Monte Carlo simulations of polyalanine using a reduced model and statistics-based interaction potentials

Alan E. van Giessen and John E. Straub
Department of Chemistry, Boston University, Boston, Massachusetts 02215

(Received 1 September 2004; accepted 22 October 2004; published online 17 December 2004)

A coarse-grained residue-residue interaction potential derived from a statistical analysis of the Protein Data Bank is used to investigate the coil-to-helix transition for polyalanine. The interaction potentials depend on the radial distance between interaction sites, as well as the relative orientation of the sites. Two types of interaction sites are present in the model: a site representing the amino acid side chain, and a site representing a ‘‘virtual backbone,’’ i.e., a site located in the peptide bond which accounts for backbone hydrogen bonding. Two chain lengths are studied and the results for the thermodynamics of the coil-to-helix transition are analyzed in terms of the Zimm–Bragg model. Results agree qualitatively and quantitatively with all-atom Monte Carlo simulations and other reduced-model Monte Carlo simulations. © *2005 American Institute of Physics.*

[DOI: 10.1063/1.1833354]

I. INTRODUCTION

Some of the most important goals in biophysics and biochemistry are to be able to theoretically predict the native structure of a protein, the relation of structural transitions to function, and the nature of protein-protein interactions intrinsic to aggregation, given only the protein's amino acid sequence. The most accurate approaches employ all-atom molecular dynamics simulations using an explicit molecular representation of the solvent. At the present time, for studies of the thermodynamics of large-scale conformational transitions, such approaches are computationally too demanding in applications involving all but small peptides and proteins. Consequently, there is an on-going effort to develop methods to predict native structures of proteins using models with a reduced number of degrees of freedom. The most appealing approach is to include solvent effects implicitly in the interaction potentials and to replace the atoms in the amino acid residue by a small number of interaction sites, thereby drastically reducing the number of particles and interactions necessary for the calculation. However, in order to implement such a strategy, it is necessary to develop a set of residue-residue interaction potentials. A straightforward approach is to take advantage of the wealth of knowledge contained in the Protein Data Bank (PDB),[1] first suggested in the seminal work of Tanaka and Scheraga.[2] Subsequently, there have been numerous efforts to use the PDB to determine a useful and accurate set of ‘‘knowledge-based’’ or ‘‘statistics-based’’ potentials of mean force.[3–6] With a few exceptions,[6] most interaction potentials have been obtained solely in terms of residue-residue contacts.

Distance-dependent interaction potentials were introduced by Sippl[6,7] using the ‘‘Boltzmann device.’’ This method assumes that the protein structures in the PDB correspond to classical equilibrium states. From this assumption, it follows that the distance between any two side chains should also correspond to the equilibrium Boltzmann distribution. A potential of mean force can then be defined by

$$U_{ij}(r) = -kT \ln\left[\frac{f_{ij}(r)}{f_{\mathrm{ref}}(r)}\right], \tag{1}$$

where k is Boltzmann's constant, T is the temperature, $f_{ij}(r)$ is the probability density for a side chain of type i to be separated by a distance r from a side chain of type j, and $f_{\mathrm{ref}}(r)$ is a reference probability density. The choice of $f_{\mathrm{ref}}(r)$ is extremely important.[8] However, in recent years, a number of studies[7–15] have evaluated the goodness of knowledge-based residue-residue potentials; it was found that pairwise additive potentials dependent only on the radial distance between residues are inadequate for protein structure prediction. One major drawback of interaction potentials which are solely distance-dependent is that they neglect the relative orientation of the side chains. This is known to be important in side chain packing in the interior of proteins.

Recently, Buchete *et al.* developed a novel set of distance- and orientation-dependent residue-residue interaction potentials[16–18] which employ a local reference frame (LRF) to account for the relative orientation of the amino acid side chains. These potentials represent all 20 naturally occurring amino acids and an additional ‘‘virtual backbone’’ interaction site located in the center of the peptide bond. A spherical harmonic analysis (SHA) and synthesis (SHS) were used to develop a continuous interaction potential suitable to use in analyzing databases of decoy structures and in Monte Carlo simulations for the prediction of native state configurations. Compared with solely distance-dependent knowledge-based potentials, these new potentials have shown significantly improved performance in identifying the native state from a collection of near-native decoy configurations. In order to test the effectiveness of knowledge-based interaction potentials, Buchete *et al.*[17] developed a knowledge-based potential for water-water interactions and used it in a Monte Carlo simulation of liquid water. They calculated the radial distribution function and found good

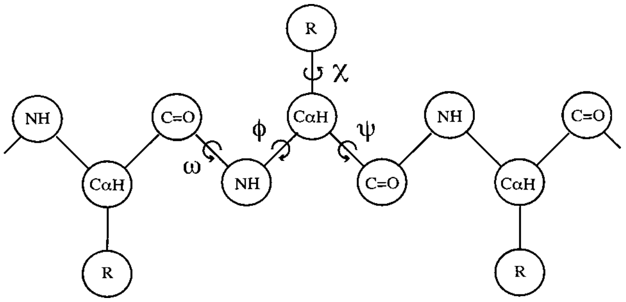

FIG. 1. Schematic representation of the peptide model showing the dihedral angles ϕ, ψ, ω, and χ. All bond lengths and bond angles are fixed.

TABLE I. Structural parameters.

van der Waals diameters	σ (Å)	σ_{local} (Å)
C_α	3.30	2.64
C'	3.56	2.94
N	2.94	2.36
C_β	4.50	4.50
Bond lengths	r (Å)	
C_α–C'	1.52	
C_α–N	1.45	
C'–N	1.33	
C_α–C_β	1.80	
Bond angles	degrees	
N–C_α–C'	111.6	
C_α–C'–N	117.5	
C'–N–C_α	120.0	
C'–C_α–C_β	110.0	
N–C_α–C_β	110.0	
Torsions	degrees	
ω	180.0	

agreement with experiment, to within the restrictions imposed by their choice of distance ranges.

In the current work, the knowledge-based residue-residue interaction potentials for all 20 amino acid residues and the "virtual backbone" are employed in Monte Carlo simulations of short polyalanine peptides. Sections II and III describe the peptide model and the interaction potentials as well as the details of the MC simulation. Results for the coil-to-helix transition are presented in Sec. IV, and our conclusions are presented in Sec. V.

II. PEPTIDE MODEL

The peptide model used in this work consists of two parts: the reduced structural model of a polypeptide chain, and the interaction potentials used to determine the energy of a given configuration.

