Application of the Wang–Landau algorithm to the dimerization of glycophorin A

Claire Gervais,[1,a)] Thomas Wüst,[2] D. P. Landau,[2] and Ying Xu[1,b)]
[1]*Department of Biochemistry and Molecular Biology, Computational Systems Biology Laboratory and Institute of Bioinformatics, The University of Georgia, Athens, Georgia 30602, USA*
[2]*Center for Simulational Physics, The University of Georgia, Athens, Georgia 30602, USA*

(Received 12 March 2009; accepted 13 May 2009; published online 5 June 2009)

A two-step Monte Carlo procedure is developed to investigate the dimerization process of the homodimer glycophorin A. In the first step, the energy density of states of the system is estimated by the Wang–Landau algorithm. In the second step, a production run is performed during which various energetical and structural observables are sampled to provide insight into the thermodynamics of the system. All seven residues LIxxGVxxGVxxT constituting the contact interface play a dominating role in the dimerization, however at different stages of the process. The *leucine* motif and to some extent the GxxxG motif are involved at the very beginning of the dimerization when the two helices come into contact, ensuring an interface already similar to the native one. At a lower temperature, the *threonine* motif stabilizes by hydrogen bonding the dimer, which finally converges toward its native state at around 300 K. The power and flexibility of the procedure employed here makes it an interesting alternative to other Monte Carlo methods for the study of similar protein systems. © *2009 American Institute of Physics*. [DOI: 10.1063/1.3148186]

I. INTRODUCTION

The process by which proteins fold into their native three-dimensional structures has been investigated extensively in the past three decades and several breakthroughs have been achieved thanks to the development of a variety of computational methods at different levels of model complexity.[1–3] On the one hand, simple coarse-grained lattice models composed of beads representing hydrophobic and polar residues[4,5] led to the conclusion that hydrophobicity is the main interaction driving the folding of soluble proteins. On the other hand, studies of specific proteins on a full-atomistic level allowed the deciphering of qualitative/quantitative rules governing the folding process (e.g., notion of cooperativity) as well as the kinetics of such systems.[6–9]

One particularly promising approach to investigate the thermodynamics of protein folding is through examining the energy landscape.[10] The exploration of funnel-shaped energy landscapes, characteristic of proteins,[5] informs us about possible folding pathways and folding intermediates and thus allows us to bridge the gap between statistical results (i.e., microscopic point of view) and experimental (i.e., macroscopic) observations.[11]

Because the energy landscape of proteins is rough (characterized by a multitude of local energy minima separated by high energy barriers), one of the main challenges lies in finding efficient simulation methods capable of sampling the entire conformational space without being trapped in local energy minima.[12] One way to overcome this difficulty is to perform Monte Carlo (MC) or molecular dynamics simulations in a so called generalized ensemble in order to facilitate the crossing of barriers and to sample energy/conformational space more efficiently. Several powerful approaches emerged for this purpose over the last two decades, such as multicanonical sampling,[13] simulated annealing,[14] parallel tempering/replica exchange,[15,16] or metadynamics.[17] Furthermore, recent efforts focused on the combination of different methods in order to overcome the deficiencies inherent to each of them as well as to make them more applicable to complex (biological) systems such as proteins; these include, e.g., combining replica exchange with multicanonical sampling[18,19] or energy space metadynamics.[20]

In this paper we present a MC procedure based on the Wang–Landau algorithm,[21–23] which allows both the prediction of the structure of the dimer glycophorin A[24] and the analysis of its dimerization thermodynamics. The procedure consists of two steps: (i) estimating the energy density of states (DOS) of the system, and (ii) performing a "production run" (multicanonical sampling), based on the determined DOS, during which various structural and energetical quantities are sampled.

Our goal is to present the methodology in a clear, self-consistent way, accessible to any scientist with a basic knowledge of MC simulations. In the first part of this paper, we familiarize the reader with the notion of DOS before explaining in detail the two-step simulation procedure, its advantages, and limitations.

In the second part of this paper, the approach is applied to the investigation of the membrane protein glycophorin A,[24] a transmembrane helix dimer often used as a testing ground for developing computational methods and energy functions for membrane proteins. Apart from successfully reaching a native structure within 0.5 Å root mean square deviation (RMSD) of the nuclear magnetic resonance (NMR)

[a)]Electronic mail: clairegervais@csbl.bmb.uga.edu.
[b)]Electronic mail: xyn@bmb.uga.edu.

structure, our approach allows us to unravel the various thermodynamic/structural changes taking place during the dimerization process. Specifically, our analysis focuses on (i) the general thermodynamics of the system, (ii) the temperature dependence of structural and energetical features, and (iii) the probability distribution of observables at several temperatures below and above the dimerization transition point. The contact motif of the homodimer is analyzed and some conclusions are drawn about the influence of each of the seven residues constituting the motif.

