited due to the non-trivial task of structural determination of membrane proteins. Among solved protein structures in the Protein Data Bank (PDB), less than 2% are membrane proteins though on average 20-30% of the genes in a genome encode membrane proteins. While there have been a number of studies using *ab initio* prediction techniques, successful prediction results have been limited to small and simple systems [1,2,3]. More sophisticated and faster algorithms are clearly desirable for more practical applications.

The major difficulties in structure prediction of transmembrane (TM) proteins come from the enormous conformational search space and the extremely hilly free-energy surface. The great specificity in helical packing requires simulations with atomistic details for realistic modeling. However, the rugged free-energy landscape and the time-consuming energy calculations at atomistic level render a thorough sampling of the conformational space near impossible with the current computing power even for modest complex systems. A more practical approach is probably a multi-scale, hierarchical method where the conformational space is explored at the residue level with a coarse-grained representation and then at the atomistic level for structure refinement. As a first step toward such a goal, in this paper, several model helical bundles with 2 or 7 helices were modeled at the residual and atomistic levels respectively. The objectives of this study include: (1) to develop effective simulation techniques at both levels; (2) to show that with careful consideration of the helix-helix and helix-lipid interactions, simulations results at the residual level correspond approximately to those from atomistic simulations, then validate the feasibility of a multi-scale approach; and (3) to provide critical insights about energy function derivations at the residual level and the development of a hierarchical scheme in the end.

2. Methodology

2.1 Wang-Landau algorithm

Wang-Landau method [4] directly obtains g(E), i.e., the number of all possible states (or configurations) at a particular energy level E of the system, in a self-consistent manner. This method is based on the observation that if a random walk is performed with a

probability proportional to the reciprocal of the density of states, i.e., $p(E) \propto 1/g(E)$, then a flat energy histogram can be obtained. Each time when an energy level E is visited, the corresponding density of states is updated by a modification factor f >1, i.e., $g(E) \to g(E)f$. During the random walk, f approaches 1 and g(E) approaches the correct densities of states. In this work, f_0 is set to 2.71828, and f_{final} is set to $\exp(10^{-6}) \approx 1.000001$.

Since all the energy levels are visited with an equal probability in the Wang-Landau method, this algorithm allows the system to escape from local energy minima (LEM) and force a broad sampling of the conformational space. In addition, it also provides the stability information of the system based on

$$P(E, T)=g(E)\exp(-E/kT)$$

where P is the probability density, k is the Boltzmann constant, and T is the temperature. Since the publication of the original paper, the Wang-Landau algorithm has had many successful applications in complex systems which have proved to be very difficult for conventional simulation techniques, such as glasses, dense fluids, polymers, and proteins [5].

2.2 Simulation details

For residue-level simulations, each amino acid residue is coarse-grained as an interacting point centered at the C_α position. Two statistical potentials were derived from known TM structures, which describe the helix-lipid and helix-helix interactions, respectively. For this study, we used 97 protein chains of 40 TM proteins from the PDB, no two of which have > 30% sequence identity. For the helix-lipid potential, the distributions of all the residues as a function of their distance along the bilayer normal to the central plane of the membrane were collected. A statistical potential is then calculated using

$$V_{lp}(a,z) = -c\ln\left(\frac{N_{obs}(a,z)}{N_{\exp}(a,z)}\right)$$

where a is the type of an amino acid, and z is the distance to the center plane along the bilayer normal direction; $N_{obs}(a,z)$ is the observed relative frequency of occurrences of amino acid type a at z, while $N_{\exp}(a,z)$ is the expected frequency if a uniform distribution is assumed. c is a constant in units of kcal/mol. The resulted potential seems to be able to describe the preference of hydrophobic residues to be located in the hydrocarbon core while the more hydrophilic residues of the same helix to be closer to the terminal regions of the lipid bilayers. The validity of this potential was tested through predicting the tilt angle with the lipid bilayer normal when a single TM helix is inserted into the membrane. Good agreements with experimental results (within 5 degree) were consistently observed for several TM helices (results not shown). For helix-helix interactions, a distance-dependant statistical potential describing inter-helical residue interactions was developed. For residue types i and j that are r Å apart, the pair-wise potential is calculated as

$$V_{pw}(i,j,r) = -c\ln\left(\frac{N_{obs}(i,j,r)}{N_{\exp}(i,j,r)}\right)$$

$N_{obs}(i,j,r)$ is the observed contact occurrences in the database, $N_{\exp}(i,j,r)$ is the expected occurrences if an ideal-gas reference state is assumed [6]. The bin size for distance is set at 1 Å and a cutoff distance of 15 Å is employed.