The structural model consists of four particles or "united atoms" per amino acid residue, shown schematically in Fig. 1. Similar structural models have been used by others.[19,20] Three of these united atoms represent the peptide backbone: one represents the amide nitrogen and its hydrogen, another the α-carbon and its hydrogen, and the third the carbonyl carbon and its oxygen. This high level of backbone representation is essential for reproducing correct secondary structure in the folded peptide.[21] The fourth united atom represents the amino acid side chain. Bond angles and bond lengths, as well as the peptide dihedral angle ω, are held fixed at values given in Table I. The only structural degrees of freedom are the ϕ and ψ angles associated with each α-carbon. A rotational degree of freedom χ describing the relative orientation of the side chain is also present for amino acid residues other than Alanine and Glycine.

There are three different potentials used to determine the potential energy of a given configuration. They are the energy due to the statistics-based residue-residue interaction potential E_{SHS}, the energy due to the dihedral angle potential E_{TOR}, and the energy from van der Waals interactions E_{vdW}. The total energy is

$$E=\lambda E_{\mathrm{SHS}}+E_{\mathrm{TOR}}+E_{\mathrm{vdW}}, \tag{2}$$

where the parameter λ will be discussed below.

The van der Waals interaction is given by

$$E_{\mathrm{vdW}}=\sum_{i,j>i}\phi_{ij}(r), \tag{3}$$

where

$$\phi_{ij}(r)=4\varepsilon\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \tag{4}$$

The cross-diameter σ_{ij} is given by $\sigma_{ij}=0.5[\sigma_i+\sigma_j]$. For interactions between particles connected by three or fewer covalent bonds, the interaction strength ε and the diameter σ_{ij} are replaced by their reduced or "local" counterparts $\varepsilon_{\mathrm{local}}$ and $\sigma_{ij,\mathrm{local}}$. The reduced parameters are introduced because these short-range interactions are better modeled by using the atomic parameters instead of the united atom parameters. Values for these parameters are given in Tables I and II.

The dihedral angle energy is given by

$$E_{\mathrm{TOR}}=E_{\phi}+E_{\psi}, \tag{5}$$

with

TABLE II. Energetic parameters.

Torsion potential	kcal/mol	
$v_{2,\phi}$	0.0	$-\pi<\phi<0$
$v_{2,\phi}$	2.0	$0<\phi<\pi$
$v_{3,\phi}$	1.5	$-\pi<\phi<-\pi/3$
$v_{3,\phi}$	6.0	$-\pi/3<\phi<\pi/3$
$v_{3,\phi}$	1.5	$\pi/3<\phi<\pi$
$v_{2,\psi}$	0.0	$-\pi<\psi<0$
$v_{2,\psi}$	0.4	$0<\psi<\pi$
$v_{3,\psi}$	5.0	$-\pi<\psi<-\pi/3$
$v_{3,\psi}$	6.0	$-\pi/3<\psi<\pi/3$
$v_{3,\psi}$	0.1	$\pi/3<\psi<\pi$
van der Waals potential	kcal/mol	
ε	0.060	
$\varepsilon_{\mathrm{local}}$	0.033	

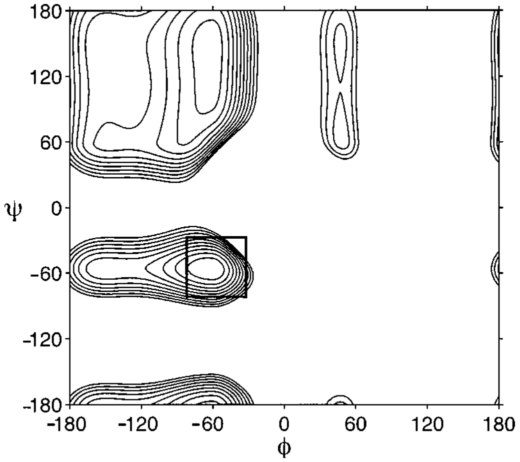

FIG. 2. Ramachandran plot for the alanine dipeptide. Contour lines are 0.5 kcal/mol apart. The global minimum is at $\phi=-62$, $\psi=-56$. The "box" defines the α-helical region (see text for details).

$$E_{\phi}=\sum_{i}\frac{1}{2}[v_{2,\phi}(1-\cos 2\phi_i)+v_{3,\phi}(1+\cos 3\phi_i)] \tag{6}$$

and

$$E_{\psi}=\sum_{i}\frac{1}{2}[v_{2,\psi}(1-\cos 2\psi_i)+v_{3,\psi}(1+\cos 3\psi_i)]. \tag{7}$$

The values for $v_{2,\phi}$, $v_{3,\phi}$, $v_{2,\psi}$, and $v_{3,\psi}$ were carefully chosen to produce, in conjunction with the van der Waals potential, a Ramachandran plot with realistic energy barriers for the alanine dipeptide, shown in Fig. 2. These values are also given in Table II.

Finally, the statistics-based interaction potentials give rise to E_{SHS}, which are thoroughly discussed in Refs. 18 and 22. Briefly, the energy due to the interaction of sites is first determined through an analysis of the PDB and the use of a "Boltzmann device" to determine the potential of mean force,

$$U_{ij}(r,\phi,\theta)=-kT\ln\left[\frac{P_{ij}(r,\phi,\theta)}{P_{\mathrm{ref}}(r,\phi,\theta)}\right], \tag{8}$$

where the polar angles ϕ and θ are defined by the local reference frame for each interaction site. Note that the ϕ in Eq. (8) should not be confused with the backbone dihedral angle in Eq. (6). The analysis is done for three distance ranges: short range, $2.0\,\text{Å}<r\leq 5.6\,\text{Å}$; intermediate range, $5.6\,\text{Å}<r\leq 9.2\,\text{Å}$; and long range, $9.2\,\text{Å}<r\leq 12.9\,\text{Å}$. The reference probability distribution, P_{ref}, is taken to be the corresponding radial or angular pair distributions obtained through an analysis of all 20 residue types.[18] Once $U_{ij}(r,\phi,\theta)$ is determined, it is then decomposed for each distance range using

$$U(\phi,\theta)=\sum_{m,n}[a_{mn}Y^{o}_{nm}(\phi,\theta)+b_{mn}Y^{e}_{nm}(\phi,\theta)], \tag{9}$$