II. COMPUTATIONAL PROCEDURE

A. The density of states and thermodynamics

In statistical mechanics, the partition function Z is a central quantity, which relates the statistics of a system to its thermodynamic properties. In the canonical ensemble (i.e., the number of particles and the volume are fixed but the system is allowed to exchange heat with the exterior), $Z(T)$ is a measure of the number of states accessible to the system at a given temperature T,

$$Z(T) = \int_x e^{-E(x)/k_B T} dx. \tag{1}$$

Here the integral runs over all protein conformations (x), which are given by the entire set of coordinates of each atom in three-dimensional space, and k_B is Boltzmann's constant. Instead of integrating over all conformations (x), $Z(T)$ can be calculated by integrating over energy (E),

$$Z(T) = \int_E g(E) e^{-E/k_B T} dE, \tag{2}$$

where $g(E)$ denotes the DOS, which is a measure of the degeneracy of the energy E. For a continuous system (e.g., a protein), the energy spectrum must be discretized in order to calculate $Z(T)$ numerically. The integrand is then replaced by a sum over discretized energy levels (bins) of a given width ΔE and $g(E)$ refers to the number of conformations having an energy between E and $E+\Delta E$,

$$Z(T) \approx \sum_E g(E) e^{-E/k_B T}. \tag{3}$$

Because $g(E)$ does not depend on temperature T, an estimate of $g(E)$ allows us to calculate the partition function $Z(T)$ and its derived thermodynamic properties at *any* temperature. For instance, the internal energy $U(T)$ (average potential energy) and the specific heat $C(T)$ (a measure of the fluctuations in the internal energy) are calculated as

$$U(T) = \frac{1}{Z(T)} \sum_E E g(E) e^{-E/k_B T} \equiv \langle E \rangle, \tag{4}$$

$$C(T) = \frac{\delta U(T)}{\delta T} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_B T^2}. \tag{5}$$

Moreover, by sampling structural properties of conformations populating a certain energy bin, it is possible to investigate the thermodynamic behavior of structural features as well and to obtain more detailed information concerning the folding process. This specific procedure is explained below.

B. Estimating the DOS by Wang-Landau sampling

Wang–Landau sampling is a MC algorithm which has proven to be successful for estimating the DOS $g(E)$ of a variety of statistical physical systems.[25–29] In this algorithm, the Metropolis acceptance criterion for the transition probability from a conformation with energy E_1 to a conformation with energy E_2 is replaced by an expression involving the instantaneous DOS,

$$p(E_1 \to E_2) = \min\left(\frac{g(E_1)}{g(E_2)}, 1 \right). \tag{6}$$

This rule implies that the simulation will have a tendency to sample conformations with small $g(E)$ (i.e., the generally less degenerate low-energy levels) with a higher probability. Ultimately, if the DOS was known, this algorithm would generate a random walk in energy space with a flat histogram, i.e., the probability to visit a conformation with energy E is uniform over the entire energy range.

Based on this idea, the *a priori* unknown DOS $g(E)$ is iteratively determined by performing a random walk in energy space seeking to sample a flat energy distribution. Initially, $g(E)=1$ for all E and a histogram $H(E)$ is established that will keep track of the number of visits at each energy level E. After each MC step, the DOS of the new conformation with energy E_2 (in case of acceptance) or of the previous one with energy E_1 (in case of rejection) is multiplied by a modification factor f (typically $f=e$ at the beginning) and the corresponding histogram entry $H(E)$ is incremented. Once the energy distribution is sufficiently flat [i.e., $H(E) \geq p\langle H(E) \rangle$, for all energies E, where $\langle H(E) \rangle$ is the average histogram and $0<p<1$ is the "flatness criterion"], the modification factor f is reduced to $\sqrt{f}$, H is reset to zero, and a new iteration begins. The process is repeated until f reaches a threshold [i.e., typically $\ln(f) \leq 10^{-6}$], below which $g(E)$ is considered to have converged toward the correct DOS. For details on the implementation of Wang–Landau sampling, we refer the reader to Ref. 25.

C. Thermodynamics of observables

Once the DOS $g(E)$ has been obtained, a second simulation with Wang–Landau sampling [i.e., the same acceptance rule as in Eq. (6)] is performed, however, this time without updating $g(E)$ anymore. The flat energy distribution during this production run enables an efficient sampling of structural and energetical information (observables), including those from conformational regions with low energies. In order to study the thermodynamic behavior of an observable $\mathcal{A}$, an estimate of the joint DOS $g(E,\mathcal{A})$ is required. For that purpose, a histogram $H(E,\mathcal{A})$ is accumulated during the production run and $g(E,\mathcal{A})$ is derived from the relation

$$H(E,\mathcal{A}) \sim \frac{g(E,\mathcal{A})}{g(E)}. \tag{7}$$

Then, the partition function is calculated by

$$Z_{\mathcal{A}}(T) = \sum_{E,\mathcal{A}} g(E,\mathcal{A}) e^{-E/k_B T} \tag{8}$$

