J Comput Aided Mol Des (2014) 28:99–109
DOI 10.1007/s10822-014-9734-2

# Computational prediction of kink properties of helices in membrane proteins

**T.-L. Mai · C.-M. Chen**

Received: 16 October 2013 / Accepted: 15 February 2014 / Published online: 21 February 2014


**Abstract** We have combined molecular dynamics simulations and fold identification procedures to investigate the structure of 696 kinked and 120 unkinked transmembrane (TM) helices in the PDBTM database. Our main aim of this study is to understand the formation of helical kinks by simulating their quasi-equilibrium heating processes, which might be relevant to the prediction of their structural features. The simulated structural features of these TM helices, including the position and the angle of helical kinks, were analyzed and compared with statistical data from PDBTM. From quasi-equilibrium heating processes of TM helices with four very different relaxation time constants, we found that these processes gave comparable predictions of the structural features of TM helices. Overall, 95 % of our best kink position predictions have an error of no more than two residues and 75 % of our best angle predictions have an error of less than 15°. Various structure assessments have been carried out to assess our predicted models of TM helices in PDBTM. Our results show that, in 696 predicted kinked helices, 70 % have a RMSD less than 2 Å, 71 % have a TM-score greater than 0.5, 69 % have a MaxSub score greater than 0.8, 60 % have a GDT-TS score greater than 85, and 58 % have a GDT-HA score greater than 70. For unkinked helices, our predicted models are also highly consistent with their crystal structure. These results provide strong supports for our assumption that kink formation of TM helices in quasi-equilibrium heating processes is relevant to predicting the structure of TM helices.

**Keywords** Helical kinks · Molecular dynamics simulation · Fold identification · Computational prediction · Quasi-equilibrium heating

## Introduction

Transmembrane (TM) proteins play a key role in a wide variety of biological processes; their functions include cell–cell contact, surface recognition, cytoskeleton contact, signaling, enzymatic activity, or transporting substances across the membrane [1]. The biological functions of TM proteins are strongly related to their three dimensional structures, which can be categorized into three classes: those with a β-barrel structure, those that cross the lipid membrane with a single α-helix, and those that transverse the membrane with an α-helix bundle. Bioinformatic analyses have demonstrated that helix bundles are much more abundant than β-barrels. The structural features of TM helices, such as position and angle of kinks, are dynamic in nature and play important roles in the function of many TM proteins. For example, TM proteins in the same superfamily usually have similar three-dimensional structures but diverse functions stemming from these multiple structural distortions.

Helical kinks could be classified as bends (change in the direction of helix axis without loss of helical character) or disruptions (change in the direction of helix axis and loss of helical character in the kink region) [2, 3]. TM helices usually have many more kinks than helices in water-soluble proteins. Proline residues present in TM helices are known to cause kinks in helices due to steric conflicts with

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-014-9734-2) contains supplementary material, which is available to authorized users.

T.-L. Mai · C.-M. Chen (✉)
Department of Physics, National Taiwan Normal University, 88 Sec. 4 Ting-Chou Rd., Taipei 116, Taiwan
e-mail: cchen@phy.ntnu.edu.tw

the preceding residue and the loss of a backbone hydrogen bond [4]. Glycine is a known helix breaker in soluble proteins because it is very mobile and usually occupies conformations other than the α-helix, due to the lack of $C_\beta$ atoms as steric restraint [5]. It has been found that the loss of helical structure is proportional to the amount of glycine in synthetic polypeptides [6]. Other amino acids, like serine and threonine, could also induce and stabilize kinks in TM helices, mainly due to the additional hydrogen bond between the polar side group of these residues and the peptide carbonyls in the previous turn of the helix [7]. This interaction is similar to the known interaction of OH moieties of water molecules that hydrogen bond the backbone carbonyls of α-helices, which might be accountable for the dished shape of solvent-exposed helices in soluble globular proteins [8, 9].

There have been several statistical analyses of helical kinks in datasets of TM proteins. An early analysis by Riek et al. [10] reported about 26 % of helices being kinked by examining 119 TM helices of 11 TM proteins. Later on, by analyzing 405 TM helices, Hall et al. [9] concluded that 44 % of TM helices are kinked, and one third of them are due to proline. Notably, a web-based Python application, MC-HELAN, was developed to analyze the kink statistics of 842 TM helices in the protein data bank of transmembrane proteins (PDBTM). This study revealed that 64 % of 842 TM helices have kinks and 33 % of these kinks are in proximity to a proline [2]. Nonetheless, our recent analysis of 1,562 TM helices in PDBTM suggested that about 59 % of these helices have at least one kink and 38 % of these kinks are associated with proline in a range of ±4 residues [3]. The effect of proline on helical kinks has been further examined by Yohannan et al. [11] with a set of 39 kinks from 10 TM protein structures to test their evolutionary hypothesis that kinks in TM helices can be traced back to ancestral proline residues. More recent studies suggested that shifting hydrogen bonds may produce flexible TM helices, and explained how evolution has been able to liberally exploit TM helix bending for the optimization of membrane protein structure, function, and dynamics [12].

In addition to the statistical analyses of kinks in TM helices, a more challenging work is to understand the kink formation in TM helices and to predict the structural features of these helical kinks, including kink position and angle. Most of earlier studies focused on predicting the kink position of TM helices. Hall et al. [9] performed molecular dynamics (MD) simulations of isolated helices to predict the position of helical kinks, which reproduced 79 % of the proline kinks and 18 % of the non-proline kinks. This study also suggested that it is easier to predict proline kinks than non-proline kinks in MD simulations, due to steric conflicts of prolines with the preceding residue. Meruelo et al. [13] proposed a neural network approach, TMKink, and achieved a result with sensitivity and specificity of 0.7 and 0.89, respectively. Kneissl et al. [14] suggested the use of string kernels for support vector machines to predict kink positions, which showed about 80 % of all helices can be correctly predicted as kinked or unkinked. However these methods fail to provide an understanding of kink formation and a reliable prediction of helical kink angles. Recently, we proposed an approach to computationally predict the structure of TM helices by combining MD simulations of isolated helices and representative structure identification from numerous simulated decoy structures [3]. This approach was tested by investigating 37 kinked helices in 29 TM proteins and 5 unkinked helices in 5 TM proteins. Most of the tested kinked helices are not associated with proline near the kink position. For unkinked helices, the predicted models are highly consistent with their crystal structure based on four different structural assessments. For kinked helices, the obtained results show an accuracy of 95 % in kink position prediction and an error less than 10° in the angle prediction of 71 % kinked helices. These results suggest that kink properties of TM helices depend mainly on the primary sequence of an isolated helix, instead of the packing dynamics of TM proteins.