where $Y^{o,e}_{nm}$ are odd and even complex spherical harmonics, and a_{mn} and b_{mn} are the expansion coefficients. For clarity, we have neglected to show the distance dependence of a and b. This spherical harmonic analysis has two important benefits: (1) once the coefficients are known, the potential can then be reconstructed for any given set of r, ϕ, and θ; and (2) the reconstructed potentials are smoothed relative to their original form given in Eq. (8). The energy of interaction between two residues is given by the spherical harmonic synthesis (SHS) formula

$$E(r,\phi,\theta)=\sum_{n=0}^{N_{\mathrm{SHS}}}\sum_{m=0}^{n}{}' P^{m}_{n}(\cos\theta)\times[a_{mn}\cos(m\phi)+b_{mn}\sin(m\phi)], \tag{10}$$

where P^{m}_{n} is the associated Legendre function. The prime notation on the second sum indicates that the $m=0$ term must be multiplied by 0.5. Note that this interaction is in general *not* symmetric, i.e., $E_{ij}\neq E_{ji}$. As used by Buchete *et al.*, $N_{\mathrm{SHS}}=13$, motivated by considerations related to extracting accurate statistical data on the relative residue orientations from experimentally resolved structures (see the Appendix of Ref. 16). However, it is possible to take fewer terms in this series. Doing so results in a further coarse-graining of the potential. There are two types of interaction sites: side chain and backbone. The side chain interaction site is located at the geometric center of the amino acid side chain while the "virtual backbone" site is located in the peptide bond midway between the amide nitrogen and the carbonyl carbon.

The introduction of the van der Waals and dihedral angle potentials is necessary in order for the polypeptide chain to adopt realistic secondary structure. The statistics-based interaction potentials by themselves cannot reproduce the correct ϕ and ψ angles and could not guard against overlapping backbone particles generated by a Monte Carlo move. However, their introduction poses a problem of energy scales: while values for the parameters for the van der Waals and dihedral angle potentials are known in specific units (e.g., kcal/mol), giving these potentials a well-defined energy scale, the corresponding energy scale for the statistics based potentials is unknown. Instead, these potentials are based on relative energies, and in their construction via the "Boltzmann device" are scaled relative to kT. Here T can be thought of as an average temperature of the crystal structures in the PDB. Clearly, such an effective temperature is not well-defined. Consequently, the statistics-based energy function must be scaled by some parameter λ. The total energy is therefore

$$E=\lambda E_{\mathrm{SHS}}+E_{\mathrm{TOR}}+E_{\mathrm{vdW}}. \tag{11}$$

In order to determine the optimal value for λ, we have carried out simulations of the thermodynamics of the coil-to-helix transition for polyalanine.

III. MONTE CARLO SIMULATION METHODOLOGY

The model described above has three degrees of freedom: the ϕ and ψ angles of the peptide backbone and the χ angle of the orientation of the side chain. Consequently, all Monte Carlo updates will be done using these degrees of freedom. Currently, the model has only been applied to polyalanine, for which the potential energy is independent of the χ angle.

Two types of moves are present in the Monte Carlo move set. This first is a simple pivot move and results in updating both the ϕ and ψ angles of one or two residues (the second being within six residues of the first). This move is a slightly modified version of that proposed by Shimada *et al.*[23,24] Schematically, the update is $\varphi \rightarrow \varphi' = \varphi + \delta\varphi$, where $\delta\varphi$ is drawn from a Gaussian distribution with a variance of 4 deg centered on zero. This results in primarily local moves, though occasionally the angles are updated in such a way as to cause a large, global change in the polypeptide configuration. In order to introduce large conformational moves in a slightly more controlled way, the pivot update occasionally updates one of the angles by drawing from a distribution centered on ±120 deg instead of 0 deg.

Though the large conformational changes that can result from a fortuitous set of ϕ, ψ updates in the pivot move are not a problem for an isolated polypeptide at high temperatures, they can result in unacceptably low acceptance rates at low temperatures, when the peptide is in its native state, or in dense phases with large numbers of steric interactions. To improve sampling in such situations, we include a local, concerted-rotation-like move discussed by Favrin *et al.*[25]

This move consists of an update to eight contiguous dihedral angles, biased in such a way as to keep the ends of the chain approximately fixed in space. Concerted rotation moves rigorously keep the ends of the chain fixed in space. While this method can generate large, local changes in chain configuration, it is a difficult and computationally complex move. The move introduced by Favrin simply biases the updates in favor of local moves. The strength of the bias can be changed, as well as the step size. In outlining this move, we closely follow the discussion of Favrin *et al.*[25]

The move works by first choosing the $\nu=8$ dihedral angles to be updated, two each from residues k, $k+1$, $k+2$, and $k+3$; these angles can be represented as a vector $\vec{\varphi}$. The next step is to identify three particles in residues $k+3$ or $k+4$ which are to remain fixed in space, ensuring the locality of the move. In our procedure, they are the α-carbon and the carbonyl carbon on residue $k+3$ and the amide nitrogen on residue $k+4$. If the position vectors of these three particles are labeled $\vec{r}_I$, where $I=1$, 2, 3, then we can define the quantity Δ, such that

$$\Delta^2 = \sum_{I=1}^{3} (\delta \vec{r}_I)^2. \tag{12}$$

The bias towards local moves is introduced by biasing toward small changes in Δ. For small changes $\delta\vec{\varphi}$, Δ can be written

$$\Delta^2 \approx \sum_{i,j,=1}^{\nu} \delta\varphi_i G_{ij} \delta\varphi_j, \tag{13}$$

where the matrix $\mathbf{G}$ has the elements

$$G_{ij} = \sum_{I=1}^{3} \frac{\partial \vec{r}_I}{\partial \varphi_i} \cdot \frac{\partial \vec{r}_I}{\partial \varphi_j}. \tag{14}$$

The first step in the move is to draw the tentative new angles, $\vec{\varphi}'$, from the distribution

$$W(\vec{\varphi} \rightarrow \vec{\varphi}') = \frac{1}{\pi^3} (\det \mathbf{A})^{1/2} \exp\{-(\vec{\varphi} - \vec{\varphi}')^T \mathbf{A} (\vec{\varphi} - \vec{\varphi}')\}, \tag{15}$$

where

$$\mathbf{A} = \frac{a}{2}(\mathbf{1} + b\mathbf{G}). \tag{16}$$

The parameters a and b play crucial roles in the move. The step size, and therefore the acceptance rate, is controlled by a. Larger values of a decrease the step size, i.e., the average value of the components of $\delta\vec{\varphi}$, and consequently increase the acceptance rate. The parameter b determines the "localness" of the move. As is readily apparent, in the limit $b=0$, the components of $\delta\vec{\varphi}$ would be independent, and the update random. The limit of large b forces the update to be strongly biased towards local moves, which keeps the ends of the peptide fixed in space.