For atomistic simulations, the CHARMM19 force field [7] is employed with the form

$$V = V_\phi + V_{vdw} + V_{ES}$$

where V_ϕ represents dihedral energy terms, and V_{vdw} and V_{ES} are the van der Waals (VDW) and the electrostatic nonbonded energy terms.

Idealized helical structure is assumed for each helical backbone, which is kept fixed during simulations. In the starting configuration, each helix is parallel to the bilayer normal and 15 Å apart with each other. Various Monte Carlo (MC) moves are used to update the packing configuration, including rigid-body translation, rotation by the center of mass, and rotation along the helix axis for each single helix. For atomistic simulation, torsional rotations are added to account for the side-chain flexibility. The orientation of a helix inside a membrane layer (i.e., N-terminal inside or outside of the cell) is included in our simulations as an input to reduce the sampling space. For the calculation of root-mean-square deviation (RMSD) of the predicted structures with respect to the experimental structure, only C_α atoms are included.

3. Results

3.1 Folding of helix bundles using residue-level energy functions

Two aforementioned residue-level energy functions were employed in MC simulations with Wang-Landau algorithm to predict the packing of two- and seven-helix systems. During the Wang-Landau sampling process, 10,000 random configurations were saved. The structure variations among these structures were then subjected to principle component analysis [8]. The energy landscape of the studied system was

then visualized by plotting the system energy using the first and second principal component (PC1 and PC2) as coordinates. Various LEMs, represented in the energy landscape as low-energy basin, could be identified, and the representative structure (the structure with the lowest energy in each basin) for each LEM was selected as possible optimal packing topology.

Shown in Fig. 1 are the prediction results for the packing of the GpA dimer. Plotted in Fig. 1(b) is the average system energy vs. RMSD from the native structure taken from the PDB file 1AFO [9]. The system energy decreases monotonically as the simulated structure becomes more similar to the native structure, down to RMSD=3.6 Å. For structures with RMSD < 3.6 Å, there are some fluctuations in the system energy curve, indicating that our residue-based energy functions are not detailed enough to differentiate native structure from its close structural neighbors. Shown in Fig. 1(a) is the energy landscape of this system plotted in PC1- and PC2-coordinates. Two large LEM regions can be identified, and labeled as A and B in the graph. Among them, LEM B corresponds to a group of structures that are most similar to the native structure, with a RMSD of 0.8 Å , and LEM A is the global energy minimum (GEM). We also simulated the packing of the first and second helices of BR and the neu dimer with similar successful results. For the first two helices of BR., a structure with RMSD 0.9 Å is identified as one of the LEMs. For the whole seven-helix BR, the energy vs. RMSD curve shows a monotonic decrease as RMSD decreases to about 6 Å. Up to 10 LEM regions can be identified from the energy landscape, one of which has structures within 5.0 Å of the native structure. Further selection and refinement of the correcting packing topology can be achieved by atomistic simulations starting from these candidate low-resolution structures.

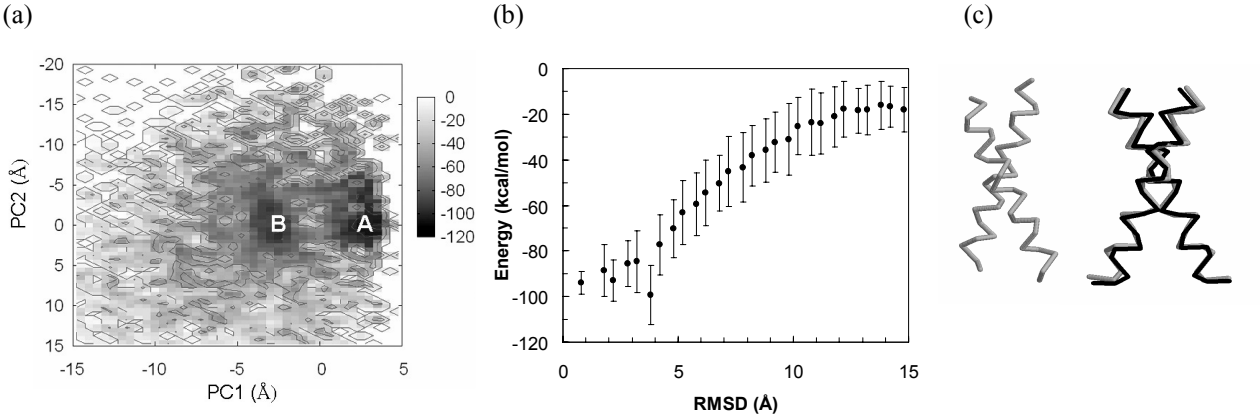

Figure 1. (a) A contour plot of the energy landscape (in units of kcal/mol) as a function of PC1 and PC2. (b) Average system energy as a function of RMSD for the packing of the GpA dimer using residue-level potentials. (c) On the left, the representative structure of global energy minimum A (RMSD=3.6 Å); on the right, the structure corresponding to local minimum B (grey sticks), with a RMSD of 0.8 Å. As a comparison, the native structure is shown by black sticks.