The structural information of TM helices is important in predicting the structure of TM proteins. According to the two-stage model of TM protein folding [15], independently stable helices are formed in the lipid membrane first, and the helices interact with each other to form a functional protein in the second stage. Our previous studies based on the two-stage model showed that the lowest-energy structure of TM proteins with slightly kinked helices (e.g. bacteriorhodopsin, halorhodopsin, and sensory rhodopsin II) is consistent with their PDB structure with a root mean square deviation (RMSD) of 1–3 Å [16–18]. However, for TM proteins with substantially kinked helices (such as bovine rhodopsin), the RMSD between predicted structure and the PDB structure could be as large as 5.5 Å when a kinked helix is approximated by an unkinked helix. Since the helical kinks caused by breaking of the backbone hydrogen bonds lead to hinge bending flexibility in these helices, it is important to reproduce these kinks in structural models of TM proteins and to understand their effects on the three-dimensional structure of TM proteins. Therefore, the structural features of TM helices would be very useful in constructing a more general model for the structure prediction of TM proteins. We note that Werner and Church have recently proposed a knowledge-based approach to model TM protein structures with improved kink modeling, which was trained with 102 crystal structures of TM proteins and tested on 14 GPCR proteins [19]. In spite of the limited data set, the predicted structures of GPCRs seem to be reasonably consistent with their crystal structures.

## Methods

### Experimental dataset of TM helices

For the purpose of comparing our computational predictions with the experimental results, we downloaded the crystal structure of TM helices from PDBTM at http://pdbtm.enzim.hu/. For protein pairs with a sequence identity greater than 95 %, only the highest resolution polypeptide chain was considered. Sets of protein sequences (of length between 40 and 10,000) from PDBTM were sifted with a protein sequence culling server, PISCES, by criteria of sequence identity (pairwise sequence identity less than 95 %) and structural quality (X-ray crystal resolution better than 4 Å and the traditional crystallographic R-factor better than 0.35) [2, 20]. On 30 November 2012, for TM helices of length 20–40 amino acids, there were 1,393 helices (448 TM chains, 239 PDB entries) in PDBTM satisfying the above criteria.

To prepare the experimental data (including position and angle) of helical kinks in PDBTM, as demonstrated in Fig. 1, a kinked helix (Fig. 1b) was first extracted from a membrane protein (Fig. 1a) in the database. The identification of kinked helices was made using MC-HELAN [2], a Monte Carlo based algorithm using heuristics to systematically detect and characterize helical kinks. Among the 1,393 helices in PDBTM, there are 808 identified kinked helices and 585 unkinked helices. We note that the reported number of kinked helices includes only those kinks converged to a unique solution in MC-HELAN, and the reported unkinked helices could have small angle kinks (but failed to converge to a unique solution). For simplicity, we have excluded those kinked helices from our study if they contain non-standard amino acids or their kinked residues are found near the two ends of helices. In this study, we have considered 696 kinked helices and 120 unkinked helices. Since some helices have more than one kink, the total number of kinks analyzed is 769. The determination of position and angle of helical kinks was then followed by calculating the intersection of two helical axes of a kinked helix as shown in Fig. 1b. In the definition of this study, a residue is considered to be found in the kinked position if it has no intra-helix hydrogen bonding.

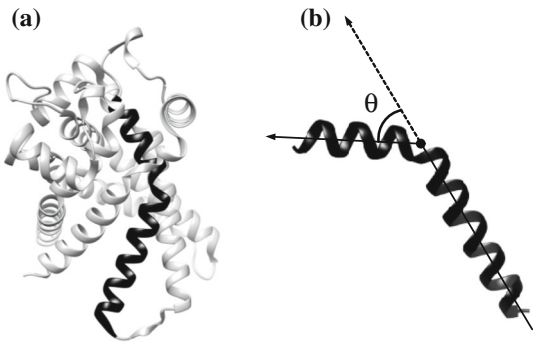


**Fig. 1** Ribbon diagrams of a TM protein (**a**) and a kinked helix extracted from the protein (**b**). The kink position (*filled circle*) and kink angle (θ) of the kinked helix are schematically illustrated in **b**

The distribution of amino acids in the kink region was examined by calculating their normalized occurrence probability in a range of ±4 residues away from the kink center. For simplicity, here we classified 20 amino acids into five groups, including charged residues (R, H, K, D, E), polar residues (S, T, N, Q, C), helix breaking residues (G, P), aromatic residues (F, Y, W), and other hydrophobic residues (A, V, I, L, M). The normalized occurrence probability of type $i$ group to be found at a particular position away from the kink center is defined as $N_i M_i^{-1}/(\sum_i N_i M_i^{-1})$, where $N_i$ is the occurrence frequency of group $i$ residues at that position and $M_i$ is its frequency to appear in kinked helices.

### Computational methods for helical kink prediction

For the computational prediction of helical kinks, we have simulated the structure of 696 kinked helices in PDBTM. In addition, we have simulated the structure of 120 unkinked helices to further validate our method for the structure prediction of unkinked helices. In our method, the selected segment sequences are initially constructed as standard helices using the program Ribosome (see http://roselab.jhu.edu/~raj/Manuals/ribosome.html). The structural refinement and simulations of TM helices were performed using Amber 11 [21] with the force field set leaprc.ff03.r1, and the decoy structures of helices from their MD simulations were analyzed by SPICKER [22]. A more detailed description of our computational methods will be given in the following paragraphs.

The charge of each TM helix was first neutralized to reduce the influence of electrostatic interaction on the formation of kinks [9]. The constructed helix was then refined by an energy minimization with 600 steps of steep descent method and 600 steps of conjugate gradient method. For simplicity, our simulations of the entire helical segment of an isolated helix (including helical sections in the membrane core, head group, and water regions) were carried out in a uniform background medium of dielectric constant 2.5 [16, 23]. A more realistic dielectric constant of the background is 2.5 in the hydrocarbon core, 10 near the ester group, 30 near the head group-water interface, and 80 in the water region [3]. Therefore, our predictions would only be reliable if those helical kinks are found in the core region. We note that, as a first order approximation, our simulation of TM protein folding in the simplified lipid environment has been shown to be able to predict the

folded structure [16, 17] and folding dynamics [24, 25] of several TM proteins, as well as the kink properties of 42 TM helices [3].

The configuration space of the helix in equilibrium was further investigated by MD simulations with Amber. In this study, four simulation scenarios with different thermostat relaxation time constants were tested to predict the structural features of TM helices. In scenarios (a)–(d), we carried out a quasi-equilibrium heating of the system with the Berendsen thermostat from 0 to 300 K within 20 ns, in which the heating process was uniformly divided into $2 \times 10^7$ time slots [26]. In each time slot, the system temperature was slightly increased by $1.5 \times 10^{-5}$ K and remained stable at the temperature for the rest of time. For a comparison of various different heating rates, we chose the value of the thermostat relaxation time ($\tau_{TR}$) to be 0.01 ps in scenario (a), 0.5 ps in scenario (b), 5 ps in scenario (c), and 1,000 ps in scenario (d). Here the Berendsen thermostat relaxation time, $\tau_{TR}$, determines how tightly the heat bath and the system are coupled together, and its value varies drastically in the 4 simulation scenarios considered. Typical values of $\tau_{TR}$ lie between 0.5 and 5 ps. In the case of a large $\tau_{TR}$ [scenario (d)], the system is weakly coupled to the bath, and the run is approximately sampling a micro-canonical ensemble.