The next step is to accept the move with the probability

$$P_{\text{accept}} = \min\left(1, \frac{W(\vec{\varphi}' \rightarrow \vec{\varphi})}{W(\vec{\varphi} \rightarrow \vec{\varphi}')} \exp\{-(E'-E)/kT\}\right). \tag{17}$$

The factor $W(\vec{\varphi}' \rightarrow \vec{\varphi})/W(\vec{\varphi} \rightarrow \vec{\varphi}')$ is included in order to satisfy detailed balance. Details and a full account of how to execute the move can be found in Favrin *et al.*[25] and in Ref. 26. In the move set, the values for a and b differ above and below the folding temperature. Below T_f, the moves are strongly biased towards local moves, with $a=500$ and $b=5.0$, while above T_f, moves biased towards a more efficient sampling of conformation space are preferred, with $a=100$ and $b=0.1$.

In order to improve the sampling of phase space, the replica exchange method[27–29] is used. In this method, several *noninteracting* replicas are simulated in parallel, each at a different temperature. At regular intervals, a Monte Carlo exchange step is attempted between two replicas, say i and j, at neighboring temperatures, T_i and T_j. The transition probability of this replica exchange is given by

$$W(X \rightarrow X') = \begin{cases} 1 & \text{if } \Delta \leqslant 0, \\ \exp\{-\Delta\} & \text{if } \Delta > 0, \end{cases} \tag{18}$$

where

$$\Delta = (\beta_i - \beta_j)[E_j - E_i]. \tag{19}$$

Here, E_i is the potential energy for replica i at temperature $\beta_i = 1/kT_i$. The temperatures are chosen to be equally spaced on a logarithmic temperature scale. Exchanges are attempted every 500 MC steps, and the replica exchange acceptance ratios vary from 15% to 40%.

IV. RESULTS AND ANALYSIS

Simulations of polyalanine were performed for two different chain lengths: 10 and 16 residues, referred to as Ala_{10} and Ala_{16}, respectively. In order to determine the value of the parameter λ, the coil-to-helix transition was used as a benchmark. Values of λ ranging from 0.05 to 1.0 were tested by running simulations with all initial configurations as α-helices. The equilibrium distribution was typically reached

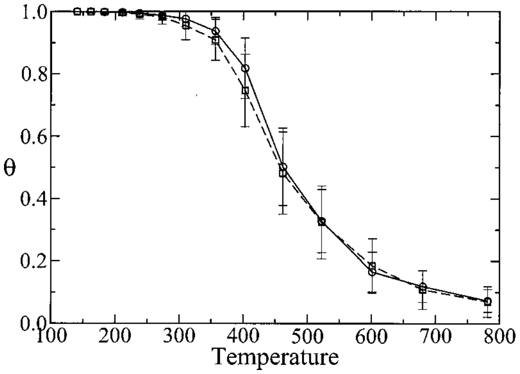

FIG. 3. Plot of the fractional helicity, θ, as a function of temperature. The solid line is for Ala_{16} and the dashed line is for Ala_{10}.

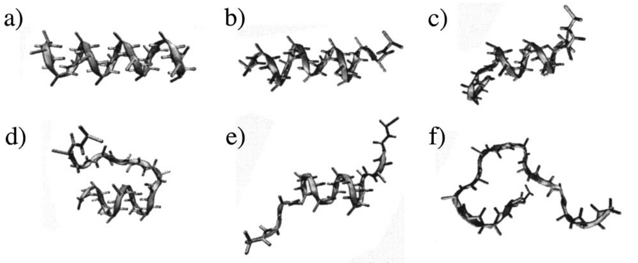

FIG. 4. Representative configurations of Ala_{16} for three different temperatures: (a) low temperature, (b)–(e) near the folding temperature, and (f) high temperature. Note that the hydrogen and oxygen atoms shown in this figure are not interaction sites in our coarse-grained peptide model. Their locations are determined by the geometry of each residue and are shown only to increase clarity.

within 1×10^6 MC steps. A further 4×10^6 MC steps were used to determine the approximate folding temperature. The value of $\lambda=0.1$ was chosen since it resulted in a reasonable transition temperature. To test the model, additional simulations were performed starting from an initial configuration set of all random coils: the ten-residue peptide reached its equilibrium distribution in 4×10^6 Monte Carlo steps. For the longer peptide, this number increased to 6×10^6. After choosing the value of λ, an additional 1×10^7 MC steps were run after the equilibrium distribution was reached in order to ensure convergence. The temperatures chosen are 141, 162, 183, 211, 238, 273, 310, 356, 403, 462, 523, 601, 680, and 782 K.

Figure 3 shows the fractional helicity, θ, for both the 10- and 16-residue polyalanine as a function of temperature (dashed and solid lines, respectively). The fractional helicity is defined by

$$\theta=\frac{N_{\mathrm{H}}}{N_{\mathrm{H}}^{\max}}, \tag{20}$$

where N_{H} is the number of helical hydrogen bonds, and $N_{\mathrm{H}}^{\max}$ is the maximum number of helical hydrogen bonds. For both peptides, θ is 1 at low temperatures, indicating that the peptide adopts a fully helical configuration, and approaches 0 at high temperatures, where it has no α-helical hydrogen bonds. The transition from $\theta=1$ to $\theta\approx0$ for the longer chain is slightly sharper, though both are centered at approximately the same temperature.

Several representative configurations of the Ala_{16} peptide are shown in Fig. 4. The low temperature α-helical conformation is shown in Fig. 4(a), while several conformations from the transition region are shown in Figs. 4(b)–4(e), and a high-temperature random coil is shown in Fig. 4(f). The average values for the ϕ and ψ angles in the helical residues are $(-58°,-53°)$, which compare favorably to the canonical crystal structure values of $(-57°,-47°)$.[30] The representative peptide conformations observed for the temperatures within the transition region all have significant helical segments.