3.2 Helix packing simulations at atomistic level

Atomistic simulations were carried out for the aforementioned systems. Besides the CHARMM19 energy terms, the helix-lipid potential at the residual level is also included to represent the lipid environment implicitly. The resulted potential energy landscape of the GpA dimer is shown in Fig. 2(a). Six distinct LEMs can be identified for the two-helix system, labeled as A~F in the plot. Representative structures of LEMs A and B are shown in Fig. 2(c). LEM A is the GEM with energy of-123.9 kcal/mol, which is at least 1.2 kcal/mol lower than the other LEMs. Its representative structure is almost identical to the experimental one with a RMSD 0.5 Å. In order to compare the stability among those LEMs, it is desired to obtain probability distribution as a function of PC1. For this purpose, a 2D Wang-Landau walk was carried out, where PC1 was employed as the second parameter besides energy. Shown in Fig. 2(b) is a plot of the probability distribution P(PC1) at different temperatures, where $P(PC1) = \sum_E g(E, PC1)\exp(-E/kT)$. Each peak in the plot corresponds to a LEM, and the most stable structure is apparently GEM A with the highest

probability density. At 300K, there are also some LEM B structures, where the contact interface is formed by residues G79 and G83 in one helix and G83 and G86 in the other. For the first two helices of BR, several LEM structures with comparable stability were observed without the helix-lipid potential. After the helix-lipid potential was added in the simulation, the energy landscape changes substantially, and a native-like structure (0.3 Å from the experimental one) was identified as the most stable one (results not shown). Attempts were also made to simulate the packing of the whole seven-helical BR protein. A GEM structure very close to the native structure (RMSD 3.0 Å) was predicted, as shown in Fig. 2(d).

Interestingly, structures of GEM A and LEM C from atomistic simulations are almost identical (RMSD < 0.5 Å) to the two structures associated with LEM B and GEM A from residue-level simulations respectively. The contribution of each residue to the inter-helical VDW interactions from two schemes also follows a similar pattern in the native-like structures, as shown in Fig. 3. Similar correspondences were also observed in other systems we studied.