According to Anfinsen's hypothesis, the native state of a protein lies at the global minimum in its free energy. In other words, the native state of a protein is an ensemble of many similar conformations with a low energy, and the target of protein structure prediction could be a representative structure or an average structure of this ensemble. In our study, we used the SPICKER fold identification algorithm to find their most representative structure from a large number of decoy structures obtained in our simulations. SPICKER was developed to identify a near-native structure of a protein from its simulation trajectory [22]. The generality of SPICKER has been assessed by analyzing 1,489 representative benchmark proteins that cover the PDB at the level of 35 % sequence identity, and the difference from the best individual decoy to native is below 1 Å in RMSD for 78 % proteins tested.

In our analyses, the decoy structures of a protein were taken from its MD trajectories obtained in simulation scenarios (a)–(d). Due to computer memory limitation, the number of decoy structures in SPICKER was limited to $10^4$, and it is crucial to consider mainly those structures possessing generic features of a kinked helix. Our previous study found that a kinked helix usually has 1–2 broken backbone hydrogen bonds with the corresponding N–O distance in the range of 4.2–8.7 Å and the distribution of

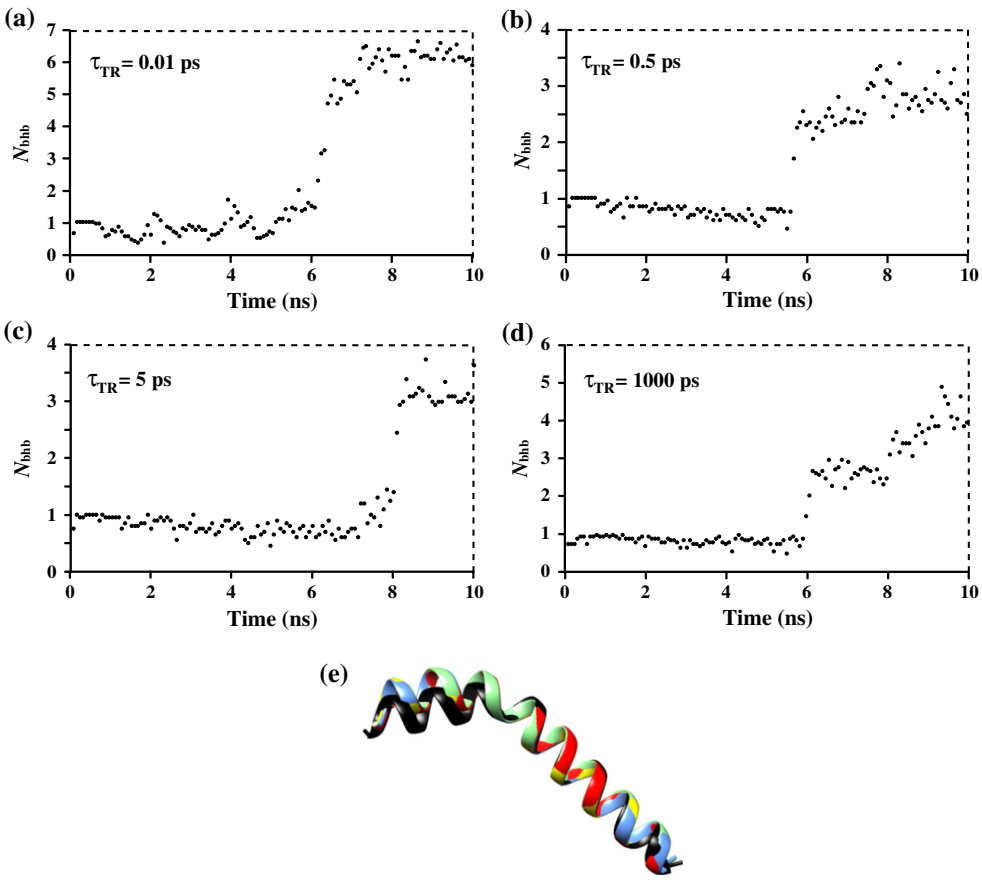


**Fig. 2** Number of broken backbone hydrogen bonds with N–O distance greater than 4.9 Å during the simulation of a kinked helix,1OKC-A:208-239, for four quasi-equilibrium heating processes with $\tau_{TR} = 0.01$ ps (**a**), 0.5 ps (**b**), 5 ps (**c**), and 1,000 ps (**d**). The structure overlap in **e** contains the PDB structure (*black*) of this helix and its predicted models from the heating processes with $\tau_{TR} = 0.01$ ps (*red*), 0.5 ps (*blue*), 5 ps (*yellow*), and 1,000 ps (*green*)

N–O distance is sharply peaked at 4.9 Å, followed by an exponential decay with increasing distance. Therefore, in choosing the decoy structures for SPICKER, we only considered helical structures with $N_{bhb} \leq 2$, where $N_{bhb}$ stands for the average number of broken hydrogen bonds with N–O distance greater than 4.9 Å. For example, in Fig. 2, we show the value of $N_{bhb}$ as a function of time during the four heating processes of the helix 1okc-A:208-239 for $\tau_{TR} = 0.01$ ps (a), 0.5 ps (b), 5 ps (c), and 1,000 ps (d). The decoy structures for SPICKER were taken from the time period of 0–6.1 ns simulation for $\tau_{TR} = 0.01$ ps, 0–5.8 ns simulation for $\tau_{TR} = 0.5$ ps, 0–8.0 ns simulation for $\tau_{TR} = 5$ ps, and 0–6.0 ns simulations for $\tau_{TR} = 1{,}000$ ps. Furthermore, the pairwise RMSD cutoff $R_{cut}$ (under which two structures are considered as clustered neighbors) in SPICKER was initially set to 7.5 Å, and then iteratively changed based on the interplay of the cutoff and the ratio of number of decoys in the most populated cluster to the total number of decoy structures. A flow chart of the SPICKER algorithm is available in Ref. [3]. In Fig. 2e, we overlapped the PDB structure of helix 1okc-A:208-239 with its predicted models from MD simulations in the above four scenarios. The consistency among the PDB structure and our predicted models suggests that the trajectory of kink formation during a quasi-equilibrium heating of an isolated helix does provide a useful description of the kink properties of TM helices.

A list of the helical features of our predicted models of 1okc-A:208-239 is presented in the supplementary Table S1, which shows a correctly predicted kink position, 7°–10° errors in the kink angle prediction, and five quantitative assessments of our predicted helical structures. These assessments include root mean square deviation (RMSD), Template Modeling score (TM-score), MaxSub score, global distance test-total score (GDT-TS), and global distance test with high accuracy (GDT-HA) [27]. Both TM-score and MaxSub score range from 0 to 1, while the GDT-TS/HA scores have a value of 0–100. In general, for high quality structure predictions, the assessment value is greater than 0.5 for TM-score, 0.8 for MaxSub score, 85 for GDT-TS score, and 70 for GDT-HA score.