It has been argued that the folding of a protein is characterized by two natural temperatures.[31] Both of these temperatures are expected to be higher for longer chain lengths, since a longer helix is energetically more stable than a shorter helix. The higher of the two temperatures, the collapse temperature T_θ, is the temperature below which the polypeptide adopts a more compact structure. T_θ is determined by the peak in the heat capacity,

$$C_v(T)=\frac{\partial E_{\text{total}}}{\partial T}=\frac{\langle E^2\rangle-\langle E\rangle^2}{k_BT^2}, \tag{21}$$

as a function of temperature. The heat capacity per particle, computed for the 10- and 16-residue peptides, is shown in Fig. 5. Averages for the statistical mechanical definition [second equality in Eq. (21)] were obtained using the weighted histogram analysis method,[32] and are in complete agreement with those from the thermodynamic defintion (first equality). There is a broad peak in the $C_v(T)$ of both peptides centered around 450 K, with the longer chain having a narrower peak and higher temperature, as expected. The locations of the peaks are $T_\theta=465$ K for Ala_{16} and $T_\theta=450$ K for Ala_{10}. The width of the transition region is approximately 200 deg, in agreement with other results for polyalanine.[33] The location of the peak is correlated with a large change in the radius of gyration (see Fig. 7 below).

The second natural temperature is the folding temperature, T_{f}, below which the polypeptide is predominantly in the native configuration. A measure of how much a given conformation differs from the native state is given by the parameter χ, called the "overlap function." There is no

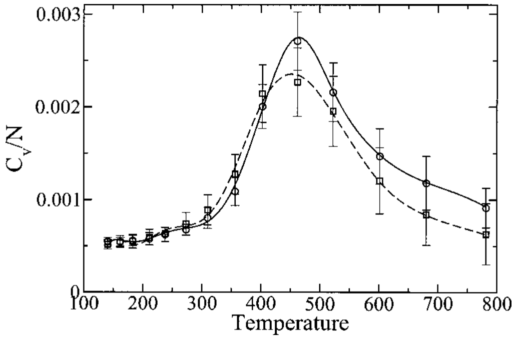

FIG. 5. The heat capacity per particle, C_v/N, as a function of temperature for Ala_{16} (solid line) and Ala_{10} (dashed line). The peak in each curve corresponds to the collapse temperature, T_θ, and is located at $T_\theta=465$ K and $T_\theta=450$ K for Ala_{16} and Ala_{10}, respectively.

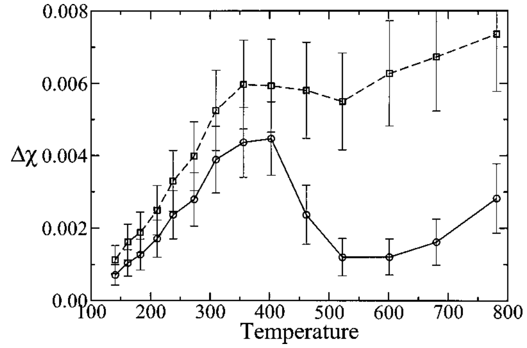

FIG. 6. The folding temperature T_{f} is given by the peak in $\Delta\chi$ versus temperature. The peaks are located at $T_{\mathrm{f}}=380$ K for Ala_{16} (solid line) and $T_{\mathrm{f}}=370$ K for Ala_{10} (dashed line).

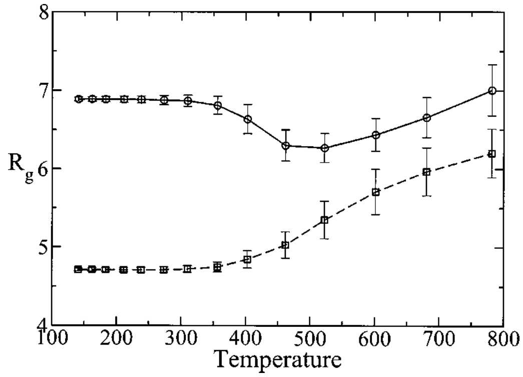

FIG. 7. The radius of gyration, in Å, as a function of temperature. The solid line is for Ala_{16} and the dashed line is for Ala_{10}.

unique way of defining such a parameter, though all reasonable definitions lead to similar results. We follow Vietshans *et al.*[31] in defining χ as

$$\chi=\frac{1}{N_\alpha^2-5N_\alpha+6}\sum_{i=1}^{N_\alpha-3}\sum_{j=i+3}^{N_\alpha}\Theta(\epsilon-|r_{ij}-r_{ij}^N|). \tag{22}$$

Here, N_α corresponds to the number of α-carbons, r_{ij} is the distance between α-carbons i and j, and r_{ij}^N is the same distance in the native state. Θ is the Heaviside function and the parameter ϵ is set to 0.5 Å. Note that χ is equal to 1 in the native state. We define the native state as a helix with ϕ and ψ angles of 58° and 53°, respectively. The fluctuations in χ are measured by

$$\Delta\chi=\langle\chi^2\rangle-\langle\chi\rangle^2. \tag{23}$$

Figure 6 shows the behavior of $\Delta\chi$ as a function of the temperature. Whereas the heat capacity has a peak at $T_\theta=465$ K for Ala_{16} and $T_\theta=450$ K for Ala_{10}, the peaks in $\Delta\chi$ are at lower temperatures: $T_{\mathrm{f}}=390$ K for Ala_{16} and $T_{\mathrm{f}}=375$ K for Ala_{10}. The difference in location of T_{f} for the two chain lengths is less than that seen in other studies.[33] As expected, both T_θ and T_{f} are higher for the longer chain. Hansmann and Okamoto[33] also determined that the transition temperature (T_c in their notation) scales with the chain length as $T_c(N)=T_c(\infty)-a\cdot\exp\{-bN\}$. Given the close agreement between our results and theirs, as discussed further below, we expect our model to show similar scaling behavior for both T_{f} and T_θ.

The behavior of $\Delta\chi$ for both chains is qualitatively similar. Two prominent features are present: a peak at the folding temperature and increasing structural fluctuations with increasing temperature. Quantitatively, the behavior of $\Delta\chi$ is quite different. The peak for Ala_{16} is very well defined and is clearly separated from the high temperature trend of increasing fluctuations. For Ala_{10}, however, these two features almost overlap. This is due to the small size of the ten-residue polypeptide. As shown by the behavior of the radius of gyration, Fig. 7, the size of Ala_{10} increases smoothly from the low-temperature helix to the high-temperature random coil. This is not so for Ala_{16}, which shows a prominent decrease in the radius of gyration at the collapse temperature before increasing with increasing T. It is this collapse which causes the structural fluctuations measured by $\Delta\chi$ to decrease for Ala_{16} and it is the absence of this collapse which results in the high value of $\Delta\chi$ at temperatures just above T_{f} for Ala_{10}.