and three thermodynamic quantities for the observable $\mathcal{A}$ can be evaluated at any temperature, namely, the distribution $p(\mathcal{A})$, the average $\langle \mathcal{A} \rangle$, and the free energy $F(\mathcal{A})$,

$$p(\mathcal{A}) = \frac{1}{Z_{\mathcal{A}}} \sum_{E} g(E,\mathcal{A}) e^{-E/k_B T}, \tag{9}$$

$$\langle \mathcal{A} \rangle = \sum_{\mathcal{A}} p(\mathcal{A}) \mathcal{A}, \tag{10}$$

$$F(\mathcal{A}) = -k_B T \ln(Z_{\mathcal{A}}). \tag{11}$$

Besides, it is often useful to examine the derivative $d\langle \mathcal{A} \rangle / dT$ in order to determine the location of possible transition points of $\mathcal{A}$.[30] Note that the transitions we found are not phase transitions in the exact thermodynamic sense, since we consider only a finite system here.[16,27]

III. APPLICATION TO GLYCOPHORIN A

A. Description of the model

The above procedure has been applied to investigate the thermodynamics of the homodimer glycophorin A.[24] The system under consideration is composed of two identical α-helices, A and B, of 22 residues each, running from E72 to Y93 (EITLIIFGVMAGVIGTILLISY). The helix backbone is a perfect α-helix and is kept fixed during our simulations. A unified atom representation is employed, where, in addition to all heavy atoms, only polar hydrogen atoms susceptible to being involved in hydrogen bonding are explicitly modeled (the total number of atoms is 378). Note that determining the DOS of a model that includes all the atomistic degrees of freedom of the membrane would not have been feasible. Therefore, the membrane is represented implicitly, that is, the interaction between membrane and protein is treated by a mean-field estimate only (see below). At the beginning of the simulation, the two helices are placed parallel to the normal of the membrane (40 Å thickness, defined as the z-axis), with their center of mass located around the center of the membrane (defined as $z=0$ Å). They are separated from each other by a distance of ~ 15 Å, with a randomly chosen orientation around their main axis. During the simulation, seven different types of MC trial moves were employed, designed to allow either global modifications of the protein or local changes in the conformation. These trial moves are (maximum magnitude and relative weights are given in parentheses, respectively): (i) translation of the protein along the z-axis (0.5 Å, 0.003), (ii) rotation of the protein around its center of mass (30°, 0.003), (iii) decrease/increase in the distance between the two helices in the xy-plane (0.01 Å, 0.004), (iv) translation of a helix (0.5 Å, 0.030), (v) rotation of a helix around its center of mass (17°, 0.030), (vi) rotation of a helix around its main axis (30°, 0.030), and (vii) dihedral rotation of the side chain of a residue (180°, 0.900).

The energy is based on the force-field CHARMM19 (Ref. 31) and a knowledge-based potential designed to take into account the membrane environment implicitly.[32,33] For our study, the energy can be decomposed into the following terms:

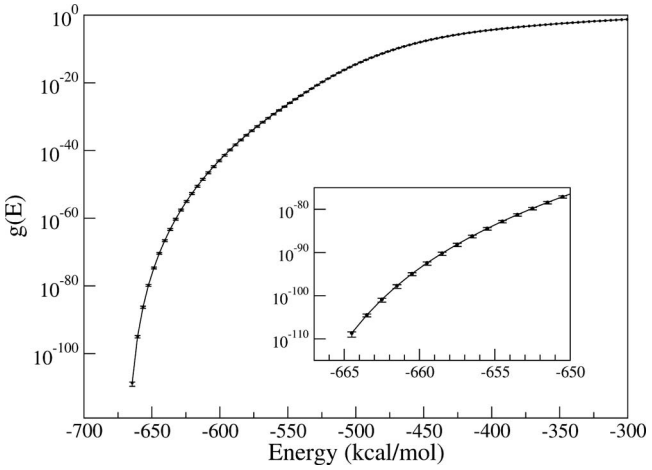

FIG. 1. Normalized DOS $g(E)$ for glycophorin A, averaged over ten runs (for clarity, only every fourth point is plotted on the main graph).

$$E = E^{A,B}_{\text{inter}} + E^{A}_{\text{intra}} + E^{B}_{\text{intra}} + E^{A}_{\text{lipid}} + E^{B}_{\text{lipid}}, \tag{12}$$

where $E^{A,B}_{\text{inter}}$ is the sum of the van der Waals and electrostatic energies between atoms of the helix A and atoms of the helix B; E^{A}_{intra} and E^{B}_{intra} define the sum of the van der Waals, electrostatic and dihedral energies within the helix A and B, respectively, see Ref. 31. Finally, E^{A}_{lipid} and E^{B}_{lipid} are the sums of the lipid-residue interactions of helixes A and B, respectively. The lipid energy is a function of the z-coordinates of the C_α atoms of the residues,