## Results and discussion

Previously we have conducted a preliminary test of predicting the structural features of 42 TM helices (37 kinked and 5 unkinked) using decoy structures generated from four simulation scenarios, including three quasi-equilibrium heating processes as described in scenarios (b)–(d) and one physical MD simulation process of TM helices with Langevin dynamics at 300 K [3]. In that test, simulations with the physical MD process and three quasi-equilibrium heating processes gave very similar accuracy in predicted structural features of TM helices, suggesting that the trajectory of kink formation during the quasi-equilibrium heating process can be used to predict position and angle of helical kinks. In the present study, we investigated a more complete set of TM helices (696 kinked and 120 unkinked) in PDBTM using four quasi-equilibrium heating processes with $\tau_{TR} = 0.01$ ps (a), 0.5 ps (b), 5 ps (c), and 1,000 ps (d). It is found that predicted kink positions of helices are mostly correct and predicted kink angles of helices are roughly consistent with their experimental value. The comparison of kink properties of 8 kinked helices in PDBTM with those of their predicted models from the four simulation scenarios is presented in the supplementary Table S2, and a complete list of our predicted kink properties for 696 kinked helices can be found in the supplementary Table S3. In this study, we have used 95 % pairwise sequence identity as a threshold for redundancy for inclusion in the dataset. Redundancy in a dataset occurs when several similar/homologous sequences are present in the same set of data, which will introduce undesirable biases in statistical analyses and also increase computational cost. Due to the efficiency of our method and for the purpose of comparing our computational predictions with their experimental values, we used this high cutoff to remove direct replications but keep as many original structures as possible.

For a comparison of the differences among our structure predictions from the four quasi-equilibrium heating processes considered, we present our predicted kink properties of 696 kinked helices in Fig. 3: Fig. 3a shows the cumulative distribution of position shift of our predicted kinks from their actual site calculated from experimental structures in PDBTM, and Fig. 3b shows the cumulative distribution of deviation of our predicted kink angles to their value derived from PDBTM. Here the cumulative distribution function of a real-valued variable X is defined as $F_X(x) = P(X \leq x)$, where the right-hand side represents the probability that the variable X takes on a value less than or equal to $x$. The best prediction in Fig. 3 is defined as the best structural model predicted from the four simulation scenarios considered. For kink position prediction, our results demonstrate about 60–70 % accuracy for these four scenarios (64.7 % for $\tau_{TR} = 0.01$ ps, 66.7 % for $\tau_{TR} = 0.5$ ps, 63.6 % for $\tau_{TR} = 5$ ps, and 59.1 % for $\tau_{TR} = 1{,}000$ ps) and 88 % accuracy for our best prediction. Overall, 95 % of our best predictions in kink position have a shift of no more than 2 residues. For our best kink angle predictions in the 615 kinked helix models with correctly predicted kink position, 60.2 % of our predictions have an error less than 10° and 73.3 % of our predictions have an error less than 15°. It is found that the simulation scenario with $\tau_{TR} = 0.5$ ps usually gives a better

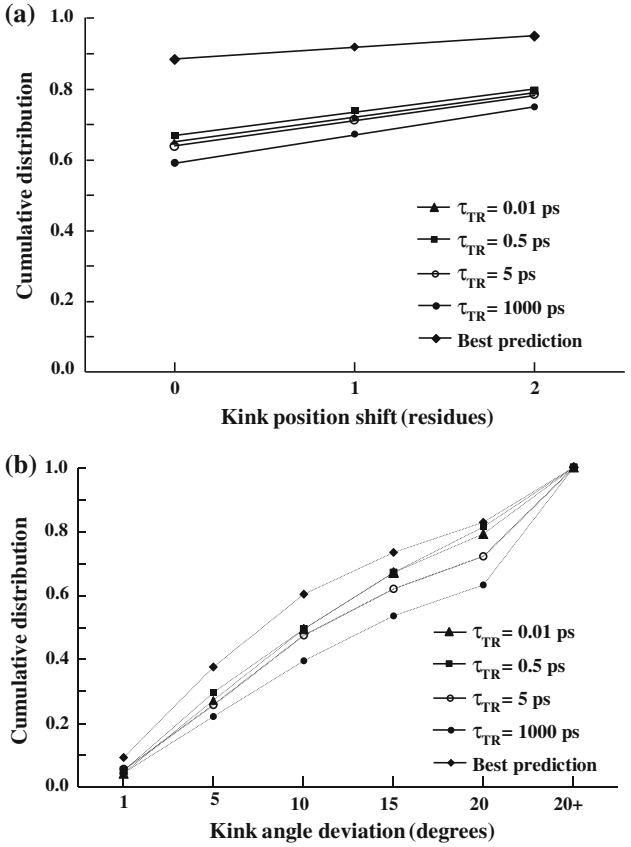


**Fig. 3** Cumulative distributions of kink position shift (**a**) and kink angle deviation (**b**) in our predicted models of 696 kinked helices in PDBTM. Quasi-equilibrium heating processes of helix folding were performed with four different thermal relaxation time constants, $\tau_{TR} = 0.01$, 0.5, 5, and 1,000 ps. The best prediction is defined as the best structural model predicted from these four simulation scenarios

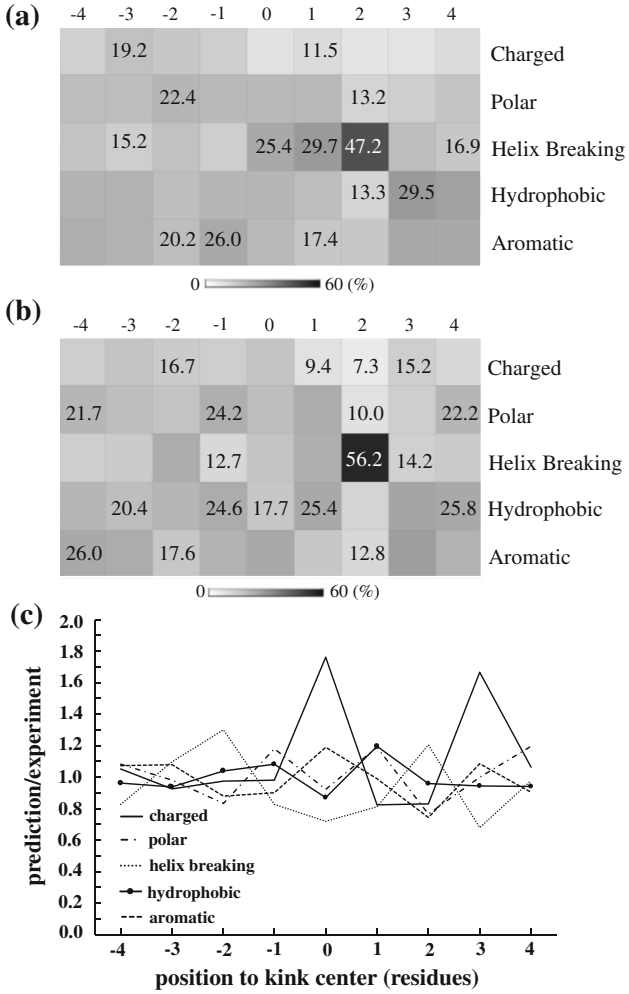


**Fig. 4** Normalized occurrence probability of five residue groups (*vertical axis*) in the range of ±4 residues away from the kink center (*horizontal axis*) that was calculated from the experimental structures in PDBTM (**a**) or from their best predicted models (**b**). A *white* to *black grayscale bar* calibrates the probability from 0 to 60 %. The curves in **c** are the ratio of the predicted probability in **b** to the experimental probability in **a** for five groups