The behavior of the energy with temperature is shown in Fig. 8 for the 16-residue polypeptide. In this plot, the energy due to the statistics-based interaction potential is shown, both as E_{SHS} and as the various contributions to E_{SHS}:E_{SS}, the contribution due to side chain-side chain interactions; E_{BB}, the contribution due to backbone-backbone interactions; and E_{SB}, the contribution due to side chain-backbone interactions. The various contributions are related by

$$E_{\mathrm{SHS}}=E_{\mathrm{SS}}+E_{\mathrm{SB}}+E_{\mathrm{BB}}. \tag{24}$$

All three quantities, and so too their sum, show a marked increase at the folding temperature. Surprisingly, the dominant contribution at low temperatures is from E_{SB}, followed closely by E_{BB}. While it is known that sidechain-backbone and backbone-backbone contacts generally account for approximately 30% of all contacts in α-helical proteins,[22] the large role played by such contacts here is due to the relatively weak interactions of the alanine sidechain, which is simply a methyl group. One would expect a much larger contribution from E_{SS} for a helical peptide with nonalanine sidechains. All three contributions to E_{SHS} show a dramatic change in the transition region, going from a large, negative contribution to a less negative or, in the case of E_{SS}, positive

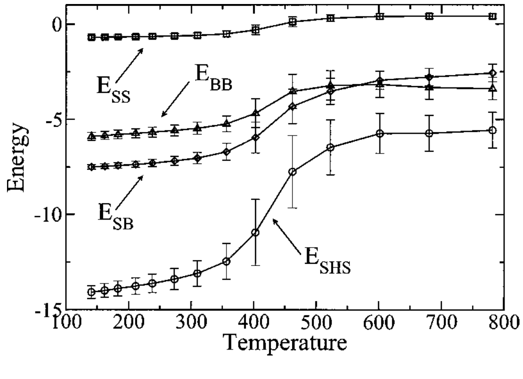

FIG. 8. The potential energy in kcal/mol for the statistics-based interactions of Ala_{16}. The circles are for the total statistical energy E_{SHS}; the squares are for the sidechain-sidechain interactions, E_{SS}; the triangles are for the backbone-backbone interactions, E_{BB}; and the diamonds are for the sidechain-backbone interactions, E_{SB}.

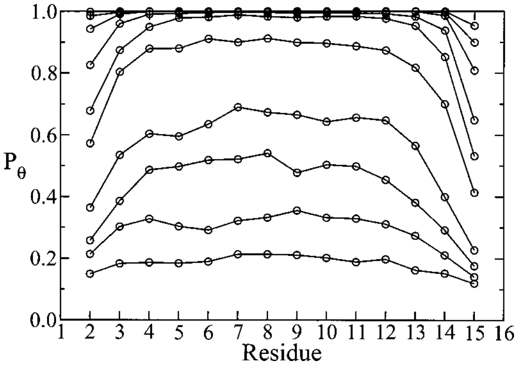

FIG. 9. The probability that a given residue is in the helical state for Ala_{16}, exclusive of the end residues. Low temperatures are at the top of the plot and high temperatures are at the bottom. The helical state is defined by the ϕ, ψ angles being within the range $-83° < \phi < -33°$ and $-78° < \psi < -28°$ (see "box" in Fig. 2). The residue index begins at the N-terminus. Note that the curves are typically asymmetric, with the C-terminus less likely to be helical than the N-terminus.

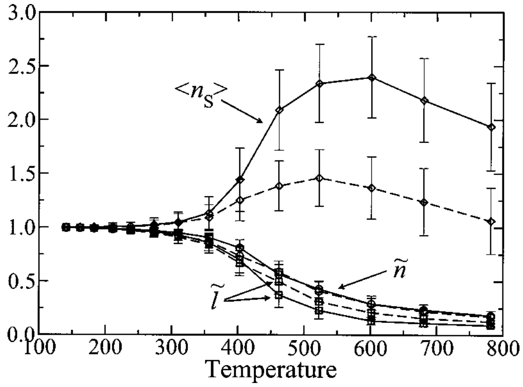

FIG. 10. A plot of the average number of helical residues, $\tilde{n}$ (circles), the average length of a helical segment, $\tilde{\ell}$ (squares), and the average number of helical segments, $\langle n_S \rangle$ (diamonds). The solid lines are for Ala_{16} and the dashed lines are for Ala_{10}.

contribution. E_{BB} even shows a slight peak at a temperature just above the folding temperature. The low temperature helix is primarily stabilized by side chain-backbone and backbone-backbone interactions.

Figure 9 shows P_θ, the probability for a given residue, exclusive of the two end residues, of Ala_{16} of being in a helical state, defined as the ϕ, ψ angles being within the range $-83° < \phi < -33°$ and $-78° < \psi < -28°$ (see the "box" in Fig. 2). The residue index begins at the N-terminus. For low temperatures, the probability is 1 for almost all the residues. For intermediate temperatures, the residues in the center of the peptide are more likely to be helical than the ends. For high temperatures, the probability is fairly flat, with only a slight peak for the central portion of the peptide. Interestingly, for all temperatures, the model predicts that the C-terminus is less likely to be in a helical state than the N-terminus. This asymmetry is a result of the angular-dependent knowledge-based potentials. The interaction potentials are anisotropic, and the LRFs are all oriented in the same direction relative to the axis of the helix. Since there is no side chain rotational degree of freedom for alanine, the residues at the N-terminus are in a different local environment than those at the C-terminus. Consequently, the probability that each terminus is in a helical state also differs.

We continue the discussion of our results with an analysis of the data in terms of the Zimm-Bragg model. In that model, for large numbers of residues, N_{res}, the average number of helical residues $\langle n_H \rangle$ and the average length of the helical segment $\tilde{\ell}$ are given by[33,34]

$$\frac{\langle n_H \rangle}{N_{\mathrm{res}}} = \frac{1}{2} - \frac{1-s}{2\sqrt{(1-s)^2+4s\sigma}}, \tag{25}$$

$$\tilde{\ell} = 1 + \frac{2s}{1-s+\sqrt{(1-s)^2+4s\sigma}}. \tag{26}$$

Here, s is related to helix propagation and σ to helix nucleation. The average number of helical residues $\langle n_H \rangle$ and the average number of helical segments $\langle n_S \rangle$ are determined directly from the simulation. The average length of a helical segment, $\tilde{\ell}$, is defined as $\tilde{\ell}\langle n_H \rangle / \langle n_S \rangle$, and is normalized such that it is 1 in the full helix. A helical segment consists of three or more consecutive residues in a helical state. Figure 10 shows the behavior of $\tilde{n} = \langle n_H \rangle / (N_{\mathrm{res}} - 2)$, $\tilde{\ell}$, and $\langle n_S \rangle$ as a function of temperature for both Ala_{10} and Ala_{16}. For low temperatures, all are equal to 1. As T increases above the folding temperature, and both $\tilde{\ell}$ and $\tilde{n}$ decrease while $\langle n_S \rangle$ increases, reaching a maximum at $T = 520\,\mathrm{K}$ for Ala_{10} and $T = 580\,\mathrm{K}$ for Ala_{16} and then decreasing. The values of $\tilde{n}$ for both chain lengths lie on top of each other. However, the average fractional length of a helical segment decreases faster for Ala_{16} than for Ala_{10}. Consequently, there are more helical segments for Ala_{16}. This is intuitively reasonable, as the longer chain has a greater likelihood of nucleating multiple helical segments.