$$E_{\text{lipid}} = \sum_{\text{residues}} \mu_{\text{type}}(|z|). \tag{13}$$

It reflects the propensity of an amino acid to be located in different regions of the membrane. Energy values $\mu_{\text{type}}(|z|)$ for each of the 20 types of amino acids at distances $|z| = \{0,1,\ldots,40\}$ Å as well as further details can be found in Ref. 32. Note that this simplified representation of the membrane-protein interaction limits our study to qualitative observations only. In particular, values for transition temperatures obtained below should be taken with care as the flexibility and deformation of the membrane at high temperature is not included in our model. However, this does not prevent us from drawing some conclusions on the behavior of the dimer, especially at low and medium temperatures, where membrane-protein interactions are likely to play a limited role on the thermodynamics of the system.[34,35]

B. Determination of the density of states and production run

The DOS is estimated from an average of ten Wang–Landau runs using a final modification factor $\ln(f_{\text{final}}) = 10^{-6}$ and a flatness criterion of 0.2 (see Fig. 1). The large range of $g(E)$ (spanning nearly 110 orders of magnitude) illustrates the complexity of the type of system investigated here. For our system, the lowest energy found during a simulation was -666.7 kcal/mol. However, the DOS of this energy bin had

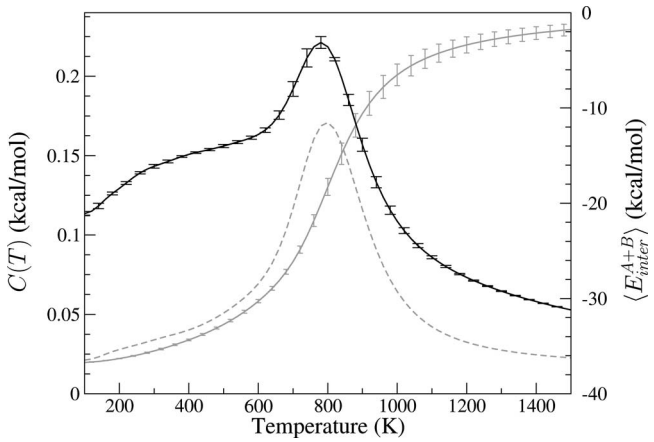

FIG. 2. Black: specific heat $C(T)$ of glycophorin A. Standard errors were estimated by a Jackknife analysis from ten runs. Note that only the temperature range for which results are reliable (i.e., $T>100$ K) is shown. Gray: variation in the interhelix energy $\langle E^{A,B}_{\text{inter}} \rangle$ (solid line) and its derivative (dashed line) as a function of temperature.

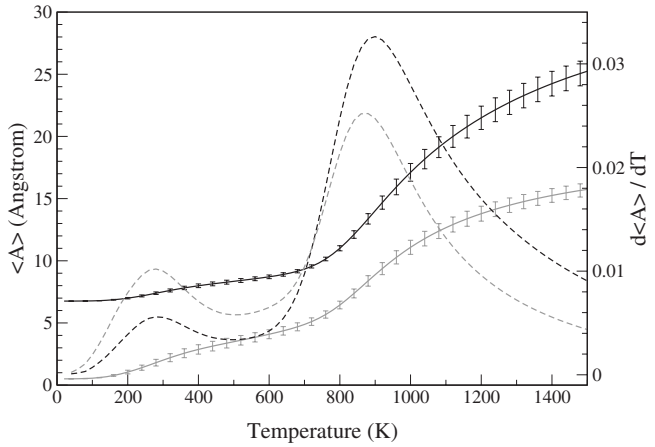

FIG. 3. Variation in $\langle A \rangle = \langle d^{A,B} \rangle$ (black solid line) and $\langle A \rangle = \langle \text{RMSD} \rangle$ (gray solid line) in function of temperature. The corresponding temperature derivatives $d\langle A \rangle / dT$ (black and gray dashed lines, respectively) highlight the two structural transitions.

only a negligible influence on the thermodynamic quantities at temperatures of interest (i.e., $T>200$ K). Therefore, limiting the energy range from -665 to -300 kcal/mol with a bin width $\Delta E=1$ kcal/mol allowed all runs to converge within about $100 h_{\text{CPU}}$ per run (AMD Opteron processors) and, at the same time, provided sufficient thermodynamic resolution. All results presented below were obtained by running and averaging five production runs of 8×10^8 MC moves each (about $100 h_{\text{CPU}}$ per run), a simulation length found sufficient to ensure the reliability of our results.

IV. RESULTS

A. Thermodynamics of global observables

The specific heat of the homodimer glycophorin A shows a significant peak at ~800 K followed by a shoulder around ~300 K (see Fig. 2). In order to investigate the nature of these two phenomena more closely, we performed production runs monitoring the interhelix interaction $E^{A,B}_{\text{inter}}$. The global structural changes taking place at the two temperatures were also monitored by examining two additional observables, namely, the distance between the centers of mass of the two helices ($d^{A,B}$) and the RMSD of the C_{alpha} atoms with respect to the experimental reference structure [model 1 of the NMR structure with PDB code 1AFO (Ref. 24)].