prediction in kink properties, but their differences are not significant. Compared with results of the $\tau_{TR} = 0.5$ ps simulations, the defined best prediction leads to a 15 % improvement in the accuracy of kink position prediction and a 5 % improvement in the accuracy of kink angle prediction. Such a trade-off between accuracy and computational cost seems practical for our prediction method, which is free of parameters and does not rely on the input of experimental structures. Compared with results of a previous study in Ref. [3], the results of this study using solely quasi-equilibrium heating processes seem to be less accurate in predicting the position and angle of kinks for a more complete data set. Physical MD simulations of isolate TM helices for this larger data set are under our current investigation. We note that, the kink position of TM helices is more accurately predicted than the kink angle. This observation might reflect the fact that the kink position of TM helices is usually not affected by the packing of helices, while the kink angle is more sensitive to considerations of the packing of helices.

In Fig. 4, we further examined the distribution of amino acids in the predicted kink region. Figure 4a shows the normalized occurrence probability of the five groups of residues from the experimental structures of helices in PDBTM, Fig. 4b shows our computational prediction of the normalized probability, and Fig. 4c compares the above two normalized probabilities by taking the ratio of the predicted probability to the experimental probability. Comparing Fig. 4a with b, our predictions are in general consistent with the experimental data: (1) there is a high prevalence of helix breaking residues at the +2 position (mainly due to proline); (2) charged residues are less likely to appear in the kink region; and (3) the normalized occurrence probability of aromatic residues has roughly similar pattern to that of other hydrophobic residues. The

consistency between our predictions and the experimental statistics can also been seen in Fig. 4c. The ratio of our predicted probability to the experimental one is mostly within the range of 0.7 and 1.3 for all groups and all positions. However, our predictions overestimated the normalized occurrence probability of charged groups at the positions of 0 and +3. A detailed comparison of the normalized probability of 20 amino acids between our prediction and the associated PDBTM statistics is available in the supplementary information (Figure S1). We note that our supplementary information shows a significant difference in the probability distributions of proline and glycine. As discussed in the introduction, the probability distribution of glycine has a small peak at the kink center and is roughly uniform in other positions, due to the lack of steric constraints from its side chain in maintaining a helical structure. On the other hand, the probability distribution of proline has a sharp peak at the +2 position since the steric conflicts of its side chain lead to a broken ($i$th, $i + 4$th) backbone hydrogen bond which distorts the helix and is further stabilized over the next two residues.

In addition to study the occurrence frequency of residue types in the kink region, it is also important to validate our prediction in the distribution of kink angles. In Fig. 5a, we compared the distributions of kink angles for 769 kinks in PDBTM that were calculated from our best predicted model of TM helices and from their experimental structures. In Fig. 5b, c, we provided similar comparisons for kinked helices that are associated with proline or glycine in the range of ±4 residues near the kink center. The solid curves in Fig. 5 show fitted distribution functions for the distribution data of calculated or predicted kink angles. More specifically, the data in Fig. 5a and c were fitted with the lognormal distribution, while the data in Fig. 5b were fitted with the Dagum distribution (type I). The angular distributions of helical kinks calculated from our predicted models and from experimental structures are highly consistent with each other for comparisons in Fig. 5a and c, but have a noticeable discrepancy for the comparison in Fig. 5b. For all helical kinks, the angular distributions peak at 10°–15°. It is also observed that the distributions of proline associated kinks peak at 20°–25°, which is larger than the peak angle of 10°–15° in glycine associated kinks. For most proline associated kinks, we found that proline initiates these kinks by breaking the ($i$th, $i + 4$th) bond and the breaking of the helical backbone bond continues to distort and is stabilized in the next two residues. A combination of small polar residues and the proline residue often results in larger kink angles. In these cases, the kink center is often two residues away from proline, as demonstrated in Fig. 4. This result is also consistent with the observation of Hall et al. [9]. Moreover, a finite portion of helical kinks are found to have very large kink angles