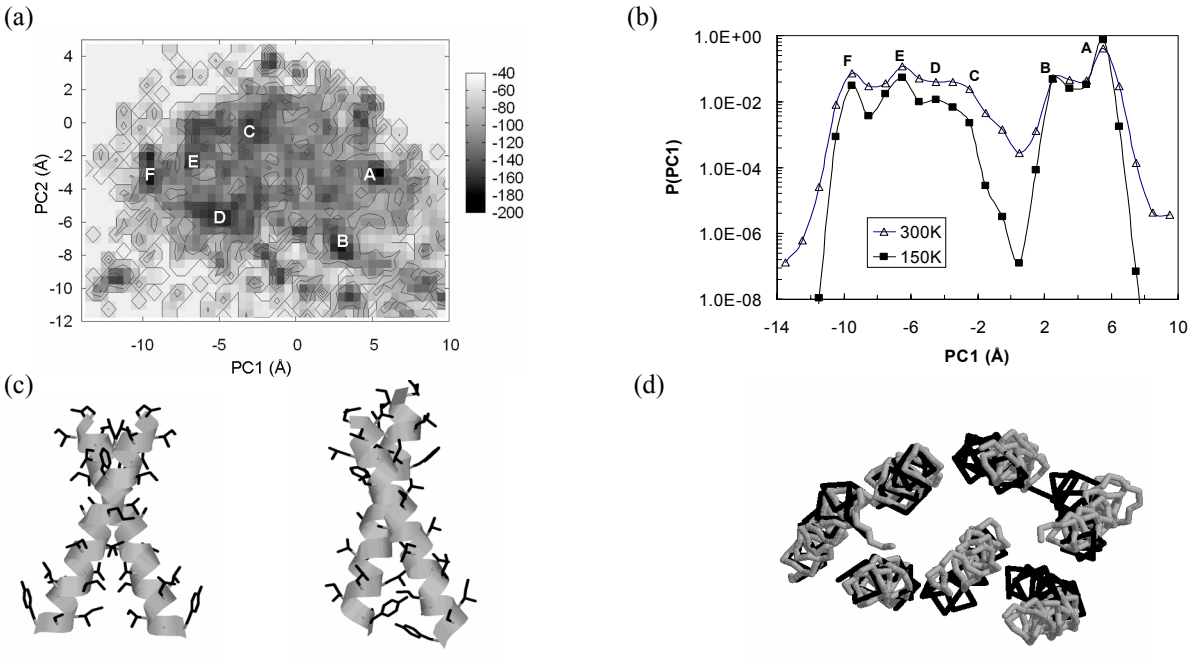

Figure 2. (a) Potential energy Landscape for the GpA dimer from atomistic simulation. (b) The probability distribution as a function of PC1 from a 2D random walk simulation for the GpA dimer. (c) Representative structures of local energy minima A and B for the GpA dimer with RMSD values of 0.5 Å (left) and 3.7 Å (right) respectively. (d) Superimposition of the predicted GEM structure (grey sticks) from atomistic simulation and native structure (black sticks) of BR.

4. Discussion

Two simulation schemes were developed to predict the optimal packing of TM helices at both residual and atomistic levels. At residual level, simulations of the packing of two two-helix systems, namely the GpA dimer and the first and second helices of BR, predicated the native-like structures (within 1 Å) as one of the LEMs. For atomistic simulations, native-like structures were predicted as the most stable structures for all the systems we studied. Comparison of simulation results from two length scales for the same system revealed a substantial correspondence. By coarse-graining all the atomistic details within a residue into an interacting point, the energy landscape is smoothed significantly for residual level simulations, and vast majority of the LEMs found in the atomistic simulations disappear or are combined into larger energy basins. However, the free-energy basin around the native structure seems to be able to maintain its characteristics after coarse-graining. For both two-helix systems we tested, native-like structures emerged as one of the LEMs. It was further observed that the

underlying interactions that stabilize the native structure were reproduced in the coarse-grained simulations in a similar pattern as in the atomistic simulations. This observation suggests that a hierarchical approach to membrane protein structure prediction, where the candidate structures are first selected at the residual level and then refined with atomistic details, is feasible.

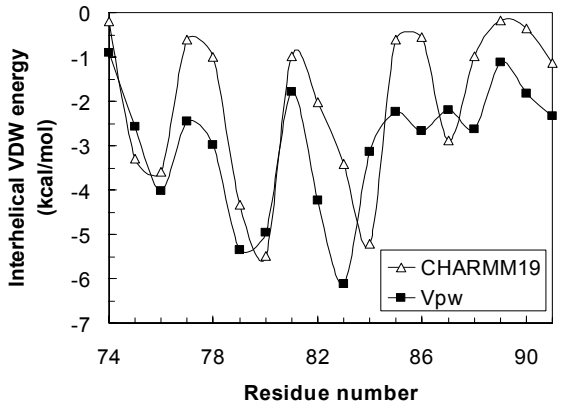

Figure 3. Contributions of residues 74 to 91 in the GpA dimer to Interhelical van der Waals interactions for native-like structures from residue-level and atomistic simulations respectively.

Acknowledgement This research was supported in part by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204), by the Georgia Cancer Coalition and US Department of Energy's GTL program (http://doegenomestolife.org/) under project, "Carbon Sequestration in Synechococcus sp.: From Molecular Machines to Hierarchical Modeling" (www.genomes2life.org).

Reference

[1] Pappu, R.V., G.R. Marshall & J.W. Ponder (1999). "A potential smoothing algorithm accurately predicts transmembrane helix packing", *Nat Struct Biol* **6**:50-55.

[2] Kim, S., A.K. Chamberlain & J.U. Bowie (2003). "A simple method for modeling transmembrane helix oligomers", *J Mol Biol* **329**:831-840.

[3] Kokubo, H. & Y. Okamoto (2004). "Prediction of membrane protein structures by replica-exchange Monte Carlo simulations: case of two helices", *J Chem Phys* **120**:10837-10847.

[4] Wang, F. & D.P. Landau (2001). "Efficient, multiple-range random walk algorithm to calculate the density of states", *Phys Rev Lett* **86**:2050-2053.

[5] Rathore, N., T.A. Knotts & J.J. de Pablo (2003). "Density of states simulations of proteins", *J Chem. Phys.* **118**:4285-4290.

[6] Zhang, C., S. Liu & Y. Zhou (2004). "Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential", *Protein Sci* **13**:391-399.

[7] Neria, E., S. Fischer & M. Karplus (1996). "Simulation of activation free energies in molecular systems", *Journal of Chemical Physics* **105**:1902-1921.

[8] Ichiye T. & M. Karplus (1991). "Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations", *Proteins* **11**:205-217.

[9] MacKenzie, K.R., J.H. Prestegard & D.M. Engelman (1997). "A transmembrane helix dimer: Structure and implications", *Science* **276**:131-133