Figure 2 shows the thermodynamic average $\langle E^{A,B}_{\text{inter}} \rangle$ and its derivative as a function of temperature. While almost no effective interaction between the helices is present at 1100 K, the system undergoes a dimerization transition at ~800 K characterized by a significant peak in the derivative. This dimerization temperature is in agreement with previous MC studies of glycophorin A.[36] However, $d\langle E^{A,B}_{\text{inter}} \rangle / dT$ does not show another peak at lower temperatures, indicating that the interhelix interaction is not responsible for the shoulder observed in the specific heat at 300 K. The averages of the two structural observables $\langle d^{A,B} \rangle$ and $\langle RMSD \rangle$ show a similar behavior at high temperatures (see Fig. 3). A large structural transition slightly above 800 K is observed, indicating that a close contact between the helices has been established. However, contrary to $\langle E^{A,B}_{\text{inter}} \rangle$, an additional peak is observed around 300 K. At this temperature, the RMSD has fallen below 2 Å and the distribution of $d^{A,B}$ shows a preference for $d^{A,B} \approx 6.7$ Å, a value close to the native distance of 6.64 Å (see Fig. 4). These observations clearly suggest that this second peak corresponds to the association of the helices into a nativelike conformation. In order to obtain representative structures at this transition temperature (300 K), (i) we located the maximum of the system energy distribution $p(E)$ at 300 K (i.e., -633 ± 10 kcal/mol) and (ii) we collected conformations within this energy range and an interhelical distance of $d^{A,B} \approx 6.7$ Å. Most of these structures were found to be similar to the experimental structure (see Fig. 5).

B. Thermodynamics of motif-based observables

The motif assumed to be responsible for the dimerization of the homodimer glycophorin A is composed of seven residues LIxxGVxxGVxxT.[37] The motif GxxxG, well known to promote dimerization in membrane proteins[37,38] and to favor helix-helix interactions in soluble proteins,[39] acts as an anchor point (Fig. 5, left). The glycine residues G79 and G83 facilitate the approach of the two helices because of their

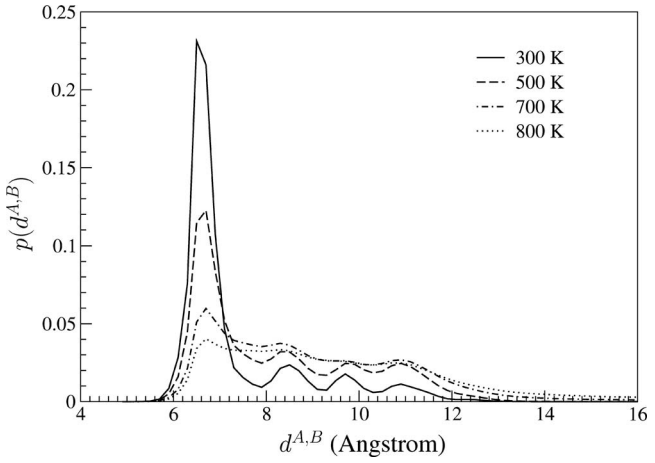

FIG. 4. Distribution of the helix-helix distance $d^{A,B}$ at various temperatures.

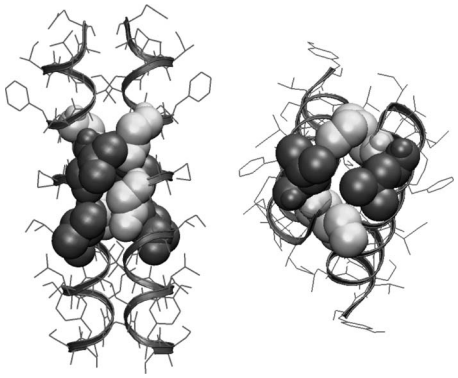

FIG. 5. Typical structure found at 300 K. Left: Side view. Glycine residues G79 and G83 (white) allow together with valine residues V80 and V84 (gray) a dense packing (Refs. 24 and 40) Right: Top view. Isoleucine I76 (white) and leucine L75 (gray) realize a hook, which stabilizes the dimer interface via hydrophobic interactions and close contact packing between branched residues. The same "handshake" pattern is observed in the GCN4 leucine zipper (Ref. 49).