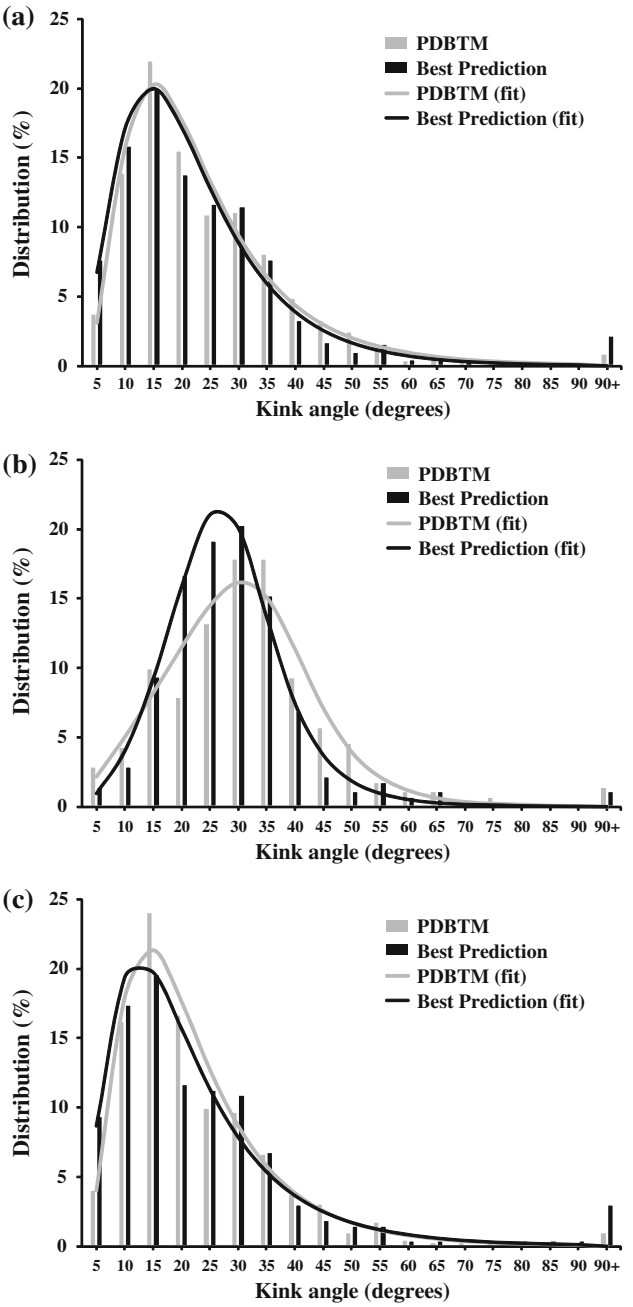


**Fig. 5** A comparison of the distributions of kink angles that were calculated from the experimental structures in PDBTM or from our best predicted models. The distributions in **a** were calculated for all 769 kinks identified, in **b** were calculated for proline associated kinks, and in (**c**) were calculated for glycine associated kinks. The histogram bin size is 5°. *Solid curves* in **a** and **c** are fitted lognormal distributions, and in **b** are fitted Dagum distributions. The values of $R^2$ (coefficient of determination) for these *fitted curves* are 0.97 (PDBTM) and 0.96 (Best) in **a**, 0.94 (PDBTM) and 0.99 (Best) in **b**, and 0.97 (PDBTM) and 0.95 (Best) in **c**

(>90°), but our predictions tend to overestimate the amount of non-proline kinks with large angles. For the test of predicting unkinked helices in our method, in Fig. 6, we

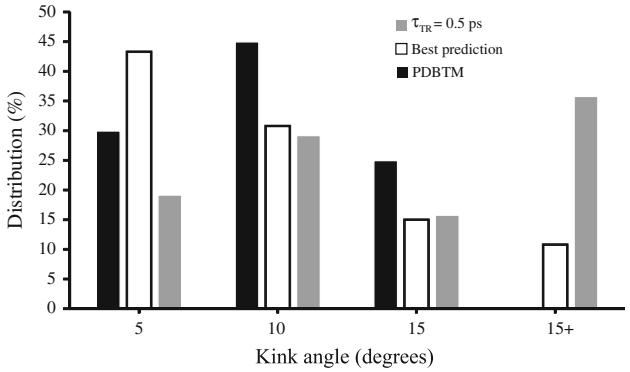


**Fig. 6** A comparison of the distribution of kink angles for "unkinked" helices that were calculated from the experimental structures or from our predicted models. The histogram bin size is 5°

listed the calculated kink angle distribution of these helices. As noted in section "Experimental dataset of TM helices", the reported "unkinked" helices in PDBTM could have small angle kinks but failed to converge to a unique solution in MC-HELAN. According to our kink angle definition in section "Experimental dataset of TM helices", we found that these "unkinked" helices in PDBTM does have small kink angles. Our best prediction in the kink angle distribution is found to be consistent with that of PDBTM, while the prediction solely using $\tau_{TR} = 0.5$ ps simulations does have a much higher percentage (35 %) for kinks with an angle greater than 15°.

To further test our method, various quantitative structure assessments, including RMSD, TM-score, MaxSub, GDT-TS, and GDT-HA, have been applied to verify the accuracy of our predicted models for 696 kinked helices and 120 unkinked helices. The RMSD of two aligned structures indicates their divergence from one another, which is sensitive to outlier regions created by poor modeling of individual loop regions in a structure that is otherwise reasonably accurate. The MaxSub test aims at identifying the maximum subset of $C_\alpha$ atoms of a model that superimposes well over the experimental structure, and produces a single normalized score that represents the quality of the model. The GDT-TS/HA tests measure the similarity between the predicted model and the experimental structure by calculating the largest set of amino acid residues' $C_\alpha$ atoms in the model structure falling within a defined distance cutoff of their position in the experimental structure. TM-score assesses the accuracy of protein structure predictions by weighting the close atom pair stronger than the distant matches. In general, TM, MaxSub and GDT scores are more sensitive to the topology fold than RMSD. A complete list of various assessment scores is available in the supplementary Tables S4 (kinked helices) and S5 (unkinked helices). For simplicity, 2 unkinked helices [1u19-A:149-169 in (a) and 2zy9-A:245-264 in (b)] and 3

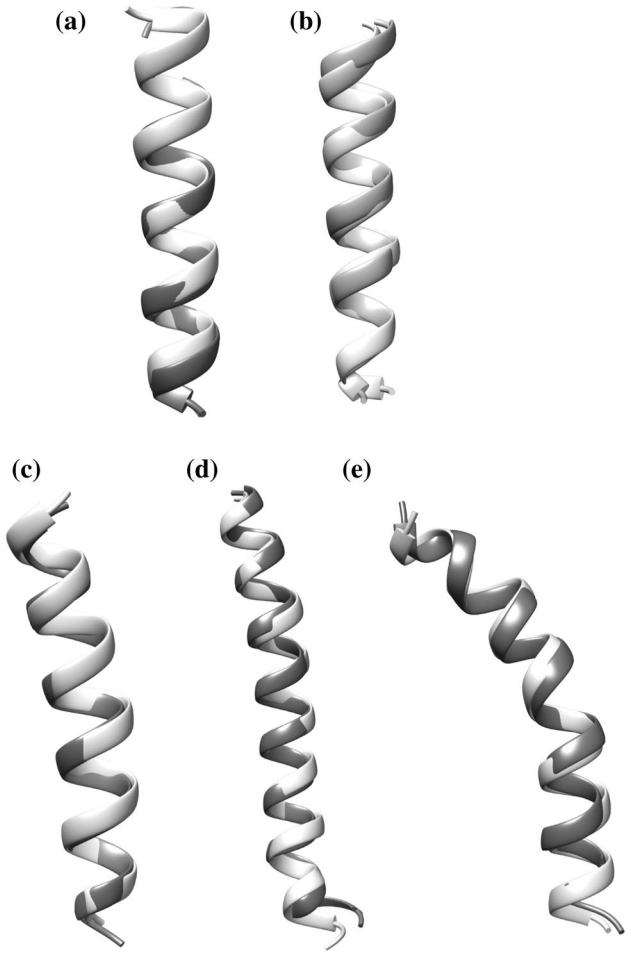


**Fig. 7** A comparison of the PDB structure and our predicted model for two unkinked helices (1u19-A:149-169 in **a** and 2zy9-A:245-264 in **b**) and three kinked helices (2bhw-A:123-144 in **c**, 3org-A:320-351 in **d**, and 3tui-A:88-113 in **e**). Here the PDB structures are *colored in light gray*, and our best predicted models are *colored in dark gray*

kinked helices [2bhw-A:123-144 in (c), 3org-A:320-351 in (d), and 3tui-A:88-113 in (e)] were randomly selected for discussion, and a comparison of their PDB structures and our predicted models were shown in Fig. 7. The overlapping structures in Fig. 7 demonstrate the good quality of our best predicted models of these five helices, whose RMSD ranges from 0.70 to 1.14 Å. The other four assessments confirm that these predicted models do have high quality in their structure. The cumulative distributions of RMSD in our predictions for 696 kinked helices (solid line) and 120 unkinked helices (dash line) are presented in Fig. 8a, and the cumulative distributions of TM-score, MaxSub score, GDT-TS score, and GDT-HA score for 696 kinked helices are shown in Fig. 8b. In Fig. 8a, the percentage of our predicted models with a RMSD less than 2Å is about 70 % for kinked helices, and is 95 % for unkinked helices. In Fig. 8b, 71 % of kinked helices have a TM-score greater than 0.5, 69 % have a MaxSub score greater