The behaviors of the parameters s and σ are shown in Fig. 11. A well-known feature of the Zimm–Bragg model is that the helicity, θ, is equal to 0.5 when $s = 1$. For lower temperatures, $s > 1$ and the peptide adopts predominantly helical configurations; for higher temperatures, $s < 1$ and the peptide is predominantly a random coil. The temperatures at which $s = 1$, 460 K for Ala_{16} and 447 K for Ala_{10}, agree well with the fractional helicity shown in Fig. 3, which crosses $\theta = 0.5$ at 462 K and 457 K, respectively. The parameter σ is

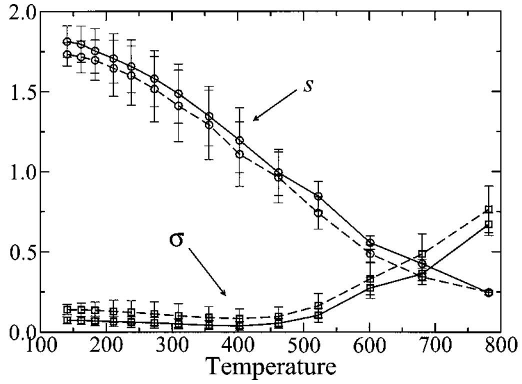

FIG. 11. The Zimm-Bragg parameters s (circles) and σ (squares) as a function of temperature for Ala_{16} (solid lines) and Ala_{10} (dashed lines).

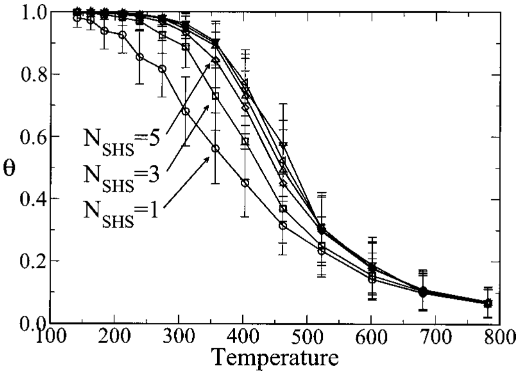

FIG. 12. A plot of the helicity, θ, as a function of the temperature for Ala_{10}. Each curve has a different number of terms in the spherical harmonic synthesis: 1 (circles), 3 (squares), 5 (diamonds), 7 (up triangle), 9 (down triangle), or 13 (left triangle). Data for $N_{\mathrm{SHS}}=1$, 3, and 5 are explicitly labeled. Data for $N_{\mathrm{SHS}}=7$, 9, and 13 lie on top of each other and are therefore not labeled.

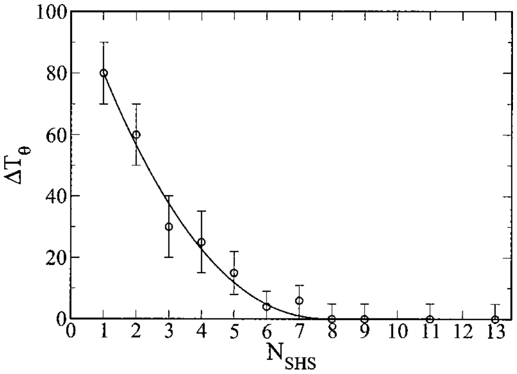

FIG. 13. A plot of ΔT_θ as a function of the number of terms in the spherical harmonic synthesis, N_{SHS}. ΔT_θ vanishes as N_{SHS}^{ν}, where $\nu=1.80$.

small at low temperatures: 0.07 for Ala_{16} and 0.12 for Ala_{10}. It decreases with increasing temperature until the folding temperature is reached, whereupon it rapidly increases. These values are in complete agreement with the all-atom multicanonical MC simulations of Hansmann and Okamoto[33] and with other Monte Carlo simulations using a coarse-grained model.[35]

As mentioned above, it is possible to further coarse-grain this model, and decrease the necessary computing time, by taking fewer terms in the spherical harmonic synthesis. Doing so has a marked effect on the location of the folding temperature. Figure 12 shows the fractional helicity for Ala_{10} for several simulations, each starting from the same set of initial configurations and each consisting of 5×10^6 MC steps. The number of terms taken in the evaluation of the statistics-based potentials for the curves plotted in Fig. 12 is $N_{\mathrm{SHS}}=1$, 3, 5, 7, 9, or the full 13 (data for $N_{\mathrm{SHS}}=2$, 4, 6, 8, 10, 11 and 12 not shown). The difference in θ for $N_{\mathrm{SHS}}=13$ between Fig. 12 and that in Fig. 3 is due to the latter being more fully converged. Figure 12 is only intended to illustrate the effect of changing N_{SHS}.

All simulations show the same behavior at high and low temperatures. Simulations with $N_{\mathrm{SHS}}<6$ show a transition region which becomes larger as N_{SHS} decreases. As would be expected, the more terms present in the SHS, the more stable the helix is, and the higher the folding temperature. The decrease in stability of the helix at low temperature is due principally to a significant decrease in the contribution to the energy primarily from E_{SB}. There is also a corresponding decrease in the magnitude of E_{BB}, which is near zero for low temperatures and $N_{\mathrm{SHS}}=1$. However, as there is still significant helical content in that case, as shown in Fig. 12, the stability of the helix must be due to the side chain-backbone bonding. The contribution to the energy from E_{SHS} for the helix decreases by 4 kcal/mol when N_{SHS} is decreased from 13 to 1. This is a significant destabilization and explains the decrease in the folding temperature.