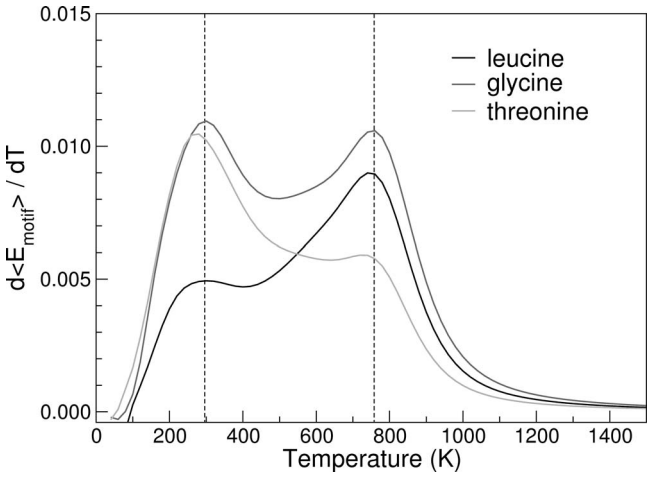

FIG. 6. Derivatives $d\langle E_{\text{motif}}\rangle/dT$ of the average intramotif energy, which highlight the two transition temperatures ~300 and ~800 K and the relative thermal stability of the motifs.

small size and their minimal entropic contribution (absence of side chain). Hence, they allow a dense packing in a "groove and ridge" fashion with the two neighboring valines V80 and V84.[24,40,41] Above this anchor point, leucine L75 and isoleucine I76 are known to be necessary for the stability of the dimer. Experimental mutagenesis studies[42] and molecular dynamics simulations[43] have shown that a mutation of L75 disrupts the helix-helix association due to a loss of favorable dispersion interactions. At the other end of the α-helix, threonine T87 stabilizes the dimer by forming an interhelical hydrogen bonding. The exact nature and location of the hydrogen bond is subject to debate and seems to depend on the environment (micelle or membrane) in which the protein is characterized.[24,44] Nonetheless, its importance in stabilizing the dimer is confirmed by both experimental and computational studies.[42,45,46]

Knowing the importance of these three strategic points at the dimer interface, we investigated their relative influence on the thermodynamics of dimerization more closely. For that purpose, we (i) calculated all interhelix residue-residue interactions, (ii) identified three motifs—defined by a group of residues—from the strengths of these interactions and the location of the residues in the dimer, and (iii) sampled the intramotif energies E_{motif} (i.e., the sum of residue-residue energies of a single motif) during a production run. Residue-residue interaction energies were obtained by performing standard Metropolis sampling at 300 K starting from a nativelike structure found during the DOS determination (RMSD=0.48 Å). The most significant interactions are shown in Table I and agree well with simulations published elsewhere.[47] The three following motifs were distinguished: The motif *leucine*, composed of the interactions of residues L75 and I76, the motif *glycine*, composed of the interactions of residues G79, V80 and G83, and the motif *threonine*, composed of the interactions of the residue T87 (see Table I for details).

The evolution with temperature of the average energies $\langle E_{\text{motif}}\rangle$, motif $\in$ {leucine, glycine, threonine} was monitored and the corresponding derivatives computed (see Fig. 6). All three motifs show qualitative thermodynamic behaviors similar to those already observed above, i.e., two peaks near ~800 and ~300 K, respectively. However, the significantly different peak magnitudes at the two transitions clearly indicate the difference in thermal stability of the three motifs. Whereas E_{leucine} undergoes mainly a transition at 800 K, E_{glycine} shows similar peak ratios at both 300 and 800 K, and $E_{\text{threonine}}$ has a major transition at 300 K. The stability can also be observed by looking, for instance, at the temperatures at which half of the corresponding ground state energies (value of $\langle E_{\text{motif}}\rangle$ at $T=0$ K) are reached. These are ~650,

TABLE I. Major residue-residue interaction energies (in kcal/mol) for the native state of glycophorin A. Energies included in the motifs leucine, glycine, and threonine are underlined, bolded, and italicized, respectively.

	L75	I76	G79	V80	A82	G83	V84	T87	I88
L75	<u>−0.77</u>	<u>−1.19</u>							
I76	<u>−1.19</u>		<u>−0.84</u>						
G79		<u>−0.84</u>	−0.63	**−0.96**					
V80			**−0.96**	**−0.97**	**−0.98**	**−1.40**			
A82				**−0.98**					
G83				**−1.40**			−1.01		
V84						−1.01	−0.81	*−1.78*	
T87							*−1.78*	*−0.34*	*−0.67*
I88								*−0.67*	

~570, and ~460 K for motifs leucine, glycine, and threonine, respectively, confirming the following order of stability: leucine > glycine > threonine.