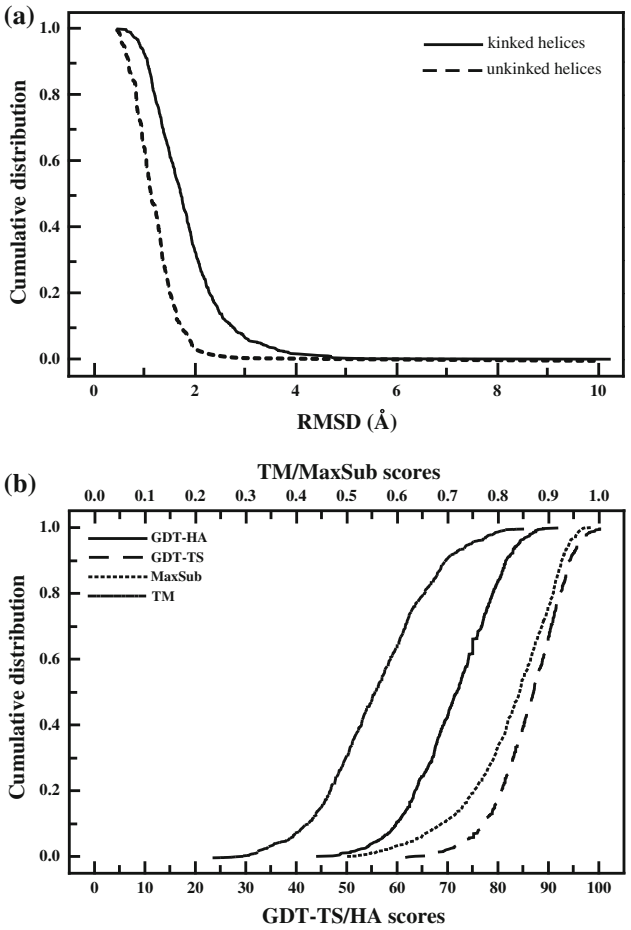


**Fig. 8** Cumulative distributions of RMSD in predictions of 696 kinked helices (*solid line*) and of 120 unkinked helices (*dash line*) are presented in **a**, and the cumulative distributions of TM-score, MaxSub score, GDT-TS score, and GDT-HA score for 696 kinked helices are shown in **b**. Notice that the cumulative distribution in **a** is accumulated from large RMSD toward small RMSD

than 0.8, 60 % have a GDT-TS score greater than 85, and 58 % have a GDT-HA score greater than 70. Here we have adopted criteria for high quality structures in various quantitative structure assessments. A further verification of our method on 14 unkinked helices is provided in Table 1, and the predictions of 120 unkinked helices are listed in the supplementary Table S5. These unkinked helices were also simulated for all four scenarios and their best predicted models were used to calculate the score for various structure assessments. The high scores from various assessments for these unkinked helices also strongly back up our approach for predicting the structure of TM helices.

## Conclusion

In this study, we have integrated MD simulations and a fold identification algorithm to investigate kink formation of TM helices for which an experimental structure was available in the PDBTM. Quasi-equilibrium heating processes of an isolated TM helix with very different thermal relaxation time constants have been simulated using AMBER. From these simulations, decoy structures possessing generic features of a kinked helix were further analyzed to find the most representative model using SPICKER. The aim of this study is to investigate the relevancy of predicted structures of TM helices from these quasi-equilibrium heating processes to their experimental structures in PDBTM. Our results showed an accuracy of 88 % in predicting the kink position of TM helices (no kink position shift), and an error less than 15° (10°) in the angle prediction of 73 % (60 %) of kinked helices. In addition, by examining the distribution of amino acids in the predicted kink region, we found a consistent pattern in the normalized occurrence probability of five residues groups (charged, polar, helix breaking, hydrophobic, and aromatic groups) between our prediction and the experimental data in PDBTM. For the distribution of kink angles in TM helices, our predicted distribution profile, including shape and peak position, resembled the profile calculated from the experimental structures.

We have also performed various quantitative structure assessments of our predicted models for 696 kinked helices and 120 unkinked helices in PDBTM. For 696 kinked helices, 70 % of our predicted models have a RMSD less than 2 Å, 71 % have a TM-score greater than 0.5, 69 % have a MaxSub score greater than 0.8, 60 % have a GDT-TS score greater than 85, and 58 % have a GDT-HA score greater than 70. These assessment results confirmed the validity of our integrated approach to predict the structure of TM helices. More specifically, this study suggest that the kink properties of TM helices depend mainly on the primary sequence of an isolated helix, instead of the dynamics of folding or the packing of TM helices. We believe these results provide some evidence for the two-stage model of TM proteins: Independently stable helices are formed in lipid membrane in the first stage, and the helices interact with others to form a functional MP in the second stage. However, as suggested by the observed error in our predicted kink angles, we note that the kink angle of TM helices could still be affected by their packing in the second stage.

To conclude, this study has achieved several unique objectives in studying kink properties of TM helices. First, the helix database in PDBTM we analyzed in this study is the largest and most up to date. Second, instead of constructing another tool for the statistical analysis of the helical features of experimental structures in PDBTM, we provide an integrated method based on thermodynamic principles to investigate the kink formation of TM helices for which an experimental structure was available in the

**Table 1** Kink angle and five structure assessments of our best predicted models for 14 unkinked helices randomly selected from PDBTM