The behavior of both C_{v} and $\Delta\chi$ were analyzed (data not shown). For $N_{\mathrm{SHS}}\geqslant 8$, the data were indistinguishable from the results using all 13 terms in the spherical harmonic synthesis. For $N_{\mathrm{SHS}}\leqslant 7$, the behavior of C_{v} and $\Delta\chi$ differed both from $N_{\mathrm{SHS}}=13$ and from each other. For the heat capacity, as N_{SHS} decreased below 8, the width of the peak broadened and the location of the maximum, which defines the collapse temperature T_θ, decreased. For $\Delta\chi$, the location of the maximum, which is the folding temperature T_{f}, and the width of the peak remained essentially constant, to within error. The height of the peak, however, decreased. Due to the short simulation length, data was insufficient to determine the location of the peak in $\Delta\chi$ for $N_{\mathrm{SHS}}\leqslant 4$. There was no peak for $\Delta\chi$ when $N_{\mathrm{SHS}}=1$ and 2, while for the same simulations, the peak in C_{v} was clearly identifiable, but extremely broad.

In order to estimate the convergence of $T_\theta(N_{\mathrm{SHS}})$ as N_{SHS} increases, we have plotted $\Delta T_0\equiv T_0(N_{\mathrm{SHS}}=13)-T_\theta(N_{\mathrm{SHS}})$ vs. N_{SHS} in Fig. 13. For values of $N_{\mathrm{SHS}}\geqslant 8$, the results are indistinguishable from those for $N_{\mathrm{SHS}}=13$, and $\Delta T_\theta=0$. For $N_{\mathrm{SHS}}\leqslant 8$, ΔT_θ vanishes as $\Delta T_\theta\sim N_{\mathrm{SHS}}^{\nu}$, where $\nu=2.25$. We conclude that, for polyalanine, it is sufficient to take only eight terms in the SHS.

The calculation of the local reference frames and of the SHS accounts for the majority of the simulation time. The total time for a 1×10^6 MC step simulation of Ala_{10} on a 1.3 GHz Pentium III processor scales as $t=B_0N_{\mathrm{SHS}}^{\mu}+B_1$, where $\mu=2.22$, and $B_0=0.71$ and $B_1=8.22$ are in minutes. When eight terms are taken in the SHS, i.e., $N_{\mathrm{SHS}}=8$, the time for 1×10^6 MC steps is 81 min. Compared to 220 min for the $N_{\mathrm{SHS}}=13$, this represents a significant savings in computation time.

V. CONCLUSION

Coarse-grained residue-residue interaction potentials derived from a statistical analysis of the Protein Data Bank have primarily been used to recognize the native structure of a protein from a set of decoy structures. In this work, we expand upon the work of Buchete *et al.*[17,18,22] and use a set of distance- and orientation-dependent interaction potentials in a Monte Carlo simulation of the coil-to-helix transition for two different polyalanine peptides. The model correctly predicts many features of the transition, including the chain length dependence of the folding temperature, and is in quantitative agreement with previous all-atom Monte Carlo studies. Along with the simulation of bulk water by Buchete *et al.*,[17] this work demonstrates the effectiveness for using knowledge-based interaction potentials to determine both the native structure of a given peptide and the thermodynamics associated with the folding transition.

ACKNOWLEDGMENTS

The authors would like to thank Dr. N.-V. Buchete for many helpful discussions. This work was funded by the NIH (R01 NS41356) and the Center for Computational Science at Boston University.

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[2] S. Tanaka and H. A. Scheraga, Macromolecules **9**, 945 (1976).
[3] J. Lee, A. Liwo, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **96**, 2025 (1999).
[4] S. Miyazawa and R. L. Jernigan, Proteins **34**, 49 (1999).
[5] A. Godzik, A. Kolinski, and J. Skolnick, Protein Sci. **4**, 2107 (1995).
[6] M. J. Sippl, J. Mol. Biol. **213**, 859 (1990).
[7] M. J. Sippl, Curr. Opin. Struct. Biol. **5**, 229 (1995).
[8] M. Betancourt and D. Thirumalai, Protein Sci. **8**, 361 (1999).
[9] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11101 (1998).
[10] A. Ben-Naim, J. Chem. Phys. **107**, 3698 (1997).
[11] D. Tobi and R. Elber, Proteins **41**, 40 (2000).
[12] D. Tobi, G. Shafran, N. Linial, and R. Elber, Proteins **40**, 71 (2000).
[13] J. Meller and R. Elber, Proteins **45**, 241 (2001).
[14] J. Meller, M. Wagner, and R. Elber, J. Comput. Chem. **23**, 111 (2002).
[15] D. Gatchell, S. Dennis, and S. Vajda, Proteins **41**, 518 (2000).
[16] N.-V. Buchete and J. E. Straub, J. Chem. Phys. **118**, 7658 (2003).
[17] N.-V. Buchete, J. E. Straub, and D. Thirumalai, Polymer **45**, 597 (2004).
[18] N.-V. Buchete, J. E. Straub, and D. Thirumalai, J. Mol. Graphics Modell. **22**, 441 (2004).
[19] S. Takada, Z. Luthey-Schulten, and P. G. Wolynes, J. Chem. Phys. **110**, 11616 (1999).
[20] A. Voegler Smith and C. K. Hall, J. Mol. Biol. **312**, 187 (2001).
[21] B. Honig and F. E. Cohen, Folding Des. **1**, R17 (1996).
[22] N.-V. Buchete, J. E. Straub, and D. Thirumalai, Protein Sci. **13**, 862 (2004).
[23] J. Shimada, E. L. Kussell, and E. I. Shakhnovich, J. Mol. Biol. **308**, 79 (2001).
[24] J. Shimada and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**, 11175 (2002).
[25] G. Favrin, A. Irbäck, and F. Sjunnesson, J. Chem. Phys. **114**, 8154 (2001).
[26] J. P. Ulmschneider and W. J. Jorgensen, J. Chem. Phys. **118**, 4261 (2003).
[27] R. H. Swendon and J.-S. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[28] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604 (1996).
[29] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **329**, 261 (2000).
[30] C. R. Cantor and P. R. Schimmel, *Biophysical Chemistry; Part I* (Freedman, New York, 1980).
[31] T. Veitshans, D. Klimov, and D. Thirumalai, Folding Des. **1**, 1 (1996).
[32] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[33] U. H. E. Hansmann and Y. Okamoto, J. Chem. Phys. **110**, 1267 (1999).
[34] B. H. Zimm and J. K. Bragg, J. Chem. Phys. **31**, 526 (1959).
[35] Y. Chen, Q. Zhang, and J. Ding, J. Chem. Phys. **120**, 3467 (2004).