V. DISCUSSION

Gathering the results for both global and motif-based observables, we can draw a "scenario" summarizing the structural and energetical changes taking place during the dimerization process of glycophorin A. Clearly, two transitions are observed, namely, at 300 and 800 K. The high temperature of 800 K can be explained by a neglect, in the various used potentials, of the excluded-volume effect induced by lipid molecules.[48] As a consequence, the attraction between the two helices is too strong and thus the temperature of dimerization is too high. However, the temperature difference between the two transitions is so large that even a possible shift in the temperature of the second transition would not lead to an overlap. One can therefore conclude positively on the existence of these two distinct dimerization stages. At 800 K, the two helices come into contact and interact with a significant interhelix energy, as exemplified by $\langle d^{A,B} \rangle$ and $\langle E_{\text{inter}}^{A,B} \rangle$, respectively. Structurally, the majority of the ensemble of dimers found at this temperature features some characteristics of the native structure. Indeed, the motif leucine and, to a lesser extent, the motif glycine already exhibit significant contributions to the overall ground state energy, an indication that the interface between the two helices brings into contact residues involved in these two motifs. This observation is not surprising considering that the GxxxG motif is known to promote dimerization of many membrane proteins.[37,38] Besides, leucine L75 and isoleucine I76 form a "hook" in the native structure, which stabilizes the dimer interface via hydrophobic interactions and close contact packing between branched residues, (see Fig. 5, right). The same leucine "handshake" pattern is observed in the GCN4 leucine zipper[49] and mutation studies confirmed the importance of L75 and I76 in stabilizing the dimer through favorable dispersive interactions.[43] Therefore, one can rationalize the early role played by the *leucine* motif in the dimerization by seeing it as a driving force to bring together the two helices and to help them in presenting the correct dimerization interface. However, an average RMSD of 5 Å with values broadly distributed between 0.5 and 9 Å indicates that the dimers around the first transition temperature (~800 K) still exhibit many different packing arrangements.

The second transition occurs around 300 K and corresponds to the convergence toward the native structure. The average RMSD falls below 2 Å and the motif *glycine* and especially the motif *threonine* undergo a transition toward the native energies. Stabilization of the dimer is affected via the formation of interhelical hydrogen bonding, as confirmed by the high interaction energy found between T87 and V84 of the opposite helix. The presence of hydrogen bonds between the β-hydroxyl group of T87 and the backbone carbonyl of V84 has indeed been observed in solid state NMR measurements of the glycophorin dimer in membrane bilayers.[45]

By an interesting mutation study, Melnyk *et al.*[50] proposed that the mechanism of dimerization of glycophorin A is essentially driven by a "long-range communication" between L75 and T87. Although in agreement with this proposition (both leucine and threonine motifs play a significant role in the stabilization of the dimer), the present study suggests that the two motifs act more in a hierarchical way rather than simultaneously, as shown by their different transition temperatures. In fact, our findings on the two-stage dimerization process of glycophorin A agree very well with the hypothesis proposed by Schneider.[51] Summarizing observations made from the formation of transmembrane helix oligomers (including glycophorin A), Schneider suggested decomposing the oligomerization into two stages. First, the contact between helices is promoted by a detailed fit between the helical surfaces, leading to close packing and van der Waals interactions. In a second stage, stabilization of the preformed dimer is obtained by electrostatic interactions, i.e., hydrogen bonding, or binding of a cofactor. In our simulations, we found indeed a first dimerization step governed by dispersive interactions (motif leucine) and close packing (motif glycine), while the second transition involved the formation of hydrogen bonds within the motif threonine.

VI. CONCLUSION

With the MC procedure based on Wang–Landau sampling presented in this work, we were able to successfully obtain a qualitative picture of the dimerization process of glycophorin A. The main advantage of this two-step approach lies in its *flexibility* as well as its *generality*. In particular, by separating the calculation of $g(E)$ and the production run, the number of different observables simultaneously sampled is principally not limited. This flexibility can be convenient when studying a system for which little is known at the beginning. Moreover, the method is widely applicable to any study of biological systems, such as the folding process of soluble proteins, polymers, DNA, or protein complexes. Therefore, it is an excellent alternative to other simulation methods used traditionally in the field of protein folding thermodynamics.

Our study of the dimerization process of glycophorin A confirmed the implication of all seven residues constituting the dimer motif. Their participation was found to happen in a hierarchical way. While residues of the motifs leucine and glycine play a role almost immediately when the two helices come into contact around 800 K, the motif threonine makes hydrogen bonding and stabilizes the dimer only at 300 K, i.e., the temperature at which convergence toward the native state occurs. This two-step dimerization process reinforces the idea that, unlike soluble proteins for which folding is mainly governed by the hydrophobic effect, the folding/oligomerization of membrane proteins is driven by a more subtle interplay between multiple types of interactions.

ACKNOWLEDGMENTS

We thank Daniel T. Seaton, James H. Prestegard, and Fang Tian for helpful discussions. This project is in part supported by Grant No. 1R01GM075331 from the National

Institutes of Health (NIH), grants from NSF (Grant Nos. DBI-0354771, ITR-IIS-0407204, CCF-0621700, and DBI-0542119) and a "Distinguished Scholar" from Georgia Cancer Coalition.