| TM helices | Assessments | | | | | Angle (°) | |
|---|---|---|---|---|---|---|---|
| | RMSD (Å) | TM score | MaxSub score | GDT-TS score | GDT-HA score | PDBTM | BEST |
| 1bcc-F:52-71 | 1.1 | 0.46 | 0.92 | 91.3 | 78.8 | 3.3 | 10.6 |
| 1ppj-F:51-71 | 1.8 | 0.54 | 0.91 | 95.2 | 86.9 | 1.3 | 25.9 |
| 1q90-B:35-55 | 0.7 | 0.51 | 0.96 | 97.6 | 86.9 | 7.1 | 4.1 |
| 1u19-A:149-169 | 0.7 | 0.52 | 0.96 | 97.6 | 88.1 | 5.7 | 2.2 |
| 2bs2-C:76-98 | 1.0 | 0.54 | 0.94 | 94.6 | 83.7 | 9.5 | 3.8 |
| 2gsm-A:91-110 | 1.4 | 0.53 | 0.89 | 92.5 | 83.8 | 13.2 | 11.4 |
| 2wjn-H:11-31 | 0.9 | 0.45 | 0.94 | 96.4 | 85.7 | 6.8 | 9.5 |
| 2zy9-A:245-264 | 1.0 | 0.51 | 0.94 | 95.0 | 83.8 | 7.6 | 11.6 |
| 3ag3-D:76-103 | 0.9 | 0.73 | 0.95 | 95.5 | 83.9 | 1.7 | 8.1 |
| 3cx5-A:133-155 | 1.5 | 0.55 | 0.88 | 90.2 | 78.3 | 6.3 | 11.0 |
| 3cx5-E:50-81 | 0.5 | 0.88 | 0.98 | 99.2 | 90.6 | 4.3 | 2.8 |
| 3l70-F:51-72 | 1.3 | 0.53 | 0.91 | 93.2 | 81.8 | 4.1 | 13.6 |
| 3sn6-B:2-25 | 1.3 | 0.66 | 0.91 | 94.8 | 84.4 | 0.7 | 9.3 |
| 4dx5-A:8-30 | 1.3 | 0.64 | 0.91 | 94.6 | 88.0 | 4.3 | 6.5 |

Here unkinked helices refer to those helices without a converged kink angle in MC-HELAN

PDBTM. Third, this method yields high quality models of TM helices and the statistical properties of helical kinks in our predicted models are consistent with those calculated from experimental structures in PDBTM. Finally, this study provides an evidence for the validity of the two-stage model of TM proteins. To go beyond our present results and as a further validation of the two-stage model, we would like to extend our study to the packing of TM helices by combining all-atom MD simulations with coarse-grained Monte-Carlo simulations. Successful prediction of the three dimensional structure of TM proteins would be helpful for us to understand their biological functions and to design feasible pharmaceutical applications.

**Acknowledgments** This work is supported by the National Science Council of Taiwan under grant of no. NSC 99-2112-M-003 -011 -MY3. We thank D.N. Langelaan for providing the MC-HELAN algorithm, and Y.-H. Huang for stimulating discussion.

## References

1. White SH, Wimley WC (1999) Membrane protein folding and stability: physical principles. Annu Rev Biophys Biomol Struct 28:319–365
2. Langelaan DN, Wieczorek M, Blouin C, Rainey JK (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. J Chem Inf Model 50(12):2213–2220. doi:10.1021/ci100324n
3. Huang YH, Chen CM (2012) Statistical analyses and computational prediction of helical kinks in membrane proteins. J Comput-Aided Mol Des 26(10):1171–1185. doi:10.1007/s10822-012-9607-5
4. von Heijne G (1991) Proline kinks in transmembrane α-helices. J Mol Biol 218(3):499–503. doi:10.1016/0022-2836(91)90695-3
5. Jacob J, Duclohier H, Cafiso DS (1999) The role of proline and glycine in determining the backbone flexibility of a channel-forming peptide. Biophys J 76(3):1367–1376. doi:10.1016/S0006-3495(99)77298-X
6. Chakrabartty A, Baldwin RL (1995) Stability of alpha-helices. Adv Protein Chem 46:141–176
7. Costantini S, Colonna G, Facchiano AM (2006) Amino acid propensities for secondary structures are influenced by the protein structural class. Biochem Biophys Res Commun 342(2):441–451. doi:10.1016/j.bbrc.2006.01.159
8. Gray TM, Matthews BW (1984) Intrahelical hydrogen bonding of serine, threonine and cysteine residues within α-helices and its relevance to membrane-bound proteins. J Mol Biol 175(1):75–81. doi:10.1016/0022-2836(84)90446-7
9. Hall SE, Roberts K, Vaidehi N (2009) Position of helical kinks in membrane protein crystal structures and the accuracy of computational prediction. J Mol Graph Model 27(8):944–950. doi:10.1016/j.jmgm.2009.02.004
10. Riek RP, Rigoutsos I, Novotny J, Graham RM (2001) Non-alpha-helical elements modulate polytopic membrane protein architecture. J Mol Biol 306(2):349–362. doi:10.1006/jmbi.2000.4402
11. Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA 101(4):959–963. doi:10.1073/pnas.0306077101
12. Cao Z, Bowie JU (2012) Shifting hydrogen bonds may produce flexible transmembrane helices. Proc Natl Acad Sci USA 109(21):8121–8126. doi:10.1073/pnas.1201298109

13. Meruelo AD, Samish I, Bowie JU (2011) TMKink: a method to predict transmembrane helix kinks. Protein Sci 20(7):1256–1264. doi:10.1002/pro.653
14. Kneissl B, Mueller SC, Tautermann CS, Hildebrandt A (2011) String kernels and high-quality data set for improved prediction of kinked helices in alpha-helical membrane proteins. J Chem Inf Model 51(11):3017–3025. doi:10.1021/ci200278w
15. Popot JL, Engelman DM (1990) Membrane protein folding and oligomerization: the two-stage model. Biochemistry 29(17): 4031–4037
16. Chen CC, Chen CM (2009) A dual-scale approach toward structure prediction of retinal proteins. J Struct Biol 165(1): 37–46. doi:10.1016/j.jsb.2008.10.001
17. Chen CC, Wei CC, Sun YC, Chen CM (2008) Packing of transmembrane helices in bacteriorhodopsin folding: structure and thermodynamics. J Struct Biol 162(2):237–247. doi:10.1016/j.jsb.2008.01.003
18. Wu HH, Chen CC, Chen CM (2012) Replica exchange Monte-Carlo simulations of helix bundle membrane proteins: rotational parameters of helices. J Comput Aided Mol Des 26(3):363–374. doi:10.1007/s10822-012-9562-1
19. Werner T, Church WB (2013) Kink characterization and modeling in transmembrane protein structures. J Chem Inf Model 53(11):2926–2936. doi:10.1021/ci400236s
20. Wang G, Dunbrack RL Jr (2003) PISCES: a protein sequence culling server. Bioinformatics 19(12):1589–1591
21. Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16):1668–1688. doi:10.1002/jcc.20290
22. Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 25(6): 865–871. doi:10.1002/jcc.20011
23. Tsong TY (1990) Electrical modulation of membrane proteins: enforced conformational oscillations and biological energy and signal transductions. Annu Rev Biophys Biophys Chem 19: 83–106. doi:10.1146/annurev.bb.19.060190.000503
24. Chen CM (2001) Lattice model of transmembrane polypeptide folding. Phys Rev E 63(1):010901. doi:10.1103/PhysRevE.63.010901
25. Chen CM, Chen CC (2003) Computer simulations of membrane protein folding: structure and dynamics. Biophys J 84(3):1902–1908
26. Hunenberger P (2005) Thermostat algorithms for molecular dynamics simulations. Adv Polym Sci 173:105–147. doi:10.1007/B99427
27. Read RJ, Chavali G (2007) Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 69(Suppl 8):27–37. doi:10.1002/prot.21662