[1] K. A. Dill, Curr. Opin. Struct. Biol. **17**, 342 (2007).
[2] C. L. Brooks III, Acc. Chem. Res. **35**, 447 (2002).
[3] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[4] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[5] K. A. Dill, Protein Sci. **8**, 1166 (1999).
[6] D. Thirumalai, D. K. Klimov, and S. A. Woodson, Theor. Chem. Acc. **1**, 23 (1997).
[7] H. Lei, C. Wu, H. Liu, and Y. Duan, Proc. Natl. Acad. Sci. U.S.A. **104**, 4925 (2007).
[8] Y. Levy, P. G. Wolynes, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **101**, 511 (2004).
[9] Y. Levy, S. S. Cho, J. N. Onuchic, and P. G. Wolynes, J. Mol. Biol. **346**, 1121 (2005).
[10] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[11] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[12] T. Wüst, D. P. Landau, C. Gervais, and Y. Xu, Comput. Phys. Commun. **180**, 475 (2009).
[13] B. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).
[14] E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992).
[15] U. H. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[16] D. P. Landau and K. Binder, *A Guide To Monte Carlo Simulations in Statistical Physics*, 2nd ed. (Cambridge University Press, New York, 2005).
[17] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[18] Y. Okamoto, J. Mol. Graphics Modell. **22**, 425 (2004).
[19] A. Mitsutake, Y. Sugita, and Y. Okamoto, Biopolymers **60**, 96 (2001).
[20] H. Li, D. Min, Y. Liu, and W. Yang, J. Chem. Phys. **127**, 094101 (2007).
[21] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[22] F. Wang and D. P. Landau, Phys. Rev. E **64**, 056101 (2001).
[23] D. P. Landau and F. Wang, Comput. Phys. Commun. **147**, 674 (2002).
[24] K. R. MacKenzie, J. H. Prestegard, and D. M. Engelman, Science **276**, 131 (1997).
[25] D. P. Landau, S.-H. Tsai, and M. Exler, Am. J. Phys. **72**, 1294 (2004).
[26] F. Rampf, K. Binder, and W. Paul, J. Polym. Sci., Part B: Polym. Phys. **44**, 2542 (2006).
[27] D. T. Seaton, S. J. Mitchell, and D. P. Landau, Braz. J. Phys. **38**, 48 (2008).
[28] T. Wüst and D. P. Landau, Comput. Phys. Commun. **179**, 124 (2008).
[29] N. Rathore, T. A. Knotts IV, and J. J. de Pablo, J. Chem. Phys. **118**, 4285 (2003).
[30] J. Kim and T. Keyes, J. Phys. Chem. B **111**, 2647 (2007).
[31] E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
[32] Z. Chen and Y. Xu, Proteins **62**, 539 (2006).
[33] Z. Chen and Y. Xu, J. Bioinform. Comput. Biol. **4**, 317 (2006).
[34] T. Haltia and E. Freire, Biochim. Biophys. Acta **1228**, 1 (1995).
[35] J. U. Bowie, Curr. Opin. Struct. Biol. **11**, 397 (2001).
[36] H. Kokubo and Y. Okamoto, J. Chem. Phys. **120**, 10837 (2004).
[37] B. Brosig and D. Langosch, Protein Sci. **7**, 1052 (1998).
[38] W. P. Russ and D. M. Engelman, J. Mol. Biol. **296**, 911 (2000).
[39] G. Kleiger, R. Grothe, P. Mallick, and D. Eisenberg, Biochemistry **41**, 5990 (2002).
[40] D. Langosch and J. Heringa, Proteins **31**, 150 (1998).
[41] A. Senes, M. Gerstein, and D. M. Engelman, J. Mol. Biol. **296**, 921 (2000).
[42] A. K. Doura, F. J. Kobus, L. Dubrovsky, E. Hibbard, and K. G. Fleming, J. Mol. Biol. **341**, 991 (2004).
[43] J. Hénin, A. Pohorille, and C. Chipot, J. Am. Chem. Soc. **127**, 8478 (2005).
[44] S. O. Smith, D. Song, S. Shekar, M. Groesbeek, M. Ziliox, and S. Aimoto, Biochemistry **40**, 6553 (2001).
[45] S. O. Smith, M. Eilers, D. Song, E. Crocker, W. Ying, M. Groesbeek, G. Metz, M. Ziliox, and S. Aimoto, Biophys. J. **82**, 2476 (2002).
[46] J. M. Cuthbertson, P. J. Bond, and M. S. P. Sansom, Biochemistry **45**, 14298 (2006).
[47] M. Mottamal, J. Zhang, and T. Lazaridis, Proteins **62**, 996 (2006).
[48] P. Lagüe, M. J. Zuckermann, and B. Roux, Faraday Discuss. **111**, 165 (1999).
[49] E. O'Shea, J. Klemm, P. Kim, and T. Alber, Science **254**, 539 (1991).
[50] R. A. Melnyk, S. Kim, A. R. Curran, D. M. Engelman, J. U. Bowie, and C. M. Deber, J. Biol. Chem. **279**, 16591 (2004).
[51] D. Schneider, FEBS Lett. **577**, 5 (